How Informative Is Wright's Estimator of the Number of Genes Affecting a Quantitative Character?

Zhao-Bang Zeng, David Houle and C. Clark Cockerham

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203

Manuscript received February 24, 1990
Accepted for publication May 24, 1990

ABSTRACT

S. Wright suggested an estimator, $\tilde{m}$, of the number of loci, m, contributing to the difference in a quantitative character between two differentiated populations, which is calculated from the phenotypic means and variances in the two parental populations and their F_1 and F_2 hybrids. The same method can also be used to estimate m contributing to the genetic variance within a single population, by using divergent selection to create differentiated lines from the base population. In this paper we systematically examine the utility and problems of this technique under the influences of unequal allelic effects and initial allele frequencies, and linkage, which are known to lead $\tilde{m}$ to underestimate m. In addition, we examine the effects of population size and selection intensity during the generations of selection. During selection, the estimator $\tilde{m}$ rapidly approaches its expected value at the selection limit. With reasonable assumptions about unequal allelic effects and initial allele frequencies, the expected value of $\tilde{m}$ without linkage is likely to be on the order of one-third of the number of genes. The estimates suffer most seriously from linkage. The practical maximum expectation of $\tilde{m}$ is just about the number of chromosomes, considerably less than the "recombination index" which has been assumed to be the upper limit. The estimates are also associated with large sampling variances. An estimator of the variance of $\tilde{m}$ derived by R. Lande substantially underestimates the actual variance. Modifications to the method can ameliorate some of the problems. These include using F_3 or later generation variances or the genetic variance in the base population, and replicating the experiments and estimation procedure. However, even in the best of circumstances, information from $\tilde{m}$ is very limited and can be misleading.

The number of genes that contribute variance in a quantitative character has important implications for evolution and for plant and animal breeding. The simplest and least expensive methods for estimating this number involve observing the means and variances of differentiated populations and their hybrids. Several statistical methods for estimating the "effective" or "minimum" number of loci segregating have been proposed (*e.g.*, CASTLE 1921; STUDENT 1934; PANSE 1940; PARK 1977a; JINKS and TOWEY 1976; COMSTOCK and ENFIELD 1981). The original method of WRIGHT (in CASTLE 1921), as elaborated by WRIGHT (1968), is the simplest and most widely used method. WRIGHT's method relates the difference in the means of two inbred lines to the variance of their F_2 and backcross populations. LANDE (1981) pointed out that WRIGHT's method could also be used with outbred populations. He suggested that the same method could be applied to artificially selected lines (high and low) from a single base population.

Since one of the main assumptions of WRIGHT's method is that one parent contains all the increasing alleles and the other parent contains all the decreasing alleles, the use of selected lines directed at making this assumption true is very appealing. Other assumptions in WRIGHT's method are additive gene action, unlinked loci, and equal allelic effects at all loci. Many authors have addressed the effects of relaxing each of these assumptions in turn (SHULL 1921; WRIGHT 1968; FALCONER 1981; LANDE 1981; MATHER and JINKS 1982). The relaxation of several of these assumptions simultaneously has not been explored. Nevertheless, it has long been clear that when these assumptions are violated the method substantially underestimates the true number of loci.

When lines are created by selection it is tempting to assume that the assumption of fixation of increasing alleles in high lines and decreasing alleles in low lines is assured. However, in the small populations of typical artificial selection experiments, genetic drift may have an important effect on fixation probability for loci of small effects. In addition, departures from the other assumptions have an impact on the process of fixation, beyond their impact on the estimation process. In this paper we explore the utility of selection lines for estimating the minimum number of loci as a function of linkage and the distributions of allelic effects and frequencies in the base population. We will, however, keep the assumption of additive allelic effects.

An additional issue is the sampling variance of the estimation process. LANDE (1981) derived an approximate expression for the sampling variance of the

effective number of loci, and pointed out the need for modest sample sizes during estimation. We report here simulations which emphasize the magnitude of sampling variance.

When all of these factors are considered simultaneously it is clear that WRIGHT's method is more apt to be misleading than illuminating.

WRIGHT'S METHOD

WRIGHT's method involves the means, μ_h and μ_l, and variances, σ_h^2 and σ_l^2, of the parents from the high and low populations, respectively, and the variance $\sigma_{F_2}^2$, of their F_2. The estimate, $\tilde{m}$, of the number of loci is given by

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

where $\sigma_s^2 = \sigma_{F_2}^2 - (\sigma_h^2 + \sigma_l^2)/2$ [see LANDE (1981) and WRIGHT (1968) for details]. There are other methods of determining σ_s^2 including the use of backcrosses and F_1 (COCKERHAM 1986) but all provide the same theoretical value in our context.

Equation 1 would give the number of loci, m, by expectation, if the four assumptions mentioned in the introduction hold. Deviations from the assumptions will usually cause an underestimate of the number of loci, so WRIGHT's method is usually said to estimate the effective or minimum number of loci.

DIVERGENT SELECTION

Consider a base population in which m loci each with two alleles are segregating. Alleles are additive within and between loci. For the ith locus the genotypic effects and frequencies in the population are given as

Genotype	A_iA_i	A_ia_i	a_ia_i
Initial Frequency	p_i^2	$2p_iq_i$	q_i^2
Effects	a_i	$a_i/2$	0

We assume that a_i and p_i are independent and are distributed among loci with density function $f(a)$, $0 < a < \infty$, and $\Pr(p)$, respectively.

We consider continuous truncation selection for high and low values of the quantitative character from the base population. Mutation is ignored.

Estimation from the fixed selection lines (at selection limit)

Let us first consider the selection limit without linkage. Let v_{ih} and v_{il} be the indicator variables for the ith locus with the properties that

$$v_{ih} = \begin{cases} 1 & \text{if the allele } A_i \\ & \text{is fixed in the high line} \\ 0 & \text{otherwise} \end{cases}$$

and

$$v_{il} = \begin{cases} 1 & \text{if the allele } A_i \\ & \text{is fixed in the low line} \\ 0 & \text{otherwise.} \end{cases}$$

Then we have $\mathscr{E}(v_{ih}) = \mathscr{E}(v_{ih}^2) = u_{ih}$ and $\mathscr{E}(v_{il}) = \mathscr{E}(v_{il}^2) = u_{il}$, where u_{ih} and u_{il} are the probabilities of the allele A_i being fixed in the high and low lines, respectively, at the selection limit, and $\mathscr{E}$ denotes expectation. The difference between the means of the high and low lines, aside from experimental error, is

$$\mu_h - \mu_l = \sum (v_{ih} - v_{il})a_i \tag{2}$$

where the summation is taken over all loci. Both parental and F_1 populations have no genetic variance and the genetic variance in the F_2 is

$$\sigma_s^2 = \sum (v_{ih} + v_{il} - 2v_{ih}v_{il})a_i^2/8 \tag{3}$$

because the chance of F_1 individuals being heterozygous for the ith locus is $v_{ih}(1 - v_{il}) + v_{il}(1 - v_{ih})$ and the allele frequency is ½ for all loci segregating in the F_2. The expected value of $\tilde{m}$ in this context is then

$$\mathscr{E}(\tilde{m}) = \mathscr{E}\left\{\frac{[\sum (v_{ih} - v_{il})a_i]^2}{\sum (v_{ih} + v_{il} - 2v_{ih}v_{il})a_i^2}\right\}. \tag{4}$$

This expectation is difficult to analyze in general. However, the ratio of expectations

$$\hat{m} = \frac{\mathscr{E}[(\sum (v_{ih} - v_{il})a_i)^2]}{\mathscr{E}[\sum (v_{ih} + v_{il} - 2v_{ih}v_{il})a_i^2]}. \tag{5}$$

is easier to analyze, and it has the property that

$$\mathscr{E}(\tilde{m}) \leq \hat{m} < \mathscr{E}(\tilde{m}) + 1$$

when the allelic effects and initial frequencies are constant among loci (see APPENDIX). As selection increases, $\hat{m}$ converges to $\mathscr{E}(\tilde{m})$. This suggests that $\hat{m}$ is a good approximation of $\mathscr{E}(\tilde{m})$. Assuming that the fixation status of one locus is independent of that at other loci, and that allelic effects as well as initial frequencies are independent among loci,

$$\hat{m} = 1 + (m - 1)\frac{(\mathscr{E}[(u_h - u_l)a])^2}{\mathscr{E}[(u_h + u_l - 2u_hu_l)a^2]} \tag{6}$$

after taking the expectation with respect to v.

