Theor Appl Genet (1993) 86:964–974

THEORETICAL AND APPLIED GENETICS

The effect of linkage on sample size determination for multiple trait selection

S. J. Schwager[1,] *, M. A. Mutschler[1], W. T. Federer[1], B. T. Scully[2]

[1] Department of Plant Breeding and Biometry, Cornell University, Ithaca, NY 14853, USA
[2] Everglades Research and Education Center, IFAS, University of Florida, Belleglade, FL 33430, USA

Received: 6 May 1992 / Accepted: 28 January 1993

Abstract. Sufficient sample sizes are needed in breeding programs to be confident, with a specified probability α, of obtaining a specified number of plants of a desired genotype in segregating populations. We develop a method of determining the minimum sample size needed to produce, with specified probability α, at least m individuals of a desired genotype. This method takes into consideration factors affecting differential selection of gametes, segregation at a single locus, and linkage among the loci of interest. We first consider the effects in the gametophyte (haploid level) of fitness and linkage on the frequencies of alleles at two linked loci, then at three or more linked loci. The probability of obtaining at least m successes, or occurrences of the desired allele, among n gametes is given by a formula based on the binomial distribution. This probability is affected by fitness and linkage through their impact on the probability that a single randomly chosen gamete is of the desired type. Using an extension of this approach, we examine the effects of the altered allelic frequencies on the likelihood of obtaining the desired genotype from a randomly chosen pair of gametes in the sporophyte (diploid level). A table and a figure show the sample size required to produce, with probability 0.95, m individuals of the desired genotype or phenotype, as a function of m and the probability that a randomly selected individual is of the desired type.

Key words: Allelic frequency – Binomial – Genotypic frequency – Minimum sample size – Probability of desired genotype

Introduction

The identification and transfer of traits controlled by one or a few genes is a common objective of many plant breeding programs. The traits of interest may involve resistance to disease or pests, or an alteration of plant type affecting the horticultural value of the cultivar and its produce.

Sufficient sample sizes are needed in breeding programs for a breeder to be confident of obtaining an adequate number of plants of the desired phenotypes in segregating populations. Excessive sample sizes are wasteful of labor, facilities, and funds, while inadequate sample sizes increase the risk of failure to unacceptable levels. Efficient use of facilities and a high assurance of success require that the breeder considers the number of individuals or lines that must be screened to obtain a desired phenotype. Hanson (1959) described a method for determining the minimum sample size needed to acquire with specified probability at least one individual of a desired genotype, assuming that the genotype is controlled in a simple Mendelian fashion. Since this method assumes that the traits being studied are controlled by one or a few genes, it also assumes that (1) the genes involved are unlinked, (2) the alleles of these genes have no influence on the survival of the genotype, and (3) the alleles at each locus are not subject to selection. If the alleles at the loci of interest are linked in coupling, have a positive effect on survival of the genotype, or are subject to positive selection, the re-

Communicated by A. L. Kahler
* BU-1031-MC in the Technical Report Series of the Biometrics Unit, Cornell University, Ithaca, New York 14853
Correspondence to: S. J. Schwager

quired sample size is decreased. If the alleles are linked in repulsion, or have a negative effect on survival of the genotype, the required sample size is increased.

It is often advantageous in breeding programs to acquire several individuals of a desired genotype, since additional traits of interest will usually be selected in later generations. Sedcole (1977) and Scully and Federer (1993) described a method of determining the minimum sample size needed to acquire with specified probability m plants of a desired genotype, assuming that the loci controlling the desired phenotypes are inherited in simple Mendelian fashion. This method allows the breeder to specify the number of plants of a desired genotype needed. However, this method is also limited, because it is based on the same assumptions as the Hanson method, except that the Scully and Federer method assumes that each trait considered within the desired phenotype is controlled by a single gene (monogenic inheritance).

