Correcting the Bias of WRIGHT's Estimates of the Number of Genes Affecting a Quantitative Character: A Further Improved Method

Zhao-Bang Zeng

Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203

Manuscript received September 16, 1991
Accepted for publication May 4, 1992

ABSTRACT

Wright's method of estimating the number of genes contributing to the difference in a quantitative character between two populations involves observing the means and variances of the two parental populations and their hybrid populations. Although simple, Wright's method provides seriously biased estimates, largely due to linkage and unequal effects of alleles. A method is suggested to evaluate the bias of Wright's estimate, which relies on estimation of the mean recombination frequency between a pair of loci and a composite parameter of variability of allelic effects and frequencies among loci. Assuming that the loci are uniformly distributed in the genome, the mean recombination frequency can be calculated for some organisms. Theoretical analysis and an analysis of the Drosophila data on distributions of effects of *P* element inserts on bristle numbers indicate that the value of the composite parameter is likely to be about three or larger for many quantitative characters. There are, however, some serious problems with the current method, such as the irregular behavior of the statistic and large sampling variances of estimates. Because of that, the method is generally not recommended for use unless several favorable conditions are met. These conditions are: the two parental populations are many phenotypic standard deviations apart, linkage is not tight, and the sample size is very large. An example is given on the fruit weight of tomato from a cross with parental populations differing in means by more than 14 phenotypic standard deviations. It is estimated that the number of loci which account for 95% of the genic variance in the F_2 population is 16, with a 95% confidence interval of 7–28, and the effect of the leading locus is 13% of the parental difference, with 95% confidence interval 8.5–25.7%.

CORRECT estimates of the number of genes contributing to the genetic variation of quantitative characters within and between populations are of fundamental importance in quantitative genetics. The original method of WRIGHT (in CASTLE 1921), as elaborated by WRIGHT (1968), for estimating the number of genes is the simplest and most widely used method. The method relates the difference in the means of two inbred lines to the variance of their F_2 and backcross populations and relies on a number of assumptions. It has been known for a long time that the estimator is seriously biased. Since it was initially proposed, many authors have devised modifications for relaxing the assumptions or otherwise extending the applicability of the method. SEREBROVSKY (1928) suggested formulae utilizing backcross data for correction of dominance effects in simplified situations. DEMPSTER and SNYDER (1950) suggested a simplified way for the correction of linkage effects. LANDE (1981) pointed out that WRIGHT's method could also be used with outbred populations and also suggested that the same method could be applied to artificially selected lines from a single base population. COCKERHAM (1986) suggested an unbiased estimator of the difference in parental lines and also a method for combining the data from parental, F_1, F_2, and backcrosses into a single least-squares estimate. All of these analyses, however, addressed only the effects of relaxing some assumptions of the method, and the modified estimates are still seriously biased.

In a previous paper (ZENG, HOULE and COCKERHAM 1990, hereinafter referred to as ZHC), we explored the utility of selection lines for estimating the number of loci and the effect of selection, linkage and the distributions of allelic effects and frequencies in the base population on the estimation. We established that unequal effects of alleles and linkage are the most important factors which create bias in the estimator. In this paper a modification for correcting the bias from unequal effects of alleles and linkage is suggested. The modification relies on the estimation of two new parameters: the mean recombination frequency between a pair of loci and a composite measure of variability of allelic effects and frequencies among loci. With the assumption that loci are uniformly (*i.e.*, randomly) distributed in the genome, the mean recombination frequency can be inferred for some organisms. However, the variability of allelic effects has to be estimated from specially designed experiments for which an example is given for the effects of *P* element inserts on bristle numbers and viability in *Drosophila melanogaster.* In the absence of information of variability of allelic effects, a method is suggested to estimate the number of loci which account for most genetic variation. This method appears to be robust and informative. The effects of dominance on the estimator and the ways for correction of the resulting bias are also analyzed. Simulations are performed to show the statistical behavior and problems of the new method and the conditions for its possible application. Finally, as an example, the method is applied to fruit weight in a cross of tomato (POWERS 1942).

WRIGHT'S ESTIMATOR

WRIGHT's method involves the means and variances of two widely diverged populations on quantitative characters and variances of their hybrid populations. If we let μ_h and μ_l be the mean values of the two parental (high P_h and low P_l) populations for a quantitative character and σ_s^2 be the genetic variance stemming from differences in gene frequencies of the parental populations in the F_2 population from a cross of the two parental populations, WRIGHT's basic formula for estimating the number of loci, m, affecting a quantitative character is

$$\tilde{m} = \frac{(\mu_h - \mu_l)^2}{8\sigma_s^2}. \qquad (1)$$

There are many ways to estimate σ_s^2 (WRIGHT 1968; LANDE 1981). For example, σ_s^2 can be estimated as

$$\sigma_s^2 = \sigma_{F_2}^2 - (\sigma_h^2 + \sigma_l^2)/2$$

where σ_h^2 and σ_l^2 are the variances of the two parental populations and $\sigma_{F_2}^2$ is the variance of the F_2 population. Assuming additivity of gene effects, the different estimates given by LANDE (1981) have the same expectation, and it is better to combine different estimates by least squares (COCKERHAM 1986). With gene interaction, however, different estimates have different expectations (see below for dominance).

The estimator (1) is unbiased if the following four assumptions are true: [1] All alleles increasing the value of the character are fixed in one (high) population and all alleles decreasing the value of the character are fixed in the other (low) population; [2] Allelic effect differences are equal at all loci; [3] All loci are unlinked; and [4] All alleles interact additively within and between loci (*i.e.*, there is no dominance and epistasis).

In reality, however, none of these assumptions are likely to be true and violations of the assumptions can seriously bias the estimator. Thus the WRIGHT's estimate of the number of loci is usually called the "segregation index," "effective" or "minimum" number of loci or "factors."

Because the estimator is seriously biased, we seek ways and methods to reduce, estimate and correct the bias in order to let the estimator to be informative. The bias of the estimator has been examined systematically by ZHC. There are many ways to reduce the bias. For example, by performing strong divergent selection on the character, highly diverged lines can be created which would largely satisfy assumption [1] (see ZHC for detailed analysis). Choosing highly diverged inbred lines or natural populations for crosses, as suggested by LANDE (1981), may also maximize this assumption. [If, however, two inbred lines are not widely diverged for the character of interest, the best way to satisfy assumption [1] is to cross the lines and then apply strong divergent selection on the character. When the initial frequencies of alleles are 0.5, which is the case in a cross of two inbred lines, selection can be very effective in fixing the alleles in the appropriate selection lines (ZHC).] However, even if assumption [1] holds, violations of other assumptions can still have drastic effects on the estimates. Our previous analysis (ZHC) showed that most of the bias comes from unequal allelic effect differences and linkage. When the number of loci, m, is very large, $\tilde{m}$ at best reflects $1/(1 - 2\bar{r})$ rather than m where $\bar{r}$ is the mean recombination frequency between pairs of loci (TURELLI 1984; ZHC). Thus $\tilde{m}$ is probably not an informative estimator.

CORRECTING THE BIAS OF THE ESTIMATOR

A method: Expressed in terms of gene frequencies and allelic effects under the additive model, Equation 1 can be written as

$$\tilde{m} = \frac{\left\{\sum_{i}^{m} (p_{ih} - p_{il})a_i\right\}^2}{\sum_{i}^{m} (p_{ih} - p_{il})^2 a_i^2 + \sum\sum_{i\neq j}^{m} (1 - 2r_{ij})(p_{ih} - p_{il})(p_{jh} - p_{jl})a_i a_j} \qquad (2)$$

(ZHC) where a_i is the difference in absolute value between the effects of two homozygotes at the ith locus, p_{ih} and p_{il} are the gene frequencies of the high valued allele at the ith locus in the high and low parental populations, and r_{ij} is the recombination frequency between loci i and j. If r_{ij} and $(p_{ih} - p_{il})(p_{jh} - p_{jl})a_i a_j$ can be assumed to be independent (which may not be appropriate under selection unless $p_{ih} - p_{il} = 1$), Equation 2 can be expressed, averaged over loci, as

$$\tilde{m} = \frac{z + (m - 1)}{z + (m - 1)(1 - 2\bar{r})} \qquad (3)$$

where $\bar{r}$ is the mean frequency of recombination among loci and

$$z = \frac{\overline{(p_{ih} - p_{il})^2 a_i^2}}{\overline{[(p_{ih} - p_{il})(p_{jh} - p_{jl})a_i a_j]}}.$$

Here an overline indicates an average over loci or pairs of loci.

This expression suggests that a modified estimator, $\tilde{m}^*$, can be defined as

$$\tilde{m}^* = \frac{2\hat{\bar{r}}\tilde{m} + (\hat{z} - 1)(\tilde{m} - 1)}{1 - \tilde{m}(1 - 2\hat{\bar{r}})} \quad \hat{\bar{r}} > 0 \qquad (4)$$

to improve the precision in estimating m, where $\hat{z}$ is an estimate of the parameter z and $\hat{\bar{r}}$ is an estimate of $\bar{r}$. Because (4) is a ratio estimator, it is still biased (slightly upward) (APPENDIX A). Numerical analysis shows that, depending on many parameter values, the magnitude of this bias is roughly on the order of $m/10$. As discussed below, the effect of this bias on the estimation is likely to be small compared with sampling effects.

How can we estimate $\bar{r}$ and z? Generally we do not know $\bar{r}$ and are not able to estimate it directly. However, if we assume that genes are uniformly (*i.e.*, randomly) distributed in the genome, $\bar{r}$ can be estimated from the number and lengths of chromosomes of the organism concerned. For example, with HALDANE's mapping function $r_{ij} = 0.5(1 - e^{-2d_{ij}})$ where d_{ij} is the map distance between loci i and j, $\bar{r}$ can be estimated as

$$\hat{\bar{r}} = \frac{1}{2} - \frac{2C - M + \sum_{i=1}^{M} e^{-2c_i}}{4C} \qquad (5)$$

(APPENDIX B), where M is the number of haploid chromosomes, c_i is the genetic length (in Morgans) of the ith chromosome and $C = \sum_{k=1}^{M} c_k$. For M haploid chromosomes each with equal length c, this is $(1.36M - 1)/2.72M$, $(1.76M - 1)/3.52M$ and $(2.19M - 1)/4.38M$ for $c = 0.5$, 1 and 1.5 Morgans, respectively.