In this paper we would like to express results as the proportion of loci detected, *i.e.*, $\mathscr{E}(\tilde{m})/m$ or $\hat{m}/m$. Instead of analyzing $\mathscr{E}(\tilde{m})/m$, we will, however, define the proportion of loci detected as

$$Z = \frac{(\mathscr{E}[(u_h - u_l)a])^2}{\mathscr{E}[(u_h + u_l - 2u_hu_l)a^2]} = d_1^2/d_2, \tag{7}$$

which ranges from 0 to 1. The above analysis indicates that

$$Z \leq \mathscr{E}(\tilde{m})/m \leq \hat{m}/m$$

TABLE 1

Expected proportion, Z, of the loci detected for various values of $S(= Nsa)$ and p from Equation 8

S	$p = 0.5$	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$
1	0.351	0.338	0.299	0.231	0.133
2	0.733	0.711	0.643	0.518	0.314
4	0.962	0.950	0.904	0.795	0.549
6	0.994	0.990	0.971	0.908	0.698
8	0.998	0.997	0.991	0.958	0.797
10	0.999	0.999	0.997	0.981	0.863

for equal allelic effects and initial frequencies. Our simulations, shown in Table 2, suggest that this inequality also holds when these assumptions are relaxed. The difference between Z and $\hat{m}/m$ is at most $1/m$.

The next step is to evaluate Z. The fixation probabilities in the high and low lines are given, by the diffusion approximation, as

$$u_h = \frac{1 - e^{-2Nsap}}{1 - e^{-2Nsa}}$$

and

$$u_l = \frac{1 - e^{2Nsap}}{1 - e^{2Nsa}}$$

(KIMURA 1957), where N is the effective population size, $s \simeq \iota/\sigma$, ι is the standardized selection differential, and σ is the phenotypic standard deviation in the base population. When a and p are constant among loci, it is easy to show that

$$Z = 1 - \frac{e^{-2Nsap} + e^{-2Nsa(1-p)}}{1 + e^{-2Nsa}} \qquad (8)$$

which increases as Nsa increases and decreases as p deviates from 0.5 (see Table 1).

Effects of variation of allelic effects: Variation of a among loci is expected to decrease the proportion of loci detected. But the distribution of a is generally unknown. To analyze the effect of inequality of allelic effects among loci on Z, we assume that a is distributed among loci with the gamma distribution

$$f(a) = \frac{\beta^\beta a^{\beta-1} e^{-a\beta}}{\Gamma(\beta)} \qquad 0 \le a < \infty, \quad 0 < \beta < \infty, \qquad (9)$$

scaled to have a unit mean, which does not influence the following results. This distribution has been used by KIMURA (1979) and HILL (1982). For this distribution $\mathscr{E}(a^2) = 1 + 1/\beta$ and the variance, $V(a)$, is $1/\beta$. Consequently, $V(a)$ decreases as β increases without change in the mean. The parameter β can then be regarded as a measure of the equality of allelic effects at different loci (HILL and RASBASH 1986). When $\beta \to \infty$, the distribution converges to the case of equal allelic effects.

With this distribution

$$d_1 = \int_0^\infty \frac{1 - e^{-2Nsap} + e^{-2Nsa} - e^{-2Nsa(1-p)}}{1 - e^{-2Nsa}} af(a)\, da$$

$$= \sum_{r=0}^{\infty} (G_1(r) + G_1(r+1) - G_1(r+p) - G_1(r+1-p))$$

$$= \mathscr{F}_1(S, \beta, p)$$

and

$$d_2 = \int_0^\infty \frac{1 + 2e^{-2Nsa} + e^{-4Nsa} - e^{-2Nsap} - e^{-2Nsa(1-p)} - e^{-2Nsa(1+p)} - e^{-2Nsa(2-p)}}{(1 - e^{-2Nsa})^2} \cdot a^2 f(a)\, da$$

$$= \frac{\beta+1}{\beta} \sum_{r=1}^{\infty} r(G_2(r-1) + 2G_2(r) + G_2(r+1) - G_2(r-1+p) - G_2(r-p) - G_2(r+p) - G_2(r+1-p))$$

$$= \frac{\beta+1}{\beta} \mathscr{F}_2(S, \beta, p)$$

where $S = Ns[\mathscr{E}(a^2)]^{1/2}$ and

$$G_i(r) = \left(\frac{2Sr}{\sqrt{\beta(\beta+1)}} + 1\right)^{-\beta-i}$$

Thus

$$Z = \frac{\beta \mathscr{F}_1^2(S, \beta, p)}{(\beta+1)\mathscr{F}_2(S, \beta, p)}. \qquad (10)$$

Note that when selection is very strong ($S \to \infty$), $Z = [\mathscr{E}(a)]^2/\mathscr{E}(a^2) = \beta/(1+\beta)$.

Figure 1A plots the curves of Z against S for different β with $p = 0.5$. It is apparent that the variation of a among loci can substantially decrease the proportion of loci detected. With $\beta = 1$, no more than half of the loci can be detected in any case. The limiting values of Z for given β are achieved in most cases at about $S = 8$. The curve with $\beta = 1000$ approximates the curve given by (8).

Effects of variation of initial allele frequencies: The distribution of p among loci depends on the history of the population. For instance, if the base population is from an unselected equilibrium population, the distribution of p is likely to be U-shaped. In this case, given that each locus in the base population has only two alleles segregating, the probability function of allele frequencies is given by

$$\Pr(j) = \frac{(2T)! l_{1,2T-j}}{2j(2T-j)! l_{2,2T}}$$

for $p = j/2T$, $j = 1, 2, \ldots, 2T-1$,

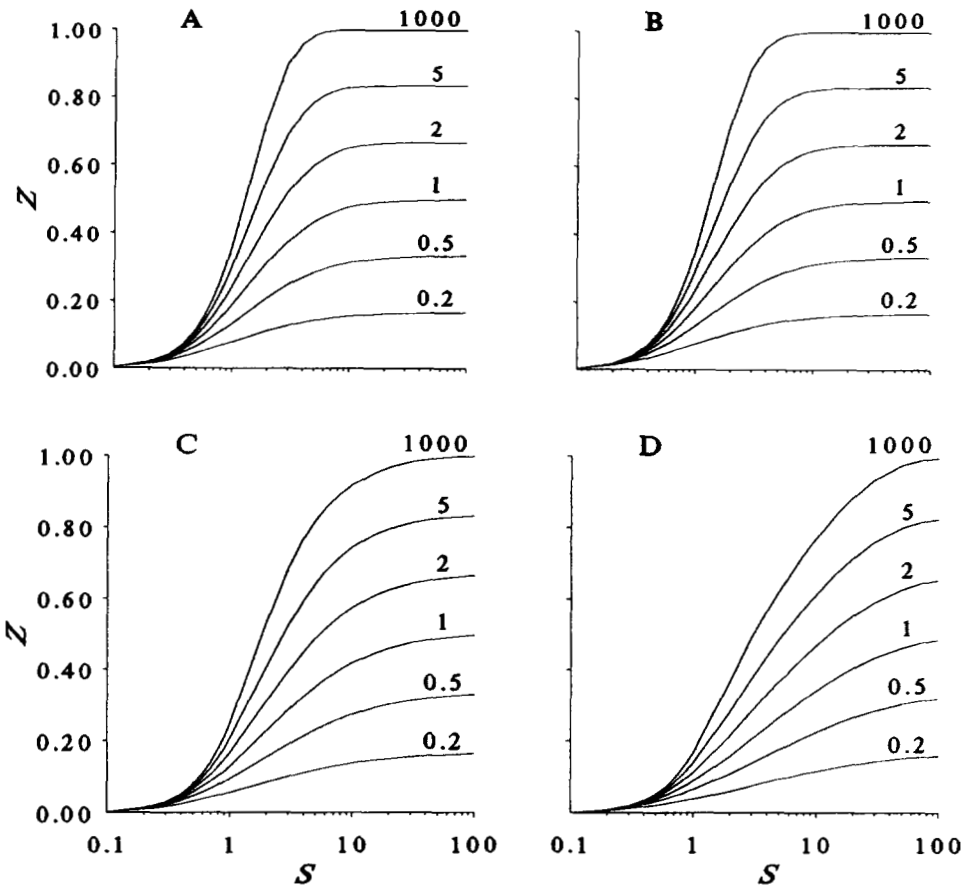

FIGURE 1.—Effects of variation of allelic effect, a, and initial allelic frequency, p, on the expected proportion of loci detected, Z, at the selection limits without linkage. Z is plotted against S for the gamma distribution of a with $\beta = 0.2$, 0.5, 1, 2, 5 and 1000 for p constant (A), or binomially (B), uniformly (C) and neutrally ($T = 20$) (D) distributed among loci with $\bar{p} = 0.5$.

where T is the initial sample size of individuals and $l_{i,2T}$ are Stirling's numbers of the first kind (EWENS 1972). Z is then given by

$$Z = \frac{\beta\{\sum_{j=1}^{2T-1} (\mathcal{F}_1(S, \beta, j/2T)\Pr(j))\}^2}{(\beta + 1)\sum_{j=1}^{2T-1} (\mathcal{F}_2(S, \beta, j/2T)\Pr(j))}. \quad (11)$$

To see the effect of variation of p on Z, we plot the graphs of Z against S in Figure 1 for p constant (A), binomially (B), uniformly (C) or neutrally ($T = 20$) (D) distributed with mean allele frequency $\bar{p} = 0.5$ for the gamma distribution of a. These distributions represent four different modes of variation of p.

When p varies among loci in the base population, there is some further reduction in Z. The decrease is negligible when p's are binomially distributed, modest when p's are uniformly distributed and drastic when p's are neutrally distributed.

Note that when p is neutrally distributed among loci, Z depends also on T. As T increases, Z decreases for given S (Figure 2) because in the neutral case the probability masses are piled up at extreme allele frequencies, notably at $1/(2T)$ and $1 - 1/(2T)$, and, as T increases, the mean fixation probability decreases if S is kept unchanged. However, it should be pointed out that in reality an increase of T will be likely to increase the parameter S, as well as the number of genes in the sample and the probability of multiple alleles at the loci in the sample, and will thus still be likely to increase the number of loci detected. We have not included these complications in our analysis.

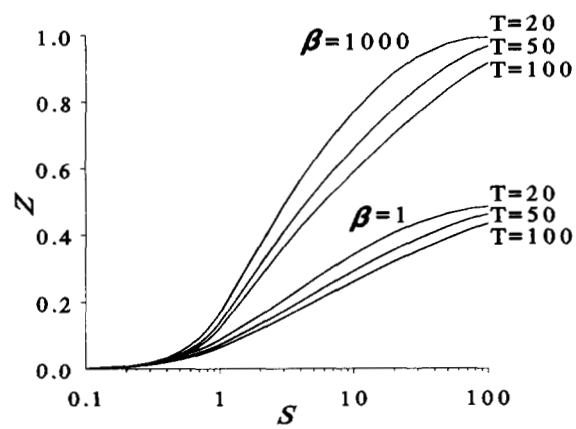

FIGURE 2.—Effect of sample size, T, on the expected proportion of loci detected, Z, for neutral initial allele frequencies and the gamma distribution of a with $\beta = 1$ and 1000.

Estimation from unfixed selection lines (in transient states)

LANDE (1981) showed that WRIGHT's method for the minimum number of loci also applies to parental populations still segregating for the loci of interest.

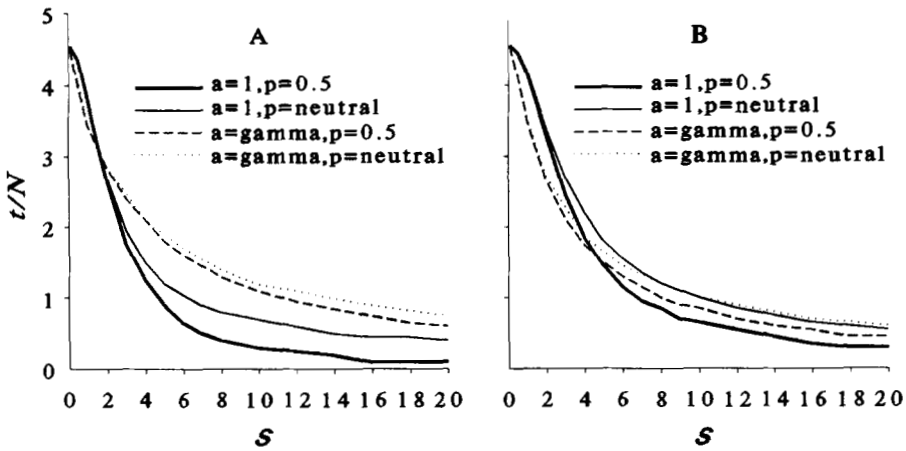

FIGURE 3.—The time required to reach 90% of Z, $t_{0.9}(Z)$, (A) and R, $t_{0.9}(R)$, (B). The allelic effect, a, is either 1 or gamma distributed with $\beta = 1$, and the initial allelic frequency, p, is either 0.5 or neutrally distributed with $T = 20$. Most calculations were done by using $N = 20$. For some large S values $N = 40$ or 80 were used to improve precision.

This raises the question as to how long divergent selection must continue before the estimate is close to that at the selection limit.

Let p_{ht} and p_{lt} be the allele frequencies of a locus in the high and low lines at the tth generation after divergent selection. Equivalent to (7), Z_t in transient states can be expressed as

$$Z_t = \frac{(\mathscr{E}[(p_{ht} - p_{lt})a])^2}{\mathscr{E}[(p_{ht} - p_{lt})^2 a^2]}. \qquad (12)$$

It initially takes the expected value of zero.

When selection is weak ($S \ll 1$), we know that the divergent selection response, $R_t = \mathscr{E}[(p_{ht} - p_{lt})a]$, increases at the rate $(1 - e^{-t/2N})$ and the denominator of (12) (which is the second moment of the selection response distribution) increases at the same rate as well. Consequently, we expect that Z_t will increase at the rate $(1 - e^{-t/2N})$ under weak selection.

As S increases, the time needed to reach the limiting value decreases. This can be evaluated by numerical analysis. We have numerically evaluated (12) by using transition probability matrices. Our method is similar to HILL and RASBASH (1986). The results are presented in Figure 3 in terms of the time to reach 90% of Z, $t_{0.9}(Z)$, at the selection limit, as compared to the same time for R, $t_{0.9}(R)$.

When a is constant, Z approaches its limit more rapidly than R for both p initially 0.5 and neutrally distributed, reflecting the fact that the denominator of (12) does not increase as fast as R. Variation of the initial allele frequencies prolongs both $t_{0.9}(Z)$ and $t_{0.9}(R)$. This is expected since the speed of increase of the mean and variance (related to the second moment) of selection response decreases as the initial allele frequencies deviate from 0.5.

When a varies among loci, $t_{0.9}(R)$ is affected relatively little, but $t_{0.9}(Z)$ increases significantly. When allelic effects are unequal among loci, a large proportion of the initial selection response is due to the alleles with relatively large effects. The change of the mean of the initial response is thus proportional to the magnitudes of these allelic effects, but the change of the second moment is proportional to square of these effects, so the denominator of (12) increases faster than R. When initial allele frequencies also vary among loci, $t_{0.9}(Z)$ increases further.

These results are for unlinked genes. With linkage, the transient behavior of Z is very different.

Linkage effects

Linkage of genes influencing the character also causes $\hat{m}$ to underestimate the actual number of genes. Its effect is two-fold. It reduces the divergent selection response and also at the same time inflates σ_s^2. Linkage inflates σ_s^2 due to the linkage disequilibrium in the F_2 gametes,

$$\sum_{i \neq j}\sum \frac{1}{8}(1 - 2r_{ij})(p_{iht} - p_{ilt})(p_{jht} - p_{jlt})a_i a_j,$$

where r_{ij} is the recombination rate between loci i and j. The reduction in response due to linkage is not very significant for large numbers of chromosomes if genes are randomly located in the genome (ROBERTSON 1970), but the inflation of σ_s^2 can be substantial.

The linkage disequilibrium among genes in the high and low lines brought about by selection also changes σ_s^2 by the amount

$$-\sum_{i \neq j}\sum \frac{1}{4} r_{ij}(2 - r_{ij})(D_{ijh} - D_{ijl})a_i a_j,$$

if σ_s^2 is estimated as $\sigma_{F_2}^2 - \frac{1}{2}(\sigma_h^2 + \sigma_l^2)$, where D_{ijh} and D_{ijl} are the linkage disequilibria between loci i and j in the high and low lines respectively. Our analysis indicates that this part of the effect is always trivial for populations initially in linkage equilibrium. In the following analysis we exclude from σ_s^2 the component due to the linkage disequilibrium in the high and low lines.