If c_i's can be estimated without bias, $\hat{\bar{r}}$ of (5) is unbiased under the assumption of the uniform distribution. Current estimates of genetic lengths of chromosomes, however, may be biased due to limited availability of genetic markers on the chromosomes. As the number of genetic markers increases, estimates of genetic lengths of the chromosomes may increase and so may $\hat{\bar{r}}$. There are two sources of sampling variation in $\hat{\bar{r}}$. One is the sampling variation of c_i's which depends on the method of estimating c_i's. The other is the finite number of underlying loci, m, involved in the study. For this part of the sampling, the sampling variance of $\hat{\bar{r}}$, $\sigma^2_{\hat{\bar{r}}}$, is analyzed in APPENDIX B, and the result shows that unless m is very small, $\sigma^2_{\hat{\bar{r}}}$ is always trivial.

If genes are not uniformly distributed in a genome, the mean recombination frequency will generally be smaller than that in (5). In the case of equal spacing of genes along the chromosomes, however, the expected mean recombination frequency is the same as in (5).

The parameter z is a function of variation of allelic effect differences and frequencies among loci. Assuming that $p_{ih} - p_{il} = 1$ (*i.e.*, assumption [1] is true) and a_i and a_j are independent, $z = \overline{a_i^2}/[\overline{a_i}]^2$. This parameter could be estimated from specially designed experiments. When individual allelic effects are normally distributed, the difference, a_i, between the two homozygote effects is then half-normally distributed, and $z = \pi/2 = 1.57$. There is, however, growing evidence to indicate that the distribution of allelic effects is likely to be highly leptokurtic (*e.g.*, MACKAY, LYMAN and JACKSON 1992) and, hence, z is likely to be larger than $\pi/2$. When $p_{ih} - p_{il} < 1$, the variation of allelic frequencies among loci will further increase the value of z. The behavior of z (which is equivalent to $1/Z$ without linkage in ZHC) with different distributions of allelic effects and initial gene frequencies under divergent selection has been analyzed in detail by ZHC. It seems that with reasonable assumptions about the distributions of allelic effects and initial gene frequencies and relatively strong selection intensity, the value of z at the selection limit is between 2 and 5 or even larger. The experimental data on the effects of P element inserts on bristle numbers in Drosophila analyzed below seem to support this argument.

If the individual allelic effects can be observed and n independent observations of a_i are available, an unbiased estimate of z is

$$\hat{z} = \frac{(n - 1)\sum_{i=1}^{n} a_i^2}{\left(\sum_{i=1}^{n} a_i\right)^2 - \sum_{i=1}^{n} a_i^2} \qquad (6)$$

and for a relatively large n the sampling variance of $\hat{z}$ can be approximated as

$$\sigma^2_{\hat{z}} \simeq \frac{1}{n}\left(\frac{[a^4]}{[a]^4} - \frac{[a^2]^2}{[a]^4} - \frac{4[a^2][a^3]}{[a]^5} + \frac{4[a^2]^3}{[a]^6}\right) \qquad (7)$$

where

$$[a^4] = a_{(4)}/n,$$

$$[a^2]^2 = [a_{(2)}^2 - a_{(4)}]/[n(n - 1)],$$

$$[a^2][a^3] = [a_{(2)}a_{(3)} - a_{(5)}]/[n(n - 1)],$$

$$[a^2]^3 = [a_{(2)}^3 - 3a_{(2)}a_{(4)} + 2a_{(6)}]/[n(n - 1)(n - 2)],$$

$$[a]^4 = [a_{(1)}^4 + 8a_{(1)}a_{(3)} + 3a_{(2)}^2 - 6a_{(1)}^2a_{(2)} - 6a_{(4)}]/ [n(n - 1)(n - 2)(n - 3)],$$

$$[a]^5 = [a_{(1)}^5 - 30a_{(1)}a_{(4)} - 20a_{(2)}a_{(3)} + 20a_{(1)}^2a_{(3)} + 15a_{(1)}a_{(2)}^2 - 10a_{(1)}^3a_{(2)} + 24a_{(5)}]/ [n(n - 1)(n - 2)(n - 3)(n - 4)],$$

$$
\begin{aligned}
[a]^6 = {} & [a^6_{(1)} + 144a_{(1)}a_{(5)} + 90a_{(2)}a_{(4)} + 40a^2_{(3)} \\
& - 90a^2_{(1)}a_{(4)} - 120a_{(1)}a_{(2)}a_{(3)} - 150a^3_{(2)} \\
& + 40a^3_{(1)}a_{(3)} + 45a^2_{(1)}a^2{}_{(2)} - 15a^4_{(1)}a_{(2)} \\
& - 120a_{(6)}]/[n(n-1)(n-2)(n-3) \\
& \cdot (n-4)(n-5)],
\end{aligned}
$$

and

$$a_{(k)} = \sum_{i=1}^{n} a_i^k \quad \text{for} \quad k = 1, 2, \cdots, 6.$$

These are used for the data discussed below.

In practice, however, necessary data for estimating z are hard to obtain. In the absence of independent estimates of z, a method is suggested below to estimate the number of loci which account for most genetic variation.

How much bias is likely to be in the estimates of $\tilde{m}$? This, of course, will depend on the values of $\tilde{m}$, $\bar{r}$ and z. Previously we have shown by simulations that in practice the expected value of $\tilde{m}$ might be about equal to the number, M, of haploid chromosomes if the number of genes, m, is much larger than M. Thus for M haploid chromosomes each with length one Morgan, the expected value of $\tilde{m}^*$ in practice may be about $1 + 2.32z(M-1)$, or seven times the number of chromosomes for $z = 3$. This can increase significantly if $\tilde{m}$ is larger than M, but less than $1/(1-2\bar{r})$ (Figure 1). On the other hand, there is relatively little difference between $\tilde{m}^*$ and $\tilde{m}$ when $\tilde{m}$ is significantly smaller than M.

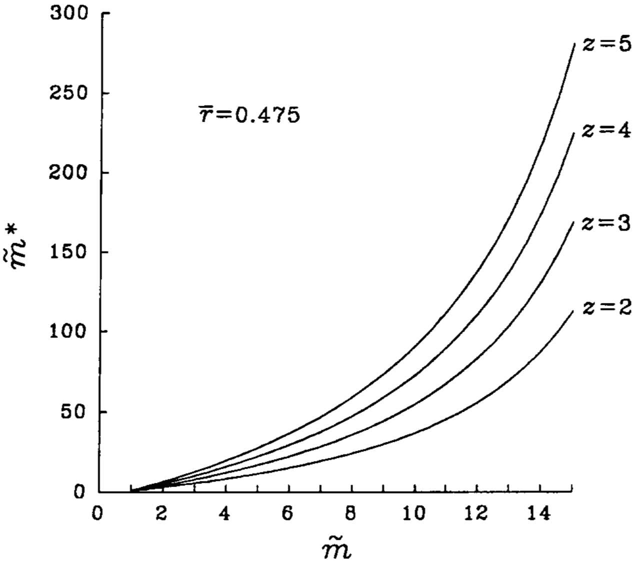

FIGURE 1.—The relation between $\tilde{m}$ and $\tilde{m}^*$ for z = 2, 3, 4, and 5. The mean recombination frequency, $\bar{r}$, used is 0.475.

Sampling variance: For WRIGHT's estimator, LANDE (1981) has given an approximate formula for calculating the sampling variance $\sigma^2_{\tilde{m}}$. The variance of the modified estimator involves the variance of estimates of $\bar{r}$ and z. By using the Taylor expansion on a ratio estimate (STUART and ORD 1987, pp. 325), the sampling variance of estimates of (4) can be approximated as

$$\sigma^2_{\tilde{m}^*} \simeq \tilde{m}^{*2} \frac{\begin{array}{c} 4\hat{\bar{r}}^2\hat{z}^2\sigma^2_{\tilde{m}} + 4\hat{z}^2(\tilde{m}-1)^2[\sigma^2_{\hat{m}} + \hat{m}^2]\sigma^2_{\bar{r}} \\ + [1 - \tilde{m}(1-2\hat{\bar{r}})]^2[\sigma^2_{\tilde{m}} + (\tilde{m}-1)^2]\sigma^2_{\hat{z}} \end{array}}{\begin{array}{c} [2\hat{\bar{r}}\tilde{m} + (\hat{z}-1)(\tilde{m}-1)]^2 \\ [1 - \tilde{m}(1-2\hat{\bar{r}})]^2 \end{array}} \tag{8}$$

There are, however, several problems in using this formula to calculate the sampling variance of $\tilde{m}^*$. Strictly speaking this approximation applies only when the denominator of the estimator (4), *i.e.*, $1 - \tilde{m}(1 - 2\hat{\bar{r}})$, is always greater than zero. When the denominator of a ratio of two random variables overlaps the zero region, the variance of the ratio does not exist and the observed sampling variance can be very large (CARSON and LANDE 1984; ZHC). This is the case for both WRIGHT's estimator and the modified estimator, and can cause serious problems in estimation. Furthermore, estimates of the sampling variance (8) depend critically on estimates of $\tilde{m}^*$. When $\tilde{m}^*$ happens to be small, $\sigma^2_{\tilde{m}^*}$ can be very small which may be misleading if it is used to construct a confidence interval of the estimate. On the other hand, when $\tilde{m}^*$ is large, $\sigma^2_{\tilde{m}^*}$ can be very large. Thus the meaning of estimates $\sigma^2_{\tilde{m}^*}$ should be interpreted cautiously.

In practice, however, $\sigma^2_{\hat{z}}$ and $\sigma^2_{\bar{r}}$ are unknown. As a rough guide, $\sigma^2_{\tilde{m}^*}$ may be further approximated by ignoring the $\sigma^2_{\bar{r}}$ and $\sigma^2_{\hat{z}}$ terms

$$\sigma^2_{\tilde{m}^*} \simeq \frac{4\hat{\bar{r}}^2\hat{z}^2\sigma^2_{\tilde{m}}\tilde{m}^{*2}}{\begin{array}{c} [2\hat{\bar{r}}\tilde{m} + (\hat{z}-1)(\hat{m}-1)]^2 \\ [1 - \tilde{m}(1-2\hat{\bar{r}})]^2 \end{array}}. \tag{9}$$

This is a minimum bound on the sampling variance. Simulation study shows, however, that most of the variation of $\tilde{m}^*$ is due to the variation of $\tilde{m}$ and relatively little to the variation of $\hat{\bar{r}}$ and $\hat{z}$. The difference between (8) and (9) is generally small. Alternatively, the sampling variance and the confidence interval of $\tilde{m}^*$ may be estimated by using bootstrap resampling of data, as CARSON and LANDE (1984) did for $\tilde{m}$.