Following the same argument leading to (7) and (12), the proportion of loci detected with linkage can

be expressed as

$$Z_t = \frac{\mathscr{E}[(p_{iht} - p_{ilt})(p_{jht} - p_{jlt})a_i a_j]}{\mathscr{E}[(p_{iht} - p_{ilt})^2 a_i^2] + (m-1)\cdot\mathscr{E}[(1-2r_{ij})(p_{iht} - p_{ilt})(p_{jht} - p_{jlt})a_i a_j]}, \quad (i \neq j) \tag{13}$$

which reduces to (12) when genes are unlinked. Note that the expectation is taken over the whole genome. Considering the fact that genes located on different chromosomes are randomly associated, the expectation can also be taken in two parts, within and between chromosomes, with appropriate weights. Suppose that there are M haploid chromosomes in the genome with the kth chromosome having the map length c_k and $\sum_{k=1}^{M} c_k = C$. Assuming the genes to be randomly located in the genome, we expect that there are $m(m-1)\sum_{k=1}^{M} c_k^2/C^2$ within-chromosome pairs and $m(m-1)(1-\sum_{k=1}^{M} c_k^2/C^2)$ between-chromosome pairs among $m(m-1)$ pair joint-product terms [deduced from multinomial distribution]. Then the weight for the within chromosome component is $\sum_{k=1}^{M} c_k^2/C^2$ and for the between-chromosome component is $(1 - \sum_{k=1}^{M} c_k^2/C^2)$.

Equation 13 is a very complex function. Most of the complexity is brought about by the unknown expected allele frequencies with linkage. Before we analyze the behavior of this equation, let us first develop a limiting, but useful, argument about the selection limit.

Limiting argument: At the selection limit with large S, Z can be approximated as

$$Z = \frac{[\mathscr{E}(a)]^2}{\mathscr{E}(a^2) + (m-1)(1-2\bar{r})[\mathscr{E}(a)]^2}$$

by letting $p_{iht} - p_{ilt} \to 1$, where $\bar{r}$ is the average recombination frequency between all pairs of loci. By our assumption of random allocation of genes in the genome, $(1-2\bar{r}) = (2C - M + \sum_{k=1}^{M} e^{-2c_k})/(2C^2)$ for HALDANE's (1919) mapping function $r_{ij} = 0.5(1 - e^{-2d_{ij}})$ where d_{ij} is the map distance between loci i and j (FRANKLIN 1970). The above equation thus suggests that, if $m \gg M\mathscr{E}(a^2)/[\mathscr{E}(a)]^2 = M(1+\beta)/\beta$, $\hat{m}$ ($\simeq Zm$) has an upper bound

$$\frac{1}{1-2\bar{r}} = \frac{2C^2}{2C - M + \sum_{k=1}^{M} e^{-2c_k}}. \tag{14}$$

For example, if each chromosome has the same map length c ($C = Mc$), this bound is $2c^2M/(2c - 1 + e^{-2c})$. For $c = 0.5$, 1 and 1.5 Morgan, it is $1.36M$, $1.76M$ and $2.19M$ respectively. Variation of chromosome map lengths will reduce this bound.

This expected upper bound is substantially less than the number of chromosome segments segregating independently in one generation, i.e., the "recombination index" of DARLINGTON (1937), which equals the haploid number of chromosomes plus the mean number of recombination events per gamete. For instance, *Drosophila melanogaster* has three regular chromosomes (with map length 0.66, 1.08 and 1.06 Morgan, respectively) and a dot chromosome. Taking into account that males have no recombination, our bound is about 3.68, but the "recombination index" for *D. melanogaster* is 9. For maize, the bound is about 18 and the "recombination index" is 36 (LEWIN 1980). However, Equation 14 is deduced by assuming that genes are randomly located in the genome. Nonrandom distribution of genes will generally further lower this bound. In the case of the equal spacing of genes along the chromosomes, the expected upper bound is the same as (14).

Theoretically, $\hat{m}$ can exceed (14) by using the genetic variance in F_3, F_4, or later generation populations to estimate σ_s^2, since the genetic variance in the F_2 population due to the linkage disequilibrium between loci i and j is reduced in each generation by a proportion r_{ij}. Averaged over loci, the linkage disequilibrium in F_t ($t \geq 2$) generation is proportional to

$$\begin{aligned}\mathscr{E}[(1-r)^{t-2}(1-2r)] &= \frac{2}{c}\int_0^c dx_2 \int_0^{x_2} \left[\frac{1}{2} + \frac{1}{2}e^{-2(x_2-x_1)}\right]^{t-2} e^{-2(x_2-x_1)}\, dx_1 \\ &= \frac{1}{c(t-1)}\left[2 - \frac{1}{2^{t-2}}\right] - \frac{1}{c^2 2^{t-2}(t-1)}\sum_{j=1}^{t-1}\binom{t-1}{j}\frac{1-e^{-2jc}}{2j}\end{aligned}$$

for a chromosome with map length c, if genes are assumed to be randomly located (see FRANKLIN 1970). This rate of reduction is not as large as that given by the average recombination frequency between genes within chromosomes; but the linkage disequilibrium can still decrease substantially. There are, however, two disadvantages in using F_3 or later generation variances to estimate σ_s^2, beyond the extra cost of continuing experiments for more generations. First, if the random mating population is not large, the genetic variance is also expected to decrease due to drift. This part of the decrease needs to be corrected in estimating σ_s^2. Second, as the genetic variance decreases, the sampling variance of estimates increases.

Returning to the effect of linkage on Z, we need to assess the expected value of Equation 13 under a variety of assumptions. This was done by simulation.

SIMULATION

Methods

The simulations consisted of three parts: the formation of a base population with the desired parameters; the selection of replicate samples from this population for high and low phenotypic values, which we refer to as the selection process; and obtaining

means and variances for the parental, F_1, and F_2 populations, which we refer to as the estimation process.

The base populations were formed by assuming that m loci were segregating for two alleles. Allelic effects were either constant or drawn from a gamma distribution with $\beta = 1$. Allele frequencies either started at 0.5 or were drawn from EWENS' (1972) neutral allele frequency distribution with $2T = 80$. Loci were either unlinked, or assigned map positions at random on chromosomes of length 100 cM. For pairs of loci on the same chromosome, r was calculated using HALDANE's mapping function. For each replicate selection run, map positions, allelic effects and allele frequencies were chosen for each locus. The expected additive genetic variance was calculated, and the environmental variance, σ_e^2, was then chosen to yield either the desired value of $S = N\iota[\overline{a^2}]^{1/2}/\sigma$ or a heritability, h^2, of 0.25.

To model the selection process, an initial sample of genotypes was drawn from this conceptual base population, assuming gametic-phase equilibrium. Phenotypes were assigned by adding a random normal deviate with variance σ_e^2 to the sum of the allelic effects for each genotype. Truncation selection for high and low phenotypes was then carried out on this initial sample. Gametes were drawn and combined at random (with selfing permitted) from selected parents to yield offspring genotypes. To model the estimation process, gametes were drawn at random from the entire selection population to form n parental genotypes, and then from these parents to form n F_1 and F_2 individuals. The means and variances of the appropriate populations were then calculated.

$\tilde{m}$ was calculated in two ways. In the first case, the genotypic values in the parental, F_1 and F_2 populations were used to calculate $\tilde{m}$. This estimate, called $\tilde{m}_g$, includes the effects of linkage and gametic disequilibrium on the estimates, but ignores the sampling of phenotypes. The second estimate, called $\tilde{m}_p$, was calculated from the sampled phenotypes, as in an actual experiment of this type,

$$\tilde{m}_p = \frac{(\mu_h - \mu_l)^2 - (\sigma_h^2 + \sigma_l^2)/n}{8\sigma_s^2}$$

which is equivalent to equation (1) with a correction for the numerator. The parameters σ_s^2, σ_h^2, and σ_l^2 were estimated using weighted least squares (COCKERHAM 1986).

Results

Comparison of analytical and simulated results: To check the accuracy of our simulations and the analytic results, we compared Z, calculated from (11), with observed $\hat{Z}$ and $\hat{\tilde{m}}_g/m$ from simulations of unlinked loci, where $\hat{\tilde{m}}_g$ is the mean of $\tilde{m}_g$. These comparisons are shown in Table 2. We consider selection to fixation for populations initially segregating at 10 loci. $\hat{Z}$ was obtained by substituting observed average fixation probabilities and allelic effects into equation (7). The observed $\hat{Z}$ agrees well with the expected Z. As expected from comparison of equations (4) and (7), $\hat{Z} \leq \hat{\tilde{m}}_g/m$ in most cases, and the estimates of $\hat{\tilde{m}}_g/m$ exceed $\hat{Z}$ by something less than $1/m$. These results even hold for unequal allelic effects and initial frequencies.

TABLE 2

Comparison of analytical predictions (11) and simulations results for 10 unlinked loci

Distribution of					
Allele frequencies	Allelic effects	S	Z	$\hat{Z}$	$\hat{\tilde{m}}_g/m \pm$ SE
Constant	Constant	1	0.352	0.352	0.400 ± 0.0072
		4	0.962	0.952	0.972 ± 0.0070
		16	0.999	1.000	1.025 ± 0.0098
	Gamma	1	0.185	0.168	0.252 ± 0.0044
		4	0.411	0.405	0.500 ± 0.0069
		16	0.490	0.498	0.576 ± 0.0102
Neutral	Constant	1	0.144	0.145	0.182 ± 0.0043
		4	0.496	0.456	0.469 ± 0.0079
		16	0.770	0.743	0.769 ± 0.0127
	Gamma	1	0.078	0.072	0.133 ± 0.0028
		4	0.211	0.190	0.271 ± 0.0053
		16	0.348	0.348	0.415 ± 0.0101

For a given S value, the total population size and the proportion selected were chosen to yield a reasonable heritability. For S = 1, 4, and 16, the total population size was 10, 20, and 60; the proportion selected was 0.2, 0.5 and 0.5; and simulation replications were 1000, 500 and 200, respectively.

Linkage effects: The effects of linkage were investigated by simulations utilizing 10 loci, which were assumed to be either unlinked, or randomly distributed on one, three or ten chromosomes. Figure 4 shows the ratio $\hat{\tilde{m}}_g/m$ at fixation as a function of S. As expected, linkage substantially reduces $\hat{\tilde{m}}_g/m$. In the extreme case where all ten loci are distributed in one chromosome, linkage essentially dominates the estimation; population size, selection intensity, allelic effects, and allele frequencies all become unimportant.

Figure 5 shows the ratios of observed components of Z, calculated as in Equation 13, with all ten loci on a single chromosome to that when all loci are unlinked. The ratio of components for the selection response ($\overline{[(p_{ih} - p_{il})(p_{jh} - p_{jl})a_i a_j]}$), and for the added F_2 genic variance ($\overline{[(p_{ih} - p_{il})a_i]^2}$) are approximately one, while the denominator of (13), the realized added F_2 genetic variance (including linkage disequilibrium), is substantially increased with linkage. This agrees with ROBERTSON's (1970) conclusion that linkage tends to have small effects on selection response. The reduction in $\hat{\tilde{m}}_g$ from linkage is essentially all due to linkage disequilibrium in the F_1 gametes.

An investigator who wants to estimate $\tilde{m}$ will not

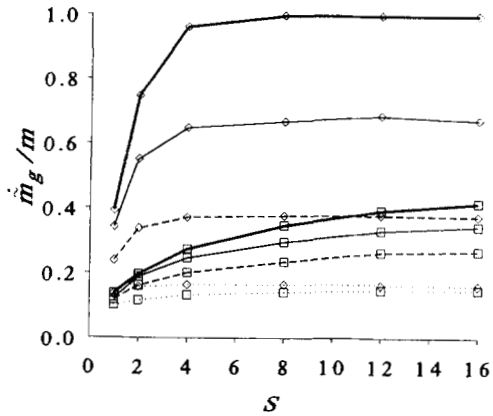

FIGURE 4.—The effect of linkage on $\hat{m}_g/m$ at fixation when m = 10. Thick solid lines denote unlinked loci, thin solid lines loci randomly located on ten chromosomes, dashes three chromosomes, dotted lines one chromosome. Diamonds denote cases where allelic effects are assumed to be equal and initial allele frequency is 0.5. Squares denote cases where allelic effects were drawn from the gamma distribution with $\beta = 1$, and allele frequencies were drawn from the EWENS' distribution ($T = 40$). To achieve a given S value while preserving a reasonable heritability, the total population size during selection and the proportion selected were varied. The proportion selected was 0.2 for $S = 1$ and 0.5 otherwise. For S values 1, 2, 4, 8, 12, and 16, the total population size was chosen to be 10, 10, 20, 30 and 40, respectively. Selection replicate numbers were 300, 250, 200, 100, 100 and 100, respectively.

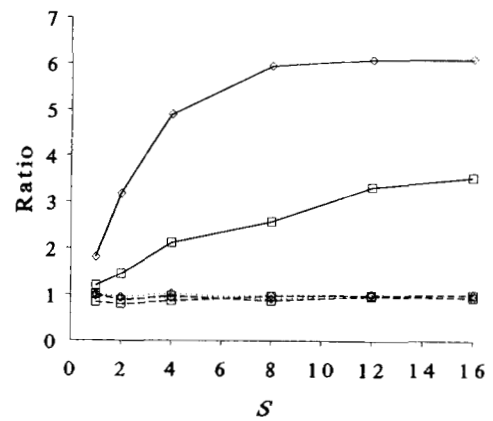

FIGURE 5.—Ratio of components of Z at fixation when ten loci are linked on a single chromosome to that when ten loci are unlinked. Dotted lines denote the numerator of Equation 13, dashed lines $[(p_{ih} - p_{il})a_i]^2$, the genic F_2 variance, and solid lines the denominator of Equation 13. Diamonds indicate equal allele frequencies and effects, and squares variable allele frequencies and effects, as described in the legend of Figure 4. Populations sizes are as in Figure 4. Selection replicate numbers were 1,000, 1,000, 500, 300, 300 and 200 for the six S values graphed.

know the magnitude of S in his experiments because the actual number of loci, their allele frequencies, and allelic effects are unknown. However, the number of chromosomes will usually be known, and heritability may be estimated. Consequently a clearer idea of the utility of this method is gained by examining the effect of changing the actual number of loci, m, on $\tilde{m}$ when chromosome number and heritability are fixed. Figure 6 shows $\hat{m}_g$ as a function of m for three linkage regimes. With linkage, as m increases, $\hat{m}_g$ underestimates m by a greater and greater amount, and becomes nearly constant. Consequently, when $m > M$, $\hat{m}_g$ is quite insensitive to the actual number of loci. Note that $\hat{m}_g$ does not become very close to the upper bound implied by Equation 14, even when there are 20 loci on only three chromosomes. For constant map length of 1 Morgan assumed here, the upper bound is 5.3 for $M = 3$, and 17.6 for $M = 10$. These results, borne out by additional simulations not shown, indicate that $\hat{m}_g \simeq M$ when $m > M$.

With linkage, the time required to obtain 90% of the limiting value of Z, $t_{0.9}(Z)$, is shorter compared with that with random recombination, especially when S is small or m is large. Also, $t_{0.9}(Z)$ is much less than $t_{0.9}(R)$ with linkage (Table 3). This is partly because the expected Z is smaller with linkage and partly because linkage disequilibrium in the F_2 after crosses of high and low lines builds up gradually as selection proceeds and linkage has little effect on selection response.