Dominance: To reduce the effects of gene interaction on the estimation, WRIGHT (1968) and LANDE (1981) suggested that, before estimation, the measurement of the character should be transformed to a scale on which genetic variance is largely additive. To test whether the phenotypes satisfy the assumption of additivity, LANDE (1981) designed a triangular test which graphs the variances against the means to give a triangular pattern with F_1 and backcross populations at the midpoints of the edges connecting the parental and F_2 populations. This is a very useful test.

However, when the mean of the F_1 population deviates significantly from the midpoint of the two parental means, or the triangular pattern is significantly distorted even after the scale transformation of the measurement, we may have to consider the effects of gene interaction, such as dominance.

Let the effects of the three genotypes for two alleles at the *i*th locus be

Genotype	A_iA_i	A_ia_i	a_ia_i
Genotypic effects	a_i	$(1 + d_i)a_i/2$	0

where d_i is the degree of dominance. When $d_i = 0$, there is no dominance; $d_i = 1$, complete dominance of allele A_i; $d_i = -1$, complete dominance of allele a_i; and so on. With the assumptions of Hardy-Weinberg and linkage equilibria in the two parental populations and no epistasis, the phenotypic means and variances of different populations with dominance are listed in APPENDIX C.

With dominance the estimator of (1) is further biased, and different estimates of the denominator of (1) using variances of different populations suggested by LANDE (1981) will be different by expectation.

If $\mu_{F_1} > \frac{1}{2}(\mu_h + \mu_l)$, SEREBROVSKY (1928) suggested that it is better to use

$$\tilde{m} = \frac{(\mu_{F_1} - \mu_l)^2}{4(\sigma^2_{B_2} - \sigma^2_{F_1})} \tag{10}$$

to minimize the bias due to dominance, where μ_{F_1} and $\sigma^2_{F_1}$ are the mean and variance of the F_1 population and $\sigma^2_{B_2}$ is the variance of the backcross ($F_1 \times P_l$) population. This estimator, however, is unbiased by dominance only when the degrees of dominance and gene frequencies are constant among loci. Otherwise the estimator is still biased downwards by dominance. For (10), the modified estimator is that given by (4), where z is now

$$z = \frac{\overline{(p_{ih} - p_{il})^2(1 + (1 - 2p_{il})d_i)^2a_i^2}}{\overline{[(p_{ih} - p_{il})(p_{jh} - p_{jl})(1 + (1 - 2p_{il})d_i)}\;\overline{(1 + (1 - 2p_{jl})d_j)a_ia_j]}}$$

or if $p_{ih} - p_{il} = 1$

$$z = \frac{\overline{(1 + d_i)^2a_i^2}}{\overline{[(1 + d_i)a_i]}^2}$$

and is therefore a measure of variation of heterozygote genotypic effects on the character among loci. Since variation of d_i among loci increases z, the effect of dominance in this case is to reduce further the expected value of the estimator (10).

If $p_{ih} - p_{il} \simeq 1$, WRIGHT (1968, p. 394) suggested using

$$\tilde{m} = \frac{(\mu_h - \mu_l)^2}{8(2\sigma^2_{F_2} - \sigma^2_{B_1} - \sigma^2_{B_2})} \tag{11}$$

to minimize the effect of dominance, where $\sigma^2_{B_1}$ is the variance of the backcross ($F_1 \times P_h$) population. In this case the modified estimator is still that given by (4), but with $\hat{\bar{r}}$ multiplied by a factor

$$y = 1 + \left(1 - 2\frac{\overline{r^2}}{\bar{r}}\right)\left(\frac{\overline{d_ia_i}}{\overline{a_i}}\right)^2$$

or if a_i and d_i are independent

$$y = 1 + \left(1 - 2\frac{\overline{r^2}}{\bar{r}}\right)(\overline{d_i})^2$$

where $\overline{r^2}$ is the mean of squares of recombination frequencies. For an uniform distribution of genes in the genome

$$\frac{\overline{r^2}}{\bar{r}} = \frac{C^2 - \frac{3}{2}C + \frac{7}{8}M - \sum_{i=1}^{M} e^{-2c_i}\left(1 - \frac{1}{8}e^{-2c_i}\right)}{2C^2 - 2C + M - \sum_{i=1}^{M} e^{-2c_i}} < \frac{1}{2},$$

if $m \geqslant M$. When $m < M$, $\overline{r^2}/\bar{r}$ can be larger or smaller than ½.

Estimation of the parameter $\overline{a_i^2}/\bar{a}_i^2$–an example: As a measure of variation of allelic effects among loci, $\overline{a_i^2}/\bar{a}_i^2$ is a key parameter for correcting the bias due to unequal effects of alleles in estimating m. To estimate $\overline{a_i^2}/\bar{a}_i^2$, however, an experiment has to be able to identify the effects of individual alleles, and the effects of identifiable alleles have to cover the whole range of the distribution of allelic effects. That means that an experiment has to be able to identify not only alleles with large effects but also alleles with small effects. Experimental data which allow $\overline{a_i^2}/\bar{a}_i^2$ to be estimated are very scarce.

Recently, MACKAY, LYMAN and JACKSON (1992) reported observations of distributions of the effects of *P* element inserts on viability and abdominal and sternopleural bristle numbers in *D. melanogaster*. From an inbred host strain background free of *P* elements, they constructed 94 third chromosome lines by *P* element mutagenesis which contained on average 3.1 stable *P* element inserts. Both homozygous and heterozygous insert lines were constructed. By comparing the chromosome lines with inserts to insert-free control lines of the inbred host strain, the homozygote and heterozygote effects of the inserts on viability and abdominal and sternopleural bristle numbers can be estimated. The estimates of $\overline{a_i^2}/\bar{a}_i^2$ and $\overline{(1 + d_i)^2a_i^2}/[\overline{(1 + d_i)a_i}]^2$ using only those chromosome lines which have single *P* element inserts are shown in Table 1 with the homozygote and heterozygote effects being estimated as

$$a_i = |HOM_i - CON|$$

$$(1 + d_i)a_i = HET_i - \frac{HOM_i + CON}{2} + \frac{|HOM_i - CON|}{2}$$

corresponding with the above notation, where HOM_i

TABLE 1

Estimates of variability of homozygote and heterozygote effects of single *P* element inserts in *D. melanogaster*

Character	n	$\overline{a_i^2}/\bar{a}_i^2$	n	$\overline{(1 + d_i)^2a_i^2}/[\overline{(1 + d_i)a_i}]^2$
Abdominal bristles	29	2.94 ± 1.02	27	4.44 ± 1.56
Sternopleural bristles	29	2.48 ± 0.84	27	—[a]
Viability	35	1.73 ± 0.20	27	6.16 ± 4.60

Sample sizes (n) are given with estimates and their standard errors. Data are kindly provided by T. F. C. MACKAY.

[a] Estimate and standard error are too large and are omitted, see MACKAY, LYMAN and JACKSON (1992) for other similar observations on this part of the data.

and HET_i are the means of the ith homozygote and heterozygote insert line and CON is the mean of control lines.

Due to small sample sizes the sampling variances of estimates are quite high. Still, the estimates of $\overline{a_i^2}/\bar{a}_i^2$ for abdominal and sternopleural bristle numbers appear to be significantly larger than 1.57, the expected value for a normal distribution of allelic effects, as indicated by the simulated 95% confidence intervals (1.94, 5.02) for abdominal bristle number and (1.75, 4.07) for sternopleural bristle number, and also by the evidence that the distributions of the insert line means are significantly negatively skewed and leptokurtic (MACKAY, LYMAN and JACKSON 1992). (The bootstrap resampling estimates tend to be biased in this case.) Because the insert line means contain some environmental deviations, these estimates are likely to be underestimates of the effects of single P element insertions on the statistics (W. G. HILL in MACKAY, LYMAN and JACKSON 1992). The estimate for viability is relatively small, presumably partly because viability is bounded at one end by zero.

There was considerable variation in the degree of dominance (MACKAY, LYMAN and JACKSON 1992), so both the estimates and sampling errors of $\overline{(1 + d_i)^2a_i^2}/[\overline{(1 + d_i)a_i}]^2$ are quite high.

PROBLEMS AND UTILITY

There are many problems in using $\tilde{m}^*$ to correct the basis of $\tilde{m}$ and to estimate m. These are discussed here:

Sensitive estimator, large sampling covariance and low efficiency: The estimate $\tilde{m}^*$ is a nonlinear function of $\tilde{m}$ and $\hat{\bar{r}}$ (Figure 1) and is very sensitive to change in values of $\tilde{m}$ and $\hat{\bar{r}}$, particularly when $\tilde{m}$ is close to $1/(1 - 2\hat{\bar{r}})$. This sensitivity of $\tilde{m}^*$ is caused by the fact that $\tilde{m}$ is a very insensitive estimator of m (ZHC). The sampling variance of estimate of mean recombination frequency is generally small (APPENDIX B), but the sampling variance of $\tilde{m}$ can be very large (ZHC). As a result the sampling variance of $\tilde{m}^*$ can be extremely large. As $\tilde{m}^*$ relies on $\tilde{m}$, it inherits the properties and also problems of $\tilde{m}$. Thus the efficiency of the estimator can be very low, even if we know the correct values of z and $\bar{r}$.

High frequency of negative estimates: The estimate $\tilde{m}^*$ is positive only when WRIGHT's estimate, $\tilde{m}$, is bounded by

$$\frac{\hat{z} - 1}{2\hat{\bar{r}} + \hat{z} - 1} < \tilde{m} < \frac{1}{1 - 2\hat{\bar{r}}},$$

otherwise it will be negative. Although by expectation

$$1 \leqslant \mathscr{E}(\tilde{m}) < \frac{1}{1 - 2\bar{r}}$$

where $\mathscr{E}$ denotes expectation, the statistic $\tilde{m}$ can be smaller than 1 or larger than $1/(1 - 2\hat{\bar{r}})$ due to sampling. This can happen quite often when the mean difference between the two parental populations is a small number of phenotypic standard deviations, or linkage is tight (*i.e.*, the number of chromosomes is small), or the sample size is small (see below), and will result in many nonsensible estimates.

Difficulty in estimating the parameter z: Few data are available for estimating the parameter z. This is a fundamental parameter in quantitative genetics. Theoretical analysis and limited available data indicate that this parameter, taking only unequal allelic effects into account, is very likely to be larger than 2 for many quantitative characters like bristle numbers in Drosophila. If we also take the possible variation of allelic frequencies among loci into account, the likely value of the parameter is even larger. This means that ignoring the variation of allelic effects and frequencies among loci can seriously bias estimates of the number of genes. Fortunately, however, $\tilde{m}^*$ is a linear function of z and the bias in $\tilde{m}^*$ from using $\hat{z}$ is proportional to the difference $\hat{z} - z$. As m is a fundamental parameter in genetics, it is very desirable to estimate the parameter z and use it to correct, at least partially, the bias due to the inequality of allelic effects and frequencies. On the other hand, in many applications what is relevant is probably not the total number of loci, as this depends on the distribution of allelic effects, but the number of loci which account for most of the genetic variation. As shown below, estimates of the latter number are largely independent of the parameter z.

To illustrate some of these problems and also to evaluate conditions for the possible utility of the method, simulations were performed. In these simulations the two parental populations are assumed to be fixed with appropriate alleles (*i.e.*, the assumption [1] is assumed to be true) and differ in means by many environmental standard deviations (since the populations are fixed, all phenotypic variation is environmental). The allelic effect differences among m loci are assumed to be identically and independently dis-

TABLE 2

Simulation results

					$\tilde{m}$				$\tilde{m}^*$				$0 < \tilde{m}^* < 1000$		
m	M	n	D	h^{2a}	Mean	SD	90% interval		Mean	SD	90% interval		Mean	SD	P(ok)^b
20	3	100	5	0.487	3.86	1.68	1.96	7.13	64.93	1205.72	−145.29	125.02	36.66	68.06	0.844
			10	0.788	3.55	0.92	2.19	5.16	39.17	600.55	5.76	154.30	41.42	71.38	0.952
			15	0.890	3.45	0.76	2.24	4.79	29.11	91.43	6.66	94.89	31.24	35.06	0.980
		300	5	0.480	3.52	0.88	2.31	5.00	42.39	259.37	6.81	127.54	37.01	60.97	0.968
			10	0.788	3.51	0.68	2.40	4.72	37.32	110.94	8.64	93.81	34.63	50.28	0.990
			15	0.889	3.47	0.67	2.43	4.57	34.82	79.36	8.88	79.71	32.60	49.47	0.994
	10	100	5	0.367	6.92	14.82	2.88	15.32	−48.19	1548.99	4.85	79.35	32.72	60.60	0.952
			10	0.698	5.97	1.90	3.41	9.41	26.57	21.84	9.98	55.17	26.57	21.84	1.000
			15	0.836	5.72	1.61	3.32	8.46	23.55	12.09	9.59	44.06	23.55	12.09	1.000
		300	5	0.356	6.27	2.46	3.47	10.53	29.98	117.05	9.84	71.59	30.25	39.75	0.994
			10	0.695	5.83	1.45	3.69	8.40	23.94	10.47	11.23	43.38	23.94	10.47	1.000
			15	0.833	5.74	1.34	3.61	8.03	23.15	9.28	10.85	39.79	23.15	9.28	1.000
	20	100	5	0.334	9.36	59.60	3.06	20.25	31.09	149.39	6.17	103.94	35.52	58.65	0.964
			10	0.661	6.85	2.57	3.99	11.75	24.31	19.36	11.11	49.36	24.31	19.36	1.000
			15	0.813	6.72	1.91	3.85	10.07	22.79	9.09	10.61	39.10	22.79	9.09	1.000
		300	5	0.331	7.35	3.85	3.94	14.00	38.92	207.61	10.92	65.75	29.88	47.07	0.998
			10	0.656	6.92	1.95	4.15	10.24	23.76	9.43	11.70	40.06	23.76	9.43	1.000
			15	0.811	6.59	1.70	4.03	9.83	22.09	7.79	11.27	37.76	22.09	7.79	1.000
100	3	500	5	0.400	4.83	1.00	3.45	6.68	12.70	1147.33	−421.72	522.18	112.05	152.26	0.704
			10	0.730	4.78	0.51	3.99	5.67	97.74	1062.97	−337.43	553.26	140.21	151.66	0.822
			15	0.857	4.76	0.45	4.02	5.48	−115.04	5775.52	−431.63	826.53	159.14	165.51	0.850
	10	300	5	0.210	7.28	203.62	6.47	33.52	566.92	11590.13	−499.03	437.67	104.87	107.10	0.730
			10	0.515	12.47	2.71	8.61	17.21	146.98	803.81	37.04	618.67	148.53	144.06	0.963
			15	0.708	12.27	1.86	9.29	15.32	171.16	652.81	52.49	323.34	134.03	87.61	0.986
		500	5	0.208	14.19	9.56	7.50	25.30	50.93	585.19	−537.55	476.49	134.88	135.90	0.798
			10	0.513	12.14	1.94	9.20	15.70	149.21	195.89	52.06	369.10	138.29	123.17	0.986
			15	0.702	12.06	1.67	9.48	14.87	116.97	285.03	54.87	266.20	121.63	71.82	0.988
	20	100	5	0.154	6.60	198.98	−79.76	84.49	143.98	2460.06	−352.15	331.03	100.38	128.59	0.674
			10	0.422	21.29	15.03	10.72	40.31	−20.72	3180.12	−584.39	620.92	140.97	139.94	0.886
			15	0.611	19.25	12.43	12.16	28.53	142.96	308.46	48.94	399.36	133.18	112.54	0.976
		300	5	0.152	29.24	244.50	−67.50	83.55	263.74	4685.48	−468.49	503.90	128.99	149.06	0.784
			10	0.419	18.97	5.43	12.26	28.89	152.83	248.25	51.37	379.00	138.14	114.34	0.980
			15	0.614	18.16	3.26	13.56	23.63	118.22	63.12	62.28	207.09	118.22	63.12	1.000
		500	5	0.151	52.66	572.42	9.40	81.93	102.73	1118.56	−517.67	450.08	124.84	130.80	0.794
			10	0.416	18.39	4.16	12.69	26.38	150.75	409.50	55.81	304.27	126.50	86.26	0.992
			15	0.614	18.00	2.80	13.66	22.72	111.98	43.48	63.04	184.48	111.98	43.48	1.000

[a] h^2 is the heritability in the F_2 population.

[b] P(ok) is the proportion of truncation.

tributed and samples from a gamma distribution

$$f(a) = \frac{\beta^{\beta} a^{\beta-1} e^{-a\beta}}{\Gamma(\beta)} \quad 0 < a < \infty, \qquad 0 < \beta < \infty \qquad (12)$$

with the shape parameter $\beta = 0.5$ which gives $z = (1 + \beta)/\beta = 3$ (see ZHC). (The definitions of m and z are inseparable. Without specification of allelic effect distribution we can not talk about the number of loci for a given amount of genetic variation. In this section m is defined and discussed in reference to the gamma distribution of allelic effects with $z = 3$.) Loci were assigned map positions at random on M chromosomes of length 100 cM. For each replication, map positions and allelic effects were chosen for each locus. The expected difference in means between the two parental populations was calculated and the environmental variance, σ_e^2, was then chosen to give the specified mean difference, $D = (\mu_h - \mu_l)/\sigma_e$, for the parental populations in environmental standard deviations. Parental, F_1 and F_2 populations were simulated each with n individuals with phenotypes assigned by adding a random normal deviate with variance σ_e^2 to the sum of the allelic effects for each genotype. The estimate $\tilde{m}$ was calculated by (1) with correction on the numerator and σ_s^2 estimated using least squares. The modified estimate $\tilde{m}^*$ was calculated with $\hat{\hat{r}}$ given by (5) and $\hat{z} = 3$. Of course, in reality genes are not completely fixed in the appropriate populations and the true parameter value of z is not known.

Table 2 gives the means, standard deviations and 90% confidence intervals for $\tilde{m}$ and $\tilde{m}^*$ as well as the results based on the truncation $0 < \tilde{m}^* < 1000$ for different values of m, M, n and D. The results depend very much on the values of m/M and D as well as n.

These parameters decide the range and locations of most estimate values of $\tilde{m}$, which in turn directly affect estimates of $\tilde{m}^*$. When D is small (and thus the ratio, h^2, of genetic variance over phenotypic variance in F_2 population is small), the interval of $\tilde{m}$ will be wide and will likely to cover the critical region of $1/(1 - 2\hat{\bar{r}})$ which would have drastic effects on $\tilde{m}^*$ and cause many estimates of $\tilde{m}^*$ to have very large positive or negative values. Also when m/M is large, many estimates of $\tilde{m}$ will be relatively high in value and near the critical region, which can cause a significant proportion of estimates of $\tilde{m}^*$ out of the truncation region even when D and n are high. This phenomenon is very troublesome for those organisms with a small number of chromosomes and tight linkage, like *D. melanogaster*.

When m/M is small and D and n are large, however, most estimates of $\tilde{m}$ are in the range of 1 and $1/(1 - 2\hat{\bar{r}})$ and the statistic $\tilde{m}^*$ behaves very well. This gives some hope to the method and suggests that under some very favorable conditions the method proposed in this paper may be informative as a way to estimate the likely magnitude of the number of genes concerned. These conditions are summarized in four parameters m, M, n and D, or more generally by the ratio $\sqrt{n}DM/m$. From simulations, it appears that when $\sqrt{n}DM/m > 15$, the lower bound of 90% interval of $\tilde{m}^*$ is likely to be positive. The 90% interval of $\tilde{m}^*$ is a good indicator of the behavior of $\tilde{m}^*$. When the interval is on the positive side, at least 90% of the estimates of $\tilde{m}^*$ are in the truncated region and the standard deviation of truncated $\tilde{m}^*$ tends to be small. Among the four parameters, m is unknown and is the subject of estimation. However, when M, D and n are all sufficiently large, we may expect that estimates of $\tilde{m}^*$ may be reliable as an indicator of the likely magnitude of m.

The averaged estimates of the standard deviation of $\tilde{m}^*$ by (8) and (9) are generally of the same magnitude as those observed and given in Table 2. Restricted to the truncated region, the averaged estimates of $\sigma_{\tilde{m}^*}$ by (8) and (9) generally underestimate the observed standard deviation of $\tilde{m}^*$. (Some concrete examples will be given in reference to the discussion on the fruit weight of tomato.) However, since an estimate of $\sigma_{\tilde{m}^*}$ strongly depends on the estimated value of $\tilde{m}^*$, a particular estimate of $\sigma_{\tilde{m}^*}$ may not be a good estimate of the standard deviation.

The message of simulation results is clear. Generally, estimates of $\tilde{m}^*$ are unpredictable and have very large sampling variances. However, under some very favorable conditions, the estimates *do* converge to the number of genes under estimation if the parameter z can be estimated reliably and gene effects are additive. These conditions are very restrictive, but not unreachable in experiments.