Sampling variance of $\tilde{m}$: There are three sources of sampling which contribute to the sampling error of estimates of $\tilde{m}$ in our simulations. They are (i) the sampling of the initial population (initial allele frequencies, effects, and map positions); (ii) the variation in chance of fixation of alleles; and (iii) the sampling of phenotypic observations. Table 3 shows the observed standard deviations of estimates of $\tilde{m}$ from simulations. An estimate, $\hat{\sigma}_{\tilde{m}_p}$, of standard deviation of $\tilde{m}$ given by LANDE (1981), which approximates the sampling standard deviation in the process of estimation, is also listed in the Table. Generally the standard deviation of $\tilde{m}_p$, $\sigma_{\tilde{m}_p}$, is large and can be very large when $m < M$, so $\hat{m}_p$, the mean of $\tilde{m}_p$, may be very far from $\hat{m}_g$. This large variance is not primarily due to the sampling of parameters or to the selection process. With samples of the size we have used, the sampling variance due to the sampling of genetic parameters and selection process, $\sigma^2_{\tilde{m}_\infty}$, is usually small and can be ignored. In almost all cases it accounts for less than 1% of the variance. The enormous variance of $\tilde{m}_p$ is largely due to a small number of cases where σ_s^2 is very near 0, leading to estimates of $\tilde{m}_p$ with very large absolute values. Typical distributions of observed σ_s^2 and $\tilde{m}_p$ are plotted in Figure 7. The distribution of σ_s is not significantly different from normal, but overlaps zero. This suggests that simply discarding those few estimates of σ_s with negative or very small values might substantially improve precision. Table 3 also shows summary statistics when all values of $\tilde{m}_p$ which are negative or more than 100 are discarded. This does indeed lower the variance, but also slightly biases the mean upward. As CARSON and LANDE (1981) pointed out, the problem of σ_s overlapping 0 leads LANDE's sampling variance to substantially underestimate the

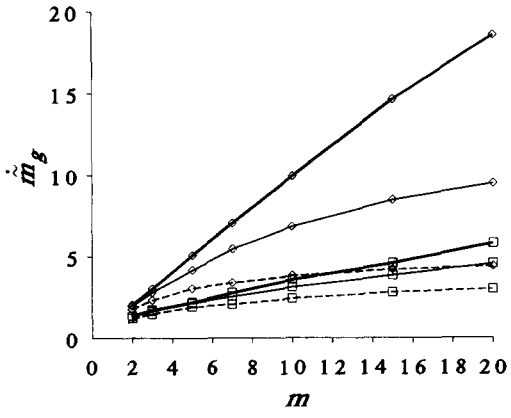

FIGURE 6.—$\hat{\bar{m}}_g$ as a function of m. Lines were selected until fixation or 200 generations had passed. Thick solid lines are for unlinked loci, thin solid lines for ten chromosomes, and dashed lines for three chromosomes. Diamonds denote equal allelic effects and frequencies, and squares variable allelic effects and frequencies. Values are means of 300 replicates.

observed variance. This is true even after the data is censored and when the variance due to sampling of genetic parameters and selection is removed.

Since the variance due to the selection process is relatively small, we investigated the effect of obtaining replicates from the same selected populations, then using averaged estimates of the numerator and denominator to obtain $\tilde{m}_p$. We simultaneously investigated the effects of estimation replicate sample size n and replicate number y on $\sigma_{\tilde{m}_p}$ for a fixed total sample size ny. To do this, we used a slightly different simulation scheme, which does not include the selection process. In these simulations, loci were assumed to have previously been fixed for appropriate alternate alleles. The environmental variance was chosen so that a population which segregated for the allelic effects represented, assuming either equal initial frequencies or neutral initial frequencies, would have a heritability of 0.25. Estimates of σ_s^2 and $(\mu_h - \mu_l)^2$ were calculated by weighted least squares within each replicate, then averaged over y replicates. Figure 8 summarizes the results with $m = 20$. The total sample size shown in the abscissa is n times y. The value of y can then be seen by dividing ny by n, depicted in the figure. For example, for the curve of $n = 32$, y ranges from 1 to 7. The standard deviations in the figure were calculated from at least 200 observations.

For unlinked loci, in part A of Figure 8, $\sigma_{\tilde{m}_p}$ decreases almost linearly with $\ln(ny)$. For a given number of individuals measured, $\sigma_{\tilde{m}_p}$ decreases somewhat if those individuals are analyzed in several replicates. However, there is still substantial variance, even at the largest total sample size, 2048. Unless sample size is approximately 500 or more, the empirical confidence intervals of $\tilde{m}_p$ do not, on the average, exclude 0. The situation is somewhat different for the case where all 20 loci are assumed to fall on three chromosomes, as in part B of Figure 8. Here replication during estimation has less effect. Increasing sample size beyond 512 has little effect. With three chromosomes, the genome behaves like three loci, each with relatively large effects. This reduces both $\tilde{m}_p$ and $\sigma_{\tilde{m}_p}$.

DISCUSSION

Our results suggest that WRIGHT's method is of little value in estimating the number of loci influencing a quantitative character. Linkage effectively prevents the expectation of such estimates from exceeding the number of chromosomes, and sampling variance may easily prevent one from concluding that the number of loci is even that large.

Ironically, our results also show that divergent selection is useful to validate the assumption of WRIGHT's method for estimating the number of loci that high and low valued alleles are fixed in the appropriate populations. However, for the reasons given above, the many estimates of the number of genes based on WRIGHT's method using one way or divergent selection must still be considered suspect.

The problem of linkage may be partially compensated for if we use the variances from F_3 or later generations for estimating σ_s^2, instead of the F_2 variance. The gametic disequilibrium in the F_2 will be reduced by recombination, which allows σ_s^2 to approach the genic variance in the population, and the value of $\tilde{m}$ will increase correspondingly. The value of this approach is counteracted by drift in the hybrid populations. It has been suggested that the additive genetic variance in the base population can be used as an estimate of σ_s^2, which may be free from the influence of gametic disequilibrium (COMSTOCK 1969; PARK 1977a; FALCONER 1981). This requires, however, that the allele frequencies in the base population be known. The only case when this could be true is following a cross between completely inbred lines, which would itself introduce disequilibria in the population.

Sampling variance of the estimate is the second major problem with WRIGHT's method. Part of the problem is that the denominator of Equation 1 can easily be negative, or very small. This is particularly true with short term selection lines which have not diverged enough. This problem will be substantially less if the parental means are many phenotypic standard deviations apart, as is true for the examples in LANDE (1981). When negative estimates occur, it is usually interpreted as a violation of some assumptions, rather than a possible consequence of sampling variance. When such estimates are simply discarded, this biases the remaining sample. Both large sample size and replications of the estimation process are useful in preventing this. With replication, we have shown that averaging the numerator and denominator of

TABLE 3

Effect of phenotypic sampling during estimation

Allelic effects, frequency[a]	M	m	$\hat{\tilde{m}}_g$	$\hat{\tilde{m}}_p$	$\sigma^b_{\tilde{m}_\infty}$	$\sigma_{\hat{\tilde{m}}_g}$	$\sigma_{\hat{\tilde{m}}_p}$	$\hat{\sigma}^c_{\tilde{m}_p}$	Censored $0 < \tilde{m}_p < 100$			Generations to 90% of	
									$\hat{\tilde{m}}_p$	$\sigma_{\hat{\tilde{m}}_p}$	P(OK)[d]	Z	R
Constant	∞	3	3.036	17.575	0.000	0.397	173.233	1.872	5.390	8.464	0.943	3	6
		20	18.607	76.932	1.215	3.044	573.554	4.704	20.806	13.700	0.897	13	17
	10	3	2.805	8.151	0.415	0.569	46.372	1.568	4.745	8.643	0.970	3	6
		20	9.527	10.685	1.256	1.824	41.092	1.404	11.554	7.712	0.993	10	17
	3	3	2.322	1.907	0.507	0.572	66.750	1.084	2.977	2.904	0.957	2	6
		20	4.432	4.688	0.460	0.732	1.576	0.411	4.688	1.380	1.000	6	17
Variable	∞	3	1.721	1.985	0.608	0.645	4.405	0.684	2.093	1.620	0.973	5	5
		20	5.808	5.349	1.601	1.784	107.341	0.909	6.976	5.605	0.973	13	12
	10	3	1.646	1.889	0.532	0.575	33.966	0.608	1.989	1.751	0.987	5	5
		20	4.583	4.627	1.233	1.379	57.503	0.692	6.172	8.053	0.983	11	12
	3	3	1.480	2.490	0.481	0.513	10.270	0.538	1.820	1.667	0.970	4	6
		20	3.018	3.224	0.625	0.753	2.824	0.363	3.357	1.663	0.997	9	12

Three hundred replicates of the selection process for each parameter set were generated. The best 8 out of 40 individuals were chosen during the selection process, which continued until fixation or 200 generations had passed. For estimation samples of 100 individuals in the parental, F_1 and F_2 populations were used. Heritability was assumed to be 0.25 in the base population.

[a] Constant: $p = 0.5$, $a = 1$; Variable: p from EWENS' distribution, and a gamma distributed.

[b] Standard deviation of $\tilde{m}$ assuming infinite sample size during the estimation process.

[c] Expected standard deviation of $\tilde{m}_p$ from LANDE's approximate formula.

[d] Proportion of replicates where $0 < \tilde{m}_p < 100$.