THE NUMBER OF LOCI WITH SIGNIFICANT EFFECTS

Since the effects of genes are not equal, we have a serious problem in discussing the number of genes, m, affecting a character. The questions often asked are: "What is meant by a locus in this context?" and "Where do we stop counting a locus as one affecting the character?" These questions are related to the parameter z. The value of z can be very high if loci with infinitesimal effects are included in the distribution, and indeed, by virtue of universal pleiotropy, m could include essentially all loci in the genome which differ in two populations. Not all these loci contribute significantly to genetic variation within and between populations, however. Thus instead of discussing the *total* number of loci, m, it might be better and more informative to estimate the number of loci which account for a specified proportion of the differences between populations or genetic variation within populations. We may call this number "the significant number of loci." The meaning of the term will become apparent. Another reason to estimate this number rather than the total number is the lack of information of the parameter value of z for a given data set. Estimation of the total number depends strongly on the value of z. However, as shown in the next section, estimation of the significant number of loci is relatively independent of z.

The reason is that there is a reverse relationship between the value of z and the proportion of loci which accounts for most of the variation. As z increases, the estimated number of loci increases but the proportion of the loci which accounts for most of the variation decreases. As a result, at some point the effects of changing z will be balanced out, which would leave the significant number of loci more or less unchanged.

EXAMPLES

There are numerous reports, in the literature, of estimates of "minimum" or "effective" numbers of loci contributing to the difference in a quantitative character between two populations. Many estimates indicate only a few "minimum" factors. However, estimates from highly diverged populations are usually about 5 to 10, with occasional values up to 20 (WRIGHT 1968; LANDE 1981). To correct the bias on these estimates due to linkage, we have to choose estimates from those organisms for which genetic lengths of chromosomes are known, so that the mean recombination frequency $\bar{r}$ can be estimated (organisms like tomato, maize, mouse and *D. melanogaster*). Two data sets on crosses between different populations or selection lines that differ greatly in quantitative characters are given in Table 3. These examples

TABLE 3

WRIGHT estimates of the number of loci from crosses between widely divergent selection lines or varieties

Populations	n	μ	σ^2	M	$\hat{\bar{r}}$	$\tilde{m}$
Tomato: fruit weight						
P_1	420	−0.137	0.0165			
B_1	932	0.249	0.0339			
F_1	475	0.710	0.0144			
F_2	932	0.653	0.0570			
B_2	931	1.163	0.0344			
P_2	456	1.689	0.0165	12	0.479	10.7 ± 0.5
Maize: percent oil in kernels						
P_1	22	0.513	0.00142			
B_1	68	0.670	0.00169			
F_1	20	0.817	0.00030			
F_2	146	0.803	0.00303			
B_2	74	0.972	0.00169			
P_2	19	1.122	0.00053	10	0.475	21.1 ± 3.1

Sample size (n), means (μ), and variances (σ^2) of the characters in parental and hybrid populations are given with the haploid number of chromosomes (M), estimates ($\hat{\bar{r}}$) of mean recombination frequency, WRIGHT's estimates ($\tilde{m}$) of the number of loci and their estimated standard deviations. Data are from WRIGHT (1968), originally from POWERS (1942) and SPRAGUE and BRIMHALL (1949).

are from the well-known experiments cited by WRIGHT (1968) with the measurements of the data transformed to a scale on which the phenotypic distribution is approximately normal.

Fruit weight of tomato: POWERS (1942) crossed two varieties of tomato that differed 56-fold (about 14 phenotypic standard deviations) in fruit weight. As shown by WRIGHT (1968) and LANDE (1981), the data after transformation give an excellent fit to the additive prediction. The estimate of $\tilde{m}$ is 10.7 ± 0.5 by weighted least squares to utilize all available data, assuming additive gene action. Tomato has 12 haploid chromosomes. Current estimates of genetic lengths of the chromosomes are given in O'BRIEN (1990). Two linkage (classical and restriction fragment length polymorphism) maps are given in pages 6.4 and 6.5 of O'BRIEN (1990). For each chromosome the estimate of chromosome length is based on the longer map, and the estimated genetic lengths for the twelve chromosomes are 211, 163, 123, 103.9, 101.1, 142.2, 89.9, 91.8, 129.2, 134, 98 and 103.7 cM. This gives a mean recombination frequency $\hat{\bar{r}}$ 0.479 by (5). The estimate of the number of loci is then about doubled by correcting the bias due to linkage.

Since z is not known for this data set, different z values were used to estimate the significant number of loci. The estimates of $\tilde{m}^*$ (after being multiplied by a factor of 0.92 for $z = 1 \sim 5$, 0.90 for $z = 10$ and 0.88 for $z = 100$ to correct the bias of the ratio estimate, see APPENDIX A) by using different z values are given in Table 4. Simulations were used to provide confidence intervals of the estimates since the estimated standard deviation is not very informative although it is given. This is particularly worthwhile for this data set which is very well suited to estimation of the number of genes, as the difference in parental populations, sample sizes and the chromosome number are all large, and also there is clear evidence for additive gene action. Assuming that the two parental populations are fixed or nearly fixed, the environmental variance is estimated to be $\hat{\sigma}_e^2 = 0.01535$ by least squares utilizing observations from all the populations. (If the parental populations are not fixed, $\hat{\sigma}_e^2$ contains some genetic variance which is assumed to be the same for both parental populations.) This gives $\hat{D} = 14.74$. In simulations $D = 14.74$ and $m = \tilde{m}^*$ were used as parameter values. Sample sizes are those given in Table 3. For $z = 1$, effects of all loci are the same. The case of $z = 1.57$ is simulated by the half-normal distribution. These two cases are given for reference. For other z values (2, 3, 5, 10 and 100) the allelic effects are assumed to be gamma distributed with $\beta = 1/(z - 1)$. Figure 2 plots these distributions with each scaled to have unit mean. The simulation procedure is the same as before except that the backcross populations are also simulated here. The results based on 1000 replications are given in Table 4.

First it is noted that for fixed D, the simulated estimates and 95% confidence intervals of $\tilde{m}$ are consistent for different z values as expected. The estimated sampling variances of $\tilde{m}$ and $\tilde{m}^*$ are about ¼ to ⅓ of variances found in the simulations, which shows that the estimated sampling variances are underestimates but consistent.

Next, these estimates are used to construct estimates of the significant number, $\tilde{m}_s^*$, of loci. For a given z and the corresponding estimate of m (or the interval of m), m random variables are sampled from the specified distribution, and then ordered. The proportions of the parental difference D ($\propto \sum_{i=1}^{m} a_i$) and the

TABLE 4

Interval estimates of the number of loci by simulations for the fruit weight of tomato

	Estimated $\hat{m}^*$			Simulated $\tilde{m}$				Simulated $\tilde{m}^*$			
z	Mean	SD	m[a]	Mean	SD	95% interval		Mean	SD	95% interval	
$\hat{\hat{r}} = 0.479$											
1.00	17.08	1.56	17	10.30	1.35	7.68	13.16	16.62	4.06	10.06	26.49
1.57	26.31	2.46	26	10.26	1.46	7.48	13.23	25.51	6.94	14.68	41.60
2.00	33.25	3.13	33	10.35	1.55	7.16	13.34	32.84	9.29	17.26	53.74
3.00	49.41	4.69	49	10.41	1.75	6.81	13.65	49.86	16.40	23.56	85.79
5.00	81.74	7.82	82	10.38	1.84	6.84	14.13	82.51	29.28	38.88	153.49
10.00	159.03	15.31	159	10.37	1.89	6.67	14.25	160.99	63.53	72.32	306.19
100.00	1546.87	149.67	1547	10.39	1.96	6.50	14.05	1539.60	579.45	655.81	2815.35
$\hat{\hat{r}} = 0.483$											
1.00	14.91	1.18	15	10.40	1.39	7.77	13.18	14.32	2.95	9.31	20.80
1.57	22.90	1.86	23	10.40	1.46	7.42	13.22	23.02	4.88	13.20	32.33
2.00	28.91	2.37	29	10.54	1.73	7.28	14.07	28.53	7.51	16.14	45.99
3.00	42.90	3.55	43	10.52	1.85	7.13	14.24	42.35	12.09	23.07	70.13
5.00	70.89	5.91	71	10.71	2.08	6.75	14.82	72.55	23.46	35.00	125.82
10.00	137.80	11.57	138	10.65	2.08	6.66	14.52	139.86	45.21	66.33	235.22
100.00	1339.45	113.11	1340	10.68	2.23	6.50	14.93	1343.39	471.45	603.36	2372.97

[a] The number of loci used for simulations.

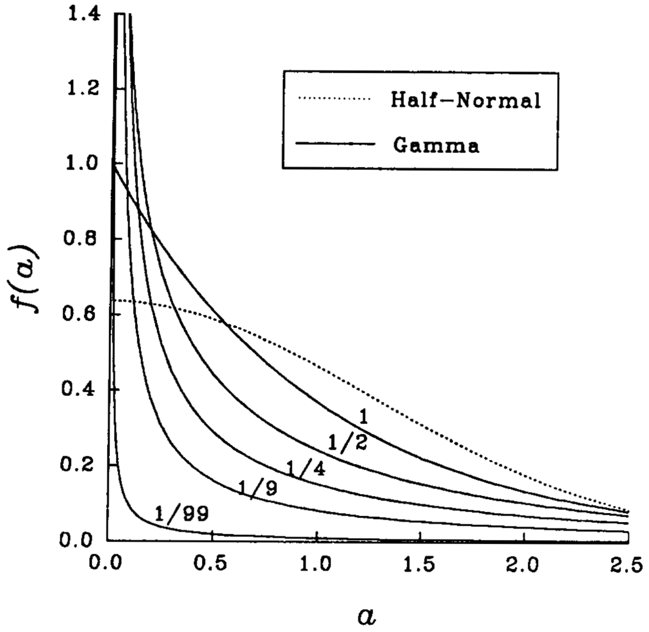

FIGURE 2.—Distributions of allelic effect differences, a, used for simulations. The dotted curve is for the half-normal distribution. The solid curves are for the gamma distribution with $\beta = 1$, ½, ¼, $1/9$ and $1/99$. All distributions are scaled to have unit mean.

genic variance, σ_g^2 ($\propto \sum_{i=1}^{m} a_i^2$), in the F_2 population accounted for by individual ordered variables are calculated. Figure 3 plots the estimates and 95% confidence intervals of $\tilde{m}_s^*$ on D (Figure 3A) and σ_g^2 (Figure 3B) for different parameter values of z. Some of these estimates are listed in Table 5. (These estimates are based on 500 replications. Variations among replications are generally very small.) This of course covers a wide range of distributions. For D, the significant numbers still depend on the distribution of allelic effects unless the specified proportion is small, say 50% of D. For σ_g^2, however, the significant numbers are largely independent of the distribution of allelic effects. This is particularly true for $z \geqslant 2$. More significantly, as z changes from 1 to 100, the estimated m changes from 17 to 1540 but the number of loci accounting for 95% of the genic variance stays more or less at 16!

As indicated in Figure 3, the current method may also be used to estimate the effects of leading loci. Figure 4 plots the estimates and 95% confidence intervals of the effects of the first five leading loci as a proportion of D (Figure 4A) and σ_g^2 (Figure 4B). Taking $2 \leqslant z \leqslant 100$, the effect of the leading locus is estimated between 12.4% and 14.9% of D which is remarkably close. Expressed in terms of the proportion of σ_g^2, the range of estimated effects of the leading locus is larger and varies between 26.7% and 37.7%.

There is, however, still another possible bias on these estimates, the bias due to the estimate of mean recombination frequency, as the genetic lengths of the chromosomes might be underestimated. To examine the possible consequence of this bias on the estimates of significant number of loci, the genetic lengths of the chromosomes are artificially amplified by 50%. This gives $\hat{\hat{r}} = 0.483$. This reduces almost all estimates by about 13% (Tables 4 and 5). Considering the magnitudes of sampling variances involved, the effect of this change is relatively small.

For this data set, the number of loci which account for 95% of the genic variance in the F_2 population is estimated to be 16 with 95% confidence interval (7, 28), and the effect of the leading locus is estimated to

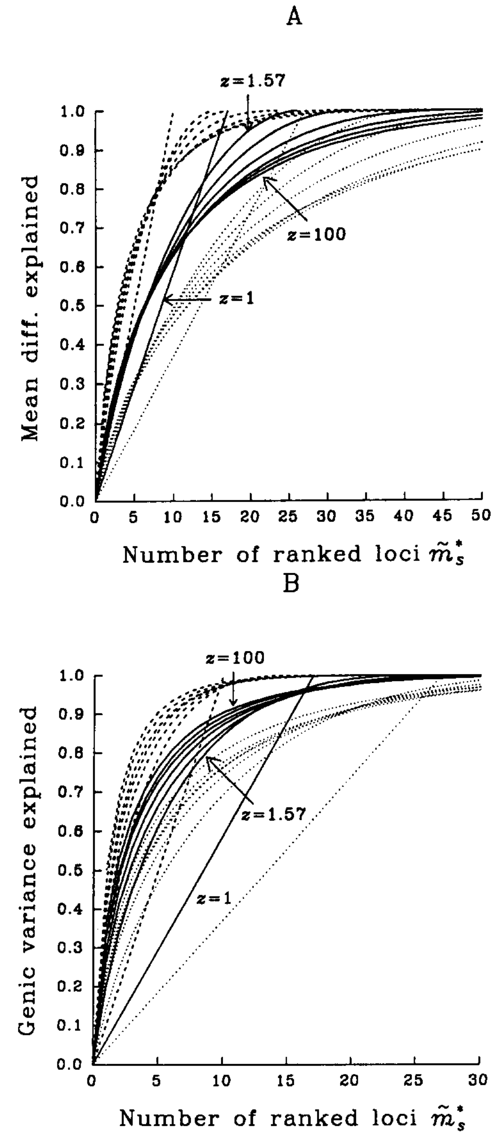

FIGURE 3.—Estimates and their 95% confidence intervals for the number of loci, $\tilde{m}_s^*$, which account for proportions of the parental difference, D, (A) and the genic variance, σ_g^2, in the F_2 population (B) for the fruit weight of tomato. The solid line and curves are the estimates, and the dotted and dashed lines and curves are the corresponding 95% confidence intervals. Different z values are simulated. Three (solid, dotted and dashed) lines are for $z = 1$. Three (solid, dotted and dashed) groups of six curves are for $z = 1.57, 2, 3, 5, 10$ and 100.

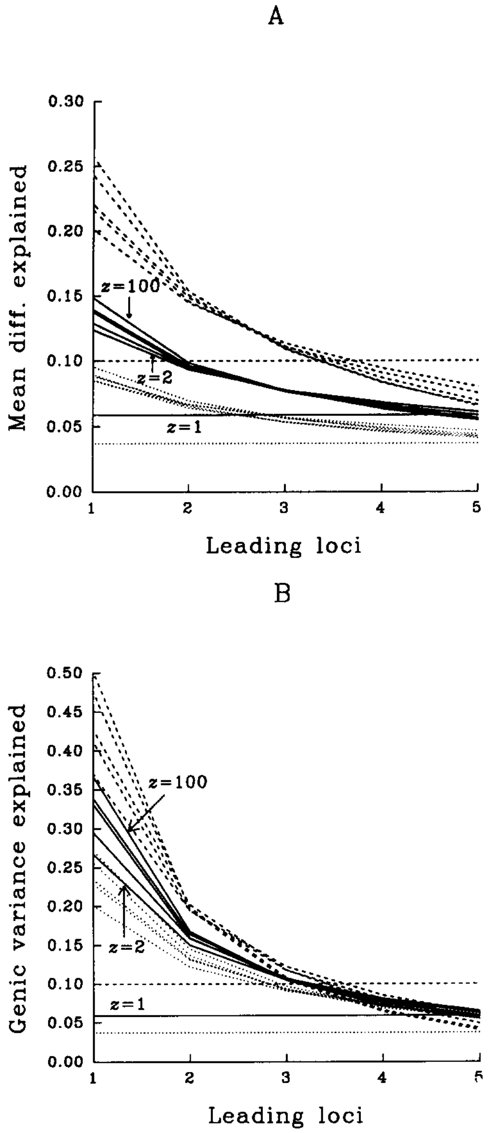

FIGURE 4.—Estimated effects and their 95% confidence intervals of the first five leading loci expressed in terms of proportion of the parental difference, D, (A) and the genic variance, σ_g^2, in the F_2 population (B) for the fruit weight of tomato. Six solid lines are the estimated values for $z = 1, 2, 3, 5, 10$ and 100. Six dashed lines and six dotted lines are for the corresponding 95% confidence intervals. The lines for $z = 1$ are drawn for reference.

be 13% of the parental difference with 95% confidence interval (8.5%, 25.7%). These results tend to be robust.

Oil content of corn: The famous Illinois long-term experiment selecting for high and low oil content in corn seeds was started in the last century and has continued for over nine decades to the present time. After selecting for over four decades, SPRAGUE and BRIMHALL (1949) reported the results of crosses between the high and low selection lines which differed roughly 8-fold (more than 9 phenotypic standard deviations) in mean oil content. The estimates of $\tilde{m}$ from the crosses are about 20 (LANDE 1981). The least squares estimate is 21.1 ± 3.1. Maize has a haploid chromosome number of 10 and the estimated mean recombination frequency $\hat{\bar{r}}$ is 0.475 (O'BRIEN 1990). The estimate 21.1 exceeds the estimated limit $1/(1 - 2\hat{\bar{r}}) = 20$ and is too large to be corrected. As the sample sizes of the experiment are relatively small, this large estimate can be attributed to sampling effect. On the other hand, a large estimate may indicate that the underlying number of loci, m, or the significant number of loci is large. There is indeed evidence to indicate that this is probably the case here. Selection response has continued almost linearly for 90 gener-

TABLE 5

Estimates of significant number, m_s^*, of loci accounting for proportions of D and σ_g^2 for the fruit weight of tomato

		D					σ_g^2				
z	m[a]	50%	75%	90%	95%	99%	50%	75%	90%	95%	99%
$\hat{r} = 0.479$											
1.00	17	9	13	16	17	17	9	13	16	17	17
	(10~27)[b]	(5~14)	(18~21)	(10~25)	(10~26)	(10~27)	(5~14)	(8~21)	(10~25)	(10~26)	(10~27)
1.57	26	7	12	18	20	24	4	8	13	16	20
	(15~42)	(4~11)	(7~20)	(10~28)	(12~32)	(14~38)	(3~6)	(5~12)	(8~20)	(9~24)	(12~32)
2.00	33	7	13	20	24	29	3	7	12	16	23
	(17~54)	(4~11)	(7~22)	(11~33)	(13~39)	(16~48)	(2~5)	(4~11)	(7~19)	(9~25)	(12~37)
3.00	50	7	14	23	29	38	3	6	12	16	25
	(24~86)	(4~11)	(7~24)	(12~39)	(14~48)	(19~64)	(2~4)	(4~10)	(7~19)	(9~26)	(13~41)
5.00	83	7	15	26	33	47	3	6	11	16	26
	(39~154)	(4~12)	(8~28)	(13~47)	(16~61)	(23~87)	(2~4)	(4~10)	(6~20)	(8~28)	(14~46)
10.00	161	7	15	27	36	55	2	6	11	15	26
	(72~306)	(4~12)	(8~28)	(13~51)	(17~68)	(26~103)	(2~4)	(3~9)	(6~19)	(8~27)	(13~46)
100.00	1540	7	16	29	39	64	2	5	10	15	26
	(656~2815)	(3~12)	(7~28)	(13~51)	(17~70)	(28~114)	(1~3)	(3~8)	(5~17)	(7~25)	(13~45)
$\hat{r} = 0.483$											
1.00	14	7	11	13	14	14	7	11	13	14	14
	(9~21)	(5~11)	(7~16)	(9~19)	(9~20)	(9~21)	(5~11)	(7~16)	(9~19)	(9~20)	(9~21)
1.57	23	6	11	16	18	21	4	7	11	14	18
	(13~32)	(4~8)	(7~15)	(9~22)	(11~25)	(12~29)	(2~5)	(4~10)	(7~15)	(8~19)	(11~25)
2.00	29	6	12	18	21	26	3	6	11	14	20
	(16~46)	(4~9)	(7~18)	(10~28)	(12~33)	(15~41)	(2~4)	(4~9)	(7~17)	(9~22)	(12~31)
3.00	42	6	12	20	24	32	3	6	10	14	21
	(23~70)	(4~10)	(7~20)	(11~32)	(14~40)	(18~53)	(2~4)	(4~9)	(6~16)	(8~22)	(13~34)
5.00	73	6	14	23	29	42	2	5	10	14	23
	(35~126)	(3~10)	(7~23)	(12~39)	(15~50)	(21~71)	(2~3)	(3~8)	(6~16)	(8~23)	(12~39)
10.00	140	6	14	24	32	49	2	5	10	14	23
	(66~235)	(3~10)	(7~22)	(12~40)	(16~53)	(24~80)	(2~3)	(3~7)	(5~15)	(7~21)	(12~37)
100.00	1343	6	14	25	34	56	2	5	9	13	23
	(603~2373)	(3~10)	(7~23)	(12~43)	(16~59)	(26~97)	(1~3)	(3~7)	(5~15)	(7~21)	(12~38)

[a] m is the number of loci estimated for the given z value.
[b] Values in brackets are 95% confidence intervals.

ations with the current selection lines, differing almost twice that reported in 1949 (J. W. DUDLEY, personal communication). Previous estimates of the number of genes by different methods ("Student" 1934; DUDLEY 1977) all indicate that the number of loci responsible for the selection response might be very large. It would be an interesting result if the current selection lines are crossed to estimate the number of genes by the present method. Since the selection lines differ widely and the number of chromosomes is not small, this experiment is well suited for estimation of the number of loci provided sample sizes can be made large.