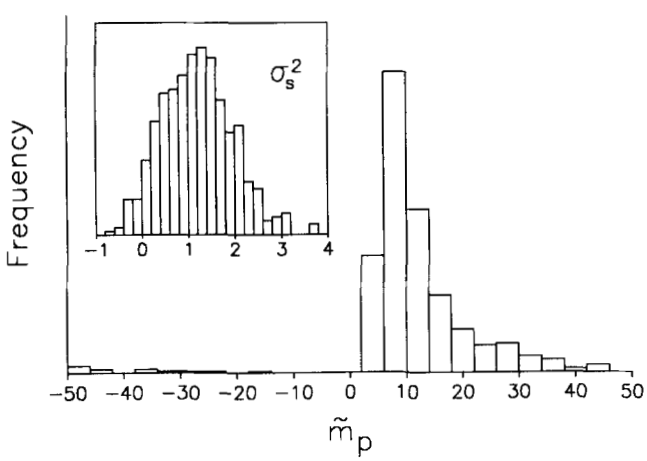

FIGURE 7.—Typical distributions of observed σ_s^2 and $\tilde{m}_p$ from simulations. Allelic effects and frequencies were assumed to be equal initially. The total population size during selection is 40. The proportion selected is 0.2. $m = 10$. $n = 100$. Simulations were run until fixation or 200 generations passed. Of the 500 replicates obtained, 12 yielded values of $\tilde{m}_p < -50$ and 32 estimates > 50.

Equation 1 over replicates decreases the variance of $\tilde{m}$. However, one of the appealing features of WRIGHT's method is that it can be estimated from data commonly collected. If experimentalists are willing to spend considerable efforts attempting to estimate the number of genes, we feel that it would be best to investigate alternative methods using genetic markers (see below), rather than doing large replicated crosses of the sort we have investigated.

LANDE (1981) has given an approximate sampling variance for estimates of $\tilde{m}$ which takes into account only the variation stemming from the process of estimation, assuming that σ_s^2 is greater than 0. With divergent selection, the variation in selection response also contributes to the sampling variance of the estimate. The small magnitude of the selection replicate standard error presented in Table 3 suggests that this additional source of error is not large for reasonable parameter combinations. Even allowing for this, the values in Table 3 make it clear that LANDE's sampling variance substantially underestimates the actual variance. The large discrepancies in Table 3 are due to estimates of very large absolute value when σ_s^2 is near 0. However, approximately 90% of our estimates do lie within two of LANDE's standard errors of the expected value of $\tilde{m}$ (results not shown), which is in agreement with CARSON and LANDE's (1984) analysis based on bootstrapping resampling. So $\hat{\sigma}_{\tilde{m}_p}$ does have some utility in assessing the reliability of $\tilde{m}_p$.

One positive result from our analysis is that selection can rapidly generate populations whose expected $\tilde{m}$ are close to those at fixation. The expected value of $\tilde{m}$ initially approaches its limit more rapidly than selection response does. The half-life of selection response for typical selection experiments is about $0.2N$ to $0.4N$ generations (FALCONER 1981). For $\tilde{m}$ the half-life is usually less than $0.2N$ generations for $S > 4$. The 90% life for $\tilde{m}$ is usually less than $1.5N$ generations, especially in the presence of linkage (Table 3). At any rate, for many selection experiments with

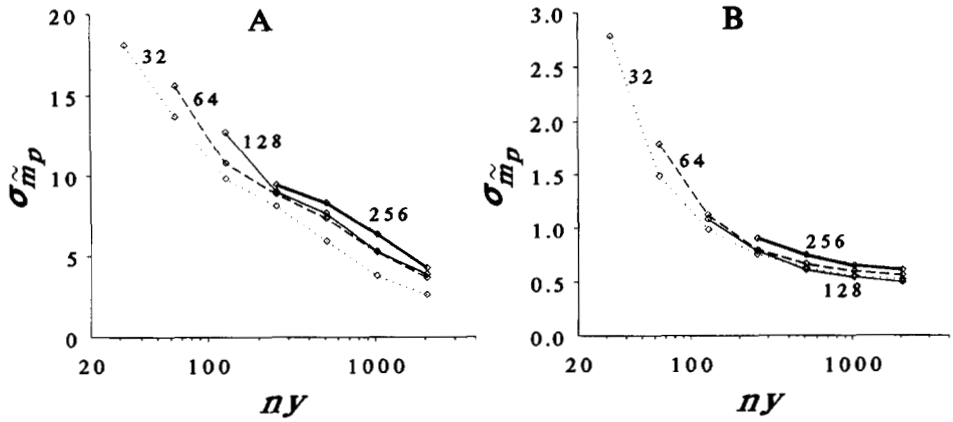

FIGURE 8.—Standard deviation of $\tilde{m}_p$, censored to exclude values of $\tilde{m}_p < -100$ or >100, for $m = 20$, as a function of replicate size, n, and estimation replicate number, y. Allele frequencies and effects are assumed equal. Part A is for unlinkled loci, and part B is for three chromosomes. Lines are labeled with values of n. See text for further explanation.

modest effective population sizes, less than 10–20 generations of selection are enough to achieve most of the limiting estimate of $\tilde{m}$. Selection for longer than this does not provide much gain based on the original genetic variation. Longer selection will tend to include the variation due to new mutations (HILL 1982; ENFIELD and BRASKERUD 1989).

We have ignored two common features of genetic systems in this paper. The first of these is dominance, and the second is natural selection. Both would tend to prolong the time to approach limiting estimates of $\tilde{m}$ as alleles with deleterious effects tend to be recessive. Dominance may either increase or decrease the expectation of $\tilde{m}$ (MATHER and JINKS 1982), while countervailing natural selection will usually decrease it by preventing fixation of deleterious alleles. However, both of these effects would tend to be overshadowed by linkage if the number of loci is large.

In addition to WRIGHT's method, there are also many other statistical methods of gene number estimation, in particular those of PANSE (1940), JINKS and TOWEY (1976) and COMSTOCK and ENFIELD (1981). Some statistical methods are just variations of WRIGHT's method, sharing some of its properties and problems (*e.g.*, STUDENT 1934; COMSTOCK 1969; PARK 1977a,b; FALCONER 1981). Compared with WRIGHT's method, PANSE's method, which uses the variance of the F_3 generation, does not have much advantage, and is more sensitive to changes in the parameters, and more laborious to estimate (MATHER and JINKS 1982). COMSTOCK and ENFIELD (1981) proposed a method to estimate gene number with multiplicative gene effects, which also uses divergent selection and subsequent crosses. They did not analyze the effect of linkage on the estimate. This effect must be large since the genetic variance in the base population, which results from a cross between two inbred lines, is used, and the biases caused by deviations from the assumptions are unknown. MAYO and HOPKINS (1985) showed that this method has the problem that it is very sensitive to small changes in the parameters. JINKS and TOWEY's method involves the determination of the proportion of individuals in the F_t generation of a cross between two inbred lines that are heterozygous at one locus, at least, by an assay of their F_{t+2} grandprogeny families. They assume that progeny from the F_2 generation on are obtained by selfing. As t increases, the estimate increases, because of recombination. JINKS and TOWEY's method tends to give larger estimates than WRIGHT's method in their experiments, but is still a kind of minimum as it is under the influence of linkage (HILL and AVERY 1978). The estimate has the problems that it is much susceptible to selection for heterozygotes or against deleterious recessives at linked genes (HILL and AVERY 1978). It is also very sensitive to unequal allelic effects and becomes very inaccurate for large gene number (MAYO 1987).

While animal and plant breeders need to be most concerned with loci which currently segregate in populations, the goal of estimating gene number in an evolutionary context is yet one step more complex. In the long term, it is the number of loci capable of expressing the type of variation studied which is of interest, rather than just the number which currently do. Even with a good estimate of the number of loci segregating in a population, it is necessary to make assumptions about the mechanisms maintaining that variation, including such factors as effective population size, in order to estimate the number of loci capable of influencing the character.

Estimation of gene number is a long standing problem. Despite its obvious importance in animal and plant breeding and in understanding evolutionary processes, we are still painfully short of a reliable method to do it. This is discouraging for all of us who depend on such estimates to design and perform experiments or build models. There may be no practical alternative to the enumeration of loci using expensive, labor-intensive techniques that utilize genetic markers (SAX 1923; THODAY 1961; TANKSLEY, MEDINA-FILHO and RICK 1982; EDWARDS, STUBER and WENDEL 1987; PATERSON *et al.* 1988; LANDER and BOTSTEIN 1989). In particular, the quantity of molecular markers which we can now imagine locating makes efforts of this kind conceivable. However, since the technique can only locate genes segregating for alleles with relatively large effects, we must have some knowledge

of distributions of allelic effects before we can estimate the number of genes.