DISCUSSION

Estimation of the number of genes responsible for the difference in quantitative characters between two extreme populations is a long standing problem. Attempts to deduce the genetics of the differences between divergent populations that are crossed, from analyses of F_1, F_2 and backcrosses, have been frustrated by the large number of possible parameters: the number, effects and frequencies of alleles, linkage, degrees of dominance and possible kinds of epistatic effects. Consequently, WRIGHT's method, though simple, provides seriously biased estimates of the number of loci. Unless the bias of the estimates can be reasonably corrected, information from the estimates is very limited.

In this paper, an attempt is made to dissect the effects of different genetic complications (except epistasis) on the estimation of the number of genes. Linkage effects are summarized by the mean recombination frequency, which is estimable, and can be corrected. Unequal effects of alleles are also summarized in a parameter z which measures the variability of allelic effects among loci. Limited data indicate that this parameter may be larger than 2. It is difficult to set an upper bound on this parameter because it is difficult to define precisely what is meant by the number of genes involved. It is helpful and informative to estimate the number of loci which account for most of the genetic variation. Under certain circumstances, estimates of this number tend to be independent of the distribution of allelic effects, showing that the concept has a nice invariant property. Another consequence of this relative invariance (particularly for $z \geqslant 2$) is that estimates of m for a given value of z, say 3, may be statistically equivalent to estimates of m with a different value of z, say 10, for certain applications.

The effects of leading loci can also be estimated by the current method. This is very important and directly related to current efforts of mapping leading quantitative trait loci (QTLs). Knowledge of likely magnitudes of effects of the leading loci gained by applying the current method can help to design mapping experiments and to determine sample sizes needed to find the QTLs.

Dominance effects can also be corrected for if genes are fixed in the appropriate populations. The effects of epistasis are more difficult to handle as they involve too many parameters and it is hard to identify the patterns of interaction in a data set. Scaling the data is a common practice to minimize the effects of possible interactions, but that does not necessarily mean that the interactions can be scaled out.

WRIGHT's estimator has also a serious sampling variance problem. This is particularly so for the proposed modified estimator. Remedies for the problem include choosing only those populations or lines which differ by *many* (say 10) phenotypic standard deviations for estimating the number of genes (or using strong divergent selection to create highly divergent lines); keeping large sample sizes (say >200 for most populations); replicating estimations if possible; and using other better methods to estimate σ_s^2 such as variance component analysis on families. The problem is likely to be more severe for those organisms which have tight linkage. If the linkage effect is a major problem, σ_s^2 may have to be estimated from other sources with linkage disequilibrium significantly reduced (ZHC). Because of these problems, the method is not recommended for general use unless these specified conditions are met.

Finally it should be pointed out that the number of loci discussed in this paper is not the number of loci which are capable of contributing to the genetic variance via mutation. The latter number is relevant to many theoretical models involving mutation. The relationship between the number of loci which contribute to the genetic variance within a current population or the difference between two current populations and the number of loci which are capable of contributing to the genetic variance via mutation depends on the mechanisms which maintain genetic variation within populations and the mechanisms which cause differentiation between populations. In any case, the latter number is substantially larger and should not be confused with the number discussed in this paper.

I am greatly indebted to TRUDY MACKAY for providing the Drosophila data for this study. The manuscript was greatly improved by comments from TRUDY MACKAY, DAVID HOULE, C. CLARK COCKERHAM, BRUCE WEIR, RUSS LANDE and ANDY CLARK. This investigation was supported in part by research grant GM 45344 from the National Institute of General Medical Sciences.

LITERATURE CITED

CARSON, H. L., and R. LANDE, 1984 Inheritance of a secondary sexual character in *Drosophila silvestris*. Proc. Natl. Acad. Sci. USA **81:** 6904–6907.

CASTLE, W. E., 1921 An improved method of estimating the number of genetic factors concerned in cases of blending inheritance. Science **54:** 223.

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

DEMPSTER, E. R., and L. A. SNYDER, 1950 A correction for linkage in the computation of number of gene differences. Science **111:** 283–285.

DUDLEY, J. W., 1977 76 generations of selection for oil and protein percentage in maize, pp. 459–473 in *Proceedings of the International Conference on Quantitative Genetics*, edited by E. POLLAK, O. KEMPTHORNE and T. B. BAILEY, JR. Iowa State University Press, Ames.

FRANKLIN, I. R., 1970 Average recombination frequencies. Genetics **66:** 709–711.

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

MACKAY, T. F. C., R. F. LYMAN and M. S. JACKSON, 1992 Effects of *P* element inserts on quantitative traits in *Drosophila melanogaster*. Genetics **130:** 315–332.

O'BRIEN, S. J., 1990 *Genetic Maps: Locus Maps of Complex Genomes*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

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

SEREBROVSKY, A. S., 1928 An analysis of the inheritance of quantitative transgressive characters. Z. Indukt. Abstammungs. Vererbungsl. **48:** 229–243.

SPRAGUE, G. F., and B. BRIMHALL, 1949 Quantitative inheritance of oil in the corn kernel. Agron. J. **41:** 30–33.

STUART, A., and J. K. ORD, 1987 *Kendall's Advanced Theory of Statistics. Vol. 1. Distribution Theory*, Ed. 5. Oxford University Press, New York.

"Student", 1934 A calculation of the minimum number of genes in Winter's selection experiment. Ann. Eugen. **6:** 77–82.

TURELLI, M., 1984 Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristles. Theor. Popul. Biol. **25:** 138–193.

WRIGHT, S., 1968 *Evolution and the Genetics of Populations. Vol. 1. Genetics and Biometrical Foundations*. University of Chicago Press, Chicago.

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.

Communicating editor: A. G. CLARK

APPENDIX A

The bias of the estimator: The analysis of Equations 2, 3 and 4 is based on taking the expectations on the numerators and denominators of the estimates $\tilde{m}$ and $\tilde{m}^*$ separately. As a result, the ratio of the expectations may be unbiased, but the expectation of the ratio (the estimate $\tilde{m}^*$) is biased. Taking

$$\tilde{m}^* = \frac{2\hat{\bar{r}}\tilde{m} + (\hat{z} - 1)(\hat{m} - 1)}{1 - \tilde{m}(1 - 2\hat{\bar{r}})} = \frac{x}{y}$$

the bias in $\tilde{m}^*$ can be approximated by a Taylor expansion with respect to x and y as

$$\text{Bias}(\tilde{m}^*) = \mathscr{E}(\tilde{m}^* - m)$$

$$\simeq \frac{m\text{Var}(y) - \text{Cov}(xy)}{[\mathscr{E}(y)]^2}$$

$$\simeq \frac{(1-2\hat{\bar{r}})[(m-1)(1-2\hat{\bar{r}})+\hat{z}]\sigma^2_{\tilde{m}} + 4(m-1)(\sigma^2_{\tilde{m}}+\tilde{m}^2)\sigma^2_{\hat{\bar{r}}}}{[1-\tilde{m}(1-2\hat{\bar{r}})]^2}$$

$$\simeq \frac{(1-2\hat{\bar{r}})[(m-1)(1-2\hat{\bar{r}})+\hat{z}]\sigma^2_{\tilde{m}}}{[1-\tilde{m}(1-2\hat{\bar{r}})]^2}$$

as the term involving $\sigma^2_{\hat{\bar{r}}}$ is very small, where $\mathscr{E}$ denotes expectation. This approximation gives a bias generally of about $m/20$. However, since the approximation underestimates the bias, the real bias must be larger than that. Simulation studies indicate that, depending on values of parameters (particularly m), this bias is generally about $m/10$.

APPENDIX B

Sampling variance of $\hat{\bar{r}}$: The mean recombination frequency, $\hat{\bar{r}}$, for m loci is defined as an average of recombination frequencies among $m(m-1)/2$ distinct gene pairs. Let r_{ij} be the recombination frequency between loci i and j, m_k, defined as a random variable, be the number of loci on the kth chromosome with the genetic length c_k, $C = \sum_{k=1}^{M} c_k$ and $m = \sum_{k=1}^{M} m_k$ where M is the number of chromosomes. Then

$$\hat{\bar{r}} = \sum_{k=1}^{M} \sum\sum_{i<j}^{m_k} \frac{2r_{ij}}{m(m-1)} + \frac{1}{2}\left(1 - \sum_{k=1}^{M} \frac{m_k(m_k-1)}{m(m-1)}\right)$$

$$= \frac{1}{2} + \sum_{k=1}^{M} \sum\sum_{i<j}^{m_k} \frac{2(r_{ij} - 1/2)}{m(m-1)}.$$

Under the assumption of uniform distribution of loci in a genome, the expected recombination frequency between two loci located on the kth chromosome is

$$\mathscr{E}_k(r) = \frac{1}{c_k^2} \int_0^{c_k} \int_0^{c_k} \frac{1}{2}[1 - e^{-2(x_2-x_1)}]\, dx_1\, dx_2$$

$$= \frac{1}{c_k^2}\left[c_k^2 - c_k + \frac{1}{2} - \frac{1}{2}e^{-2c_k}\right]$$

and the expected mean recombination frequency among m loci is then

$$\mathscr{E}(\hat{\bar{r}}) = \mathscr{E}\left(\frac{1}{2} + \sum_{k=1}^{M} \frac{m_k(m_k-1)[\mathscr{E}_k(r) - 1/2]}{m(m-1)}\right)$$

$$= \frac{1}{2} - \frac{1}{4C^2}\left[2C + M - \sum_{k=1}^{M} e^{-2c_k}\right]$$

where $\mathscr{E}$ outside the bracket denotes for expectation with respect to m_k's which are multinomially distributed.

The second moment of mean recombination frequency is defined as

$$\mathscr{E}(\hat{\bar{r}}^2) = \mathscr{E}\left\{\frac{1}{2} + \sum_{k=1}^{M} \sum\sum_{i<j}^{m_k} \frac{2(r_{ij} - 1/2)}{m(m-1)}\right\}^2.$$

To derive it, let us first consider the second moment of recombination frequency for a pair of loci on the kth chromosome

$$\mathscr{E}_k(r^2) = \frac{1}{c_k^2} \int_0^{c_k} \int_0^{c_k} \frac{1}{4}[1 - e^{-2(x_2-x_1)}]^2\, dx_1\, dx_2$$

$$= \frac{1}{4c_k^2}\left[c_k^2 - \frac{3}{2}c_k + \frac{7}{8} - \frac{1}{2}e^{-2c_k} + \frac{1}{8}e^{-4c_k}\right].$$

The expected joint recombination frequency for three genes in two pairs on the kth chromosome is defined as

$$\mathscr{E}_k(r_{12}r_{23}) = \frac{1}{c_k^3} \int_0^{c_k} \int_0^{c_k} \int_0^{c_k} \frac{1}{4}[1 - e^{-2(x_2-x_1)}] \cdot [1 - e^{-2(x_3-x_2)}]\, dx_1\, dx_2\, dx_3$$

$$= \frac{3}{16c_k^3}\left[\frac{4}{3}c_k^3 - 4c_k^2 + 6c_k - 4 + 4e^{-2c_k} + 2c_k e^{-2c_k}\right].$$

Under the assumption of no interference, the joint recombination frequency between four genes in two pairs on the same chromosome is however independent, *i.e.*,

$$\mathscr{E}_k(r_{12}r_{34}) = \frac{1}{c_k^4} \int_0^{c_k} \int_0^{c_k} \int_0^{c_k} \int_0^{c_k} \frac{1}{4}[1 - e^{-2(x_2-x_1)}] \cdot [1 - e^{-2(x_4-x_3)}]\, dx_1\, dx_2\, dx_3\, dx_4$$

$$= [\mathscr{E}_k(r)]^2,$$

so is the joint frequency for two gene pairs involving genes on different chromosomes. Thus

$$\mathscr{E}(\hat{\bar{r}}^2) = \mathscr{E}\left\{\frac{1}{4} + \sum_{k=1}^{M} \frac{m_k(m_k-1)}{m(m-1)}\left[\mathscr{E}_k(r) - \frac{1}{2}\right]\right.$$

$$+ \sum_{k=1}^{M} \frac{2m_k(m_k-1)}{m^2(m-1)^2}\left[\mathscr{E}_k(r^2) - \mathscr{E}_k(r) + \frac{1}{4}\right]$$

$$+ \sum_{k=1}^{M} \frac{4m_k(m_k-1)(m_k-2)}{m^2(m-1)^2}\left[\mathscr{E}_k(\{r_{12}r_{23}) - \mathscr{E}_k(r) + \frac{1}{4}\right]$$

$$
\begin{aligned}
&+\sum_{k=1}^{M}\frac{m_k(m_k-1)(m_k-2)(m_k-3)}{m^2(m-1)^2}\left[\mathscr{E}_k(r)-\frac{1}{2}\right]^2\\
&+2\sum_{k<1}^{M}\sum\frac{m_k(m_k-1)m_l(m_l-1)}{m^2(m-1)^2}\left[\mathscr{E}_k(r)-\frac{1}{2}\right]\\
&\left.\cdot\left[\mathscr{E}_l(r)-\frac{1}{2}\right]\right\}\\
&=\frac{1}{4}+\sum_{k=1}^{M}q_k^2\left[\mathscr{E}_k(r)-\frac{1}{2}\right]+\frac{2}{m(m-1)}\\
&\cdot\sum_{k=1}^{M}q_k^2\left[\mathscr{E}_k(r^2)-\mathscr{E}_k(r)+\frac{1}{4}\right]\\
&+\frac{4(m-2)}{m(m-1)}\sum_{k=1}^{M}q_k^3\left[\mathscr{E}_k(r_{12}r_{23})-\mathscr{E}_k(r)+\frac{1}{4}\right]\\
&+\frac{m-2)(m-3)}{m(m-1)}\sum_{k=1}^{M}q_k^4\left[\mathscr{E}_k(r)-\frac{1}{2}\right]^2\\
&+\frac{2(m-2)(m-3)}{m(m-1)}\sum_{k<1}^{M}\sum q_k^2q_l^2\left[\mathscr{E}_k(r)-\frac{1}{2}\right]\left[\mathscr{E}_l(r)-\frac{1}{2}\right]
\end{aligned}
$$

where $q_k = c_k/C$. The variance of $\hat{\bar{r}}$ is

$$
\begin{aligned}
\sigma_{\bar{r}}^2 &= \mathscr{E}(\hat{\bar{r}}^2)-[\mathscr{E}(\hat{\bar{r}})]^2\\
&=\frac{2}{m(m-1)}\sum_{k=1}^{M}\left\{q_k^2[\mathscr{E}_k(r^2)-(\mathscr{E}_k(r))^2]\right.\\
&\quad\left.+[q_k^2-q_k^4]\left[\mathscr{E}_k(r)-\frac{1}{2}\right]^2\right\}\\
&\quad+\frac{4(m-2)}{m(m-1)}\sum_{k=1}^{M}\left\{q_k^3[\mathscr{E}_k(r_{12}r_{23})-(\mathscr{E}_k(r))^2]\right.\\
&\quad\left.+[q_k^3-q_k^4]\left[\mathscr{E}_k(r)-\frac{1}{2}\right]^2\right\}\\
&\quad-\frac{4(2m-3)}{m\,(m-1)}\sum_{k<l}^{M}\sum q_k^2q_l^2\left[\mathscr{E}_k(r)-\frac{1}{2}\right]\left[\mathscr{E}_l(r)-\frac{1}{2}\right].
\end{aligned}
$$

The value of this sampling variance is generally small. Depending on M and C, the values of $\sigma_{\bar{r}}^2$ is on the fourth decimal point for $m = 10$, and on the sixth decimal point for $m = 100$.

APPENDIX C

This appendix lists the phenotypic means and variances of parental, hybrid and backcross populations with dominance and no epistasis. The parental populations are assumed to be in Hardy-Weinberg and linkage equilibrium. The total phenotypic variance in each population is assumed to be the sum of the genetic variances of m loci, plus a noninheritable environmental variance, σ_e^2, supposing that genetic and environmental effects are independent. Thus for the high (P_h) population

$$
\mu_h = \mu + \sum_i p_{ih}a_i[1 + (1 - p_{ih})d_i]
$$

$$
\begin{aligned}
\sigma_h^2 &= 1/2\sum_i p_{ih}(1-p_{ih})a_i^2[(1+(1-2p_{ih})d_i)^2\\
&\quad+2p_{ih}(1-p_{ih})d_i^2]+\sigma_e^2\\
&=\sigma_{gh}^2+\sigma_e^2,
\end{aligned}
$$

where μ is the ground mean. For the low (P_l) population

$$
\mu_l = \mu + \sum_i p_{il}a_i[1 + (1 - p_{il})d_i]
$$

$$
\begin{aligned}
\sigma_l^2 &= 1/2\sum_i p_{il}(1-p_{il})a_i^2[(1+(1-2p_{il})d_i)^2\\
&\quad+2p_{il}(1-p_{il})d_i^2]+\sigma_e^2\\
&=\sigma_{gl}^2+\sigma_e^2.
\end{aligned}
$$

For the F_1 (P_h × P_l) population

$$
\mu_{F_1} = 1/2\mu_h + 1/2\mu_l + 1/2\sum_i (p_{ih}-p_{il})^2a_id_i
$$

$$
\begin{aligned}
\sigma_{F_1}^2 &= 1/2\sigma_{gh}^2 + 1/2\sigma_{gl}^2 + \sum_i (p_{ih}-p_{il})^2(1-p_{ih}-p_{il})a_i^2d_i\\
&\quad\cdot[1+1/2(1-p_{ih}-p_{il})d_i]+\sigma_e^2.
\end{aligned}
$$

For the F_2 (F_1 × F_1) population

$$
\mu_{F_2} = 1/2\mu_h + 1/2\mu_l + 1/4\sum_i (p_{ih}-p_{il})^2a_id_i
$$

$$
\begin{aligned}
\sigma_{F_2}^2 &= 1/2\sigma_{gh}^2+1/2\sigma_{gl}^2+1/8\sum_i(p_{ih}-p_{il})^2a_i^2[1+6(1-p_{ih}-p_{il})d_i\\
&\quad+3(1-p_{ih}-p_{il})^2d_i^2+1/2(p_{ih}-p_{il})^2d_i^2]\\
&\quad+1/8\sum_{i\neq j}\sum(1-2r_{ij})(p_{ih}-p_{il})(p_{jh}-p_{jl})a_ia_j\\
&\quad\cdot\{[1+(1-p_{ih}-p_{il})d_i][1+(1-p_{jh}-p_{jl})d_j]\\
&\quad+1/2(1-2r_{ij})(p_{ih}-p_{il})(p_{jh}-p_{jl})d_id_j\}+\sigma_e^2.
\end{aligned}
$$

For the B_1 (F_1 × P_h) population

$$
\mu_{B_1} = 3/4\mu_h + 1/4\mu_l + 1/4\sum_i (p_{ih}-p_{il})^2a_id_i
$$

$$
\begin{aligned}
\sigma_{B_1}^2 &= 3/4\sigma_{gh}^2+1/4\sigma_{gl}^2+1/16\sum_i(p_{ih}-p_{il})^2a_i^2\{[1+(1-2p_{ih})d_i]^2\\
&\quad+4(1-p_{ih}-p_{il})d_i[2+(1-p_{ih}-p_{il})d_i]\}\\
&\quad+1/16\sum_{i\neq j}\sum(1-2r_{ij})(p_{ih}-p_{il})(p_{jh}-p_{jl})a_ia_j\\
&\quad\cdot[1+(1-2p_{ih})d_i][1+(1-2p_{jh})d_j]+\sigma_e^2.
\end{aligned}
$$

For the B_2 (F_1 × P_l) population

$$
\mu_{B_2} = 1/4\mu_h + 3/4\mu_l + 1/4\sum_i (p_{ih}-p_{il})^2a_id_i
$$

$$
\begin{aligned}
\sigma_{B_2}^2 &= 1/4\sigma_{gh}^2+3/4\sigma_{gl}^2+1/16\sum_i(p_{ih}-p_{il})^2a_i^2\{[1+(1-2p_{il})d_i]^2\\
&\quad+4(1-p_{ih}-p_{il})d_i[2+(1-p_{ih}-p_{il})d_i]\}\\
&\quad+1/16\sum_{i\neq j}\sum(1-2r_{ij})(p_{ih}-p_{il})(p_{jh}-p_{jl})a_ia_j\\
&\quad\cdot[1+(1-2p_{il})d_i][1+(1-2p_{jl})d_j]+\sigma_e^2.
\end{aligned}
$$