We thank TRUDY MACKAY and two reviewers for their comments, and J. POOLE for assistance in making figures. This investigation was supported in part by research grant GM 11546 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.

COMSTOCK, R. E., 1969 Number of genes affecting growth in mice, pp. 137–148 in *Genetic Lectures*. Oregon State University, Corvallis.

COMSTOCK, R. E., and F. D. ENFIELD, 1981 Gene number estimation when multiplicative genetic effects are assumed—growth in flour beetles and mice. Theor. Appl. Genet. **59:** 373–379.

DARLINGTON, C. D., 1937 The biology of crossing-over. Nature **140:** 759–761.

EDWARDS, M. D., C. W. STUBER and J. F. WENDEL, 1987 Molecular-marker-facilitated investigations of quantitative trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics **116:** 113–125.

ENFIELD, F. D., and O. BRASKERUD, 1989 Mutational variance for pupa weight in *Tribolium castaneum*. Theor. Appl. Genet. **77:** 416–420.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

FALCONER, D. S., 1981 *Introduction to Quantitative Genetics*, 2nd Ed. Longman, London.

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

HALDANE, J. B. S., 1919 The combination of linkage values, and the calculation of distance between loci of linked factors. J. Genet. **8:** 299–309.

HILL, W. G., 1982 Predictions of response to artificial selection from new mutations. Genet. Res. **40:** 255–278.

HILL, W. G., and P. J. AVERY, 1978 On estimating the number of genes by genotype assay. Heredity **40:** 397–403.

HILL, W. G., and J. RASBASH, 1986 Models of long term artificial selection in finite population. Genet. Res. **48:** 41–50.

JINKS, J. L., and P. M. TOWEY, 1976 Estimating the number of genes in a polygenic system by genotype assay. Heredity **37:** 69–81.

KIMURA, M., 1957 Some problems of stochastic processes in genetics. Ann. Math. Stat. **28:** 882–901.

KIMURA, M., 1979 Model of effectively neutral mutations in which selective constraint is incorporated. Proc. Natl. Acad. Sci. USA **76:** 3440–3444.

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

LANDER, E. S., and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

LEWIN, B., 1980 *Gene Expression. Vol. 2, Eucaryotic Chromosomes*. John Wiley & Sons, New York.

MATHER, K., and J. L. JINKS, 1982 *Biometrical Genetics*, 3rd Ed. Chapman & Hall, London.

MAYO, O., 1987 *The Theory of Plant Breeding*, 2nd Ed. Clarendon, Oxford.

MAYO, O., and A. M. HOPKINS, 1985 Problems in estimating the minimum number of genes contributing to quantitative variation. Biom. J. **2:** 181–187.

PANSE, V. G., 1940 Application of genetics to plant breeding. II. The inheritance of quantitative characters and plant breeding. J. Genet. **40:** 283–302.

PARK, Y. C., 1977a Theory for the number of genes affecting quantitative characters. I. Estimation of and variance of the estimates of gene number for quantitative traits controlled by additive genes having equal effect. Theor. Appl. Genet. **50:** 153–161.

PARK, Y. C., 1977b Theory for the number of genes affecting quantitative characters. II. Biases from drift, dominance, inequality of gene effects, linkage disequilibrium and epistasis. Theor. Appl. Genet. **50:** 163–172.

PATERSON, A. H., E. S. LANDER, J. D. HEWITT, S. PETERSON, S. E. LINCOLN and S. D. TANKSLEY, 1988 Resolution of quantitative traits into Mendelian factors by using a complete RFLP linkage map. Nature **335:** 721–726.

ROBERTSON, A., 1970 A theory of limits in artificial selection with many selected loci, pp. 246–288 in *Mathematical Topics in Population Genetics*, Vol. 1, edited by K. KOJIMA. Springer-Verlag, Berlin,

SAX, K., 1923 The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics **8:** 552–560.

SHULL, G. H., 1921 Estimating the number of genetic factors concerned in blending inheritance. Am. Nat. **55:** 556–564.

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

TANKSLEY, S. D., H. MEDINA-FILHO and C. M. RICK, 1982 Use of naturally-occurring enzyme variation to detect and map genes controlling quantitative traits in an interspecific backcross of tomato. Heredity **49:** 11–25.

THODAY, J. M., 1961 Location of polygenes. Nature **191:** 368–370.

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

Communicating editor: A. G. CLARK

APPENDIX

C. J. JIANG

Let the allelic effect, a, and initial frequency, p, be constant over loci. At the selection limit, the outcome of fixation for the loci can be summarized as follows:

Class	Low line		High line		No. of loci	Probability
1	aa	$1-u_l$	AA	u_h	x	$P=u_h(1-u_l)$
2	AA	u_l	aa	$1-u_h$	y	$Q=(1-u_h)u_l$
3	AA	u_l	AA	u_h	$m-x-y$	$1-P-Q$
4	aa	$1-u_l$	aa	$1-u_h$		

where u_h and u_l are the probabilities of an allele with effect a being fixed in the high and low lines respectively. Thus $u_h - u_l = (x - y)a$, $\sigma_s^2 = (x + y)a^2/8$ and

$$\tilde{m} = \frac{(x - y)^2}{x + y}.$$

By the assumption that the fixation status of one locus is independent of that at other loci, x and y are trinomially distributed with the probability

$$\Pr(x, y \mid m) = \frac{m!}{x!y!(m - x - y)!} P^x Q^y (1 - P - Q)^{m-x-y}.$$

Since

$$\mathscr{E}\frac{(x - y)^2}{x + y} = \mathscr{E}(x + y) - 4\mathscr{E}\frac{xy}{x + y}$$

$$= m(P + Q) - 4\mathscr{E}\frac{xy}{x + y},$$

we need only evaluate the last term.

$$\mathscr{E}\frac{xy}{x + y}$$

$$= \sum_{x,y=0}^{m} \frac{xy}{x + y} \frac{m!}{x!y!(m - x - y)!} P^x Q^y (1 - P - Q)^{m-x-y}$$

$$= PQ \sum_{x,y=1}^{m} \frac{1}{x + y} \frac{m!}{(x - 1)!(y - 1)!(m - x - y)!} P^{x-1} Q^{y-1} (1 - P - Q)^{m-x-y}.$$

Letting $w = (x - 1) + (y - 1)$ and $P_1 = P + Q$ and translating the trinomial distribution of x and y to binomial distribution of w give

$$\mathscr{E}\frac{xy}{x + y}$$

$$= PQ \sum_{w=0}^{m-2} \frac{1}{w + 2} \frac{m}{w!(m - w - 2)!} P_1^w (1 - P_1)^{m-w-2}$$

$$= \frac{PQ}{P_1^2} \sum_{w=0}^{m-2} \frac{(w + 1)m!}{(w + 2)!(m - w - 2)!} P_1^{w+2} (1 - P_1)^{m-w-2}$$

$$= \frac{PQ}{P_1^2} \left\{ \sum_{w+2=0}^{m} \frac{(w + 2 - 1)m!}{(w + 2)!(m - w - 2)!} P_1^{w+2} (1 - P_1)^{m-w-2} - (0 - 1)\frac{m!}{0!m!} P_1^0 (1 - P_1)^m - (1 - 1)\frac{m!}{1!(m - 1)!} P_1 (1 - P_1)^{m-1} \right\}$$

$$= \frac{PQ}{P_1^2} (mP_1 - 1 + (1 - P_1)^m)$$

$$= \frac{mPQ}{P + Q} - \frac{PQ}{(P + Q)^2} (1 - (1 - P - Q)^m).$$

Thus

$$\mathscr{E}(\tilde{m}) = \frac{m(P - Q)^2}{P + Q} + \frac{4PQ}{(P + Q)^2}[1 - (1 - P - Q)^m].$$

The ratio of expectations is

$$\hat{m} = \frac{\mathscr{E}(x - y)^2}{\mathscr{E}(x + y)}$$

$$= \frac{m(P - Q)^2}{P + Q} + 1 - \frac{(P - Q)^2}{P + Q}.$$

Since $p(1 - p) \leq P \leq 1$ and $0 \leq Q \leq p(1 - p)$, it is easy to show that

$$\mathscr{E}(\tilde{m}) \leq \hat{m} < \mathscr{E}(\tilde{m}) + 1.$$