The assumptions upon which the Hanson, Sedcole, and Scully and Federer methods are based limit their usefulness, because these assumptions are not satisfied for many characteristics and in many breeding programs. A locus involved in the phenotype of interest may not segregate in Mendelian fashion because of reduced fitness of either a gametophyte or a sporophyte carrying the desired allele at the locus of interest. Non-Mendelian segregation may also be the result of linkage of the desired allele to an undesired allele at a locus that alters allelic frequencies in viable gametes. Examples of loci affecting the segregation of alleles include *Ge* (*Gamete eliminator*), *Gp* (*Gamete promoter*), and *X* (*gametophytic factor*) in tomato (Rick 1965; Pelham 1968, 1970; Laterrot 1975), pollen killer in wheat, spore killer in *Neurospora*, and segregation distorter in *Drosophila*. In these systems, there is a severe reduction in the frequency of one allelic form of the active locus and against the alleles tightly linked in coupling to the eliminated allele at the active locus. Aberrant segregation is also frequently observed in segregating populations derived from interspecific crosses (Zamir and Tadmor 1986). This is an important concern, because many breeding programs involve the transfer of desired traits from wild species into domesticated crop species.

In accordance with Mendel's first law, each gamete from a diploid individual may be viewed as an independent trial in which one of two alleles is chosen at the locus being considered. The number of occurrences of a given allele in a population of gametes from a single heterozygous individual follows a binomial distribution. Mendel's first law is therefore the basis for the pivotal role of binomial theory in genetics problems and in questions concerning sample size. In accordance with Mendel's second law, the segregation of alleles at two or more loci is independent, assuming that the loci are functionally unlinked. This assumption may or may not be met, depending upon the genomic locations of the loci under consideration. Mendel's second law deals with the joint occurrence of events at two or more loci, which can be handled by standard probability theory for independent events.

The goal of this paper is to extend previous methods to permit determination of the minimal size of a segregating population that must be screened to ensure a desired level of confidence of obtaining at least m plants with the desired genotype at a set number of loci. Factors affecting differential selection of gametes, segregation at a single locus, and linkage among the loci of interest are taken into consideration. We will first consider the effects in the gametophyte (haploid level) of fitness and linkage on the frequencies of alleles at two linked loci, then continue to the consideration of three or more linked loci. Then we will consider the effects of the altered allelic frequencies on the likelihood of obtaining a desired genotype or phenotype in the sporophyte (diploid level). In the exposition, we assume for simplicity that each trait is controlled by an allele at one locus. However, the results presented here are completely applicable under more general conditions, when multiple genes control a single trait. Although we view these results from a plant breeding perspective, the method is not limited to the plant kingdom and is generally applicable in genetic research.

The model for two linked loci

We now develop a model for the effect of linkage relationships between two loci. We will examine first the frequencies of the haploid (gamete) genotypes, and then the frequencies of the diploid (sporophyte) genotypes. Throughout this paper, we focus on the case of two possible alleles at each locus. This is not a restriction for most loci in diploid species. Considering all possible gametes produced by an individual, let p denote the relative frequency of a desired allele at a specified locus. If differential survival and mutation are absent in a pool of gametes from a heterozygous individual, the relative frequency of each allele occurring at any locus is $p = 0.5$. The probability distribution of the number of occurrences, x, of a desired allele in a sample of n gametes is given by the binomial formula

$$P(\text{exactly } x \text{ successes}) = f(x; n, p)$$

$$= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$$\text{for } x = 0, 1, \ldots, n. \quad (1)$$

The probability that there are at least m successes, or occurrences of the desired allele, in n gametes is

$$P(\text{at least } m \text{ successes}) = \sum_{x=m}^{n} f(x; n, p) \quad \text{for } m = 0, 1, \ldots, n. \quad (2)$$

The probability of at least one occurrence of the desired allele is

$$P(\text{at least one success}) = 1 - f(0; n, p) = 1 - (1-p)^n. \quad (3)$$

The binomial model (1) gives the probability of obtaining any specified number of successes, x, in a series of n independent observations, each of which has success probability p. In this general setting, (2) and (3) also hold. The definition of a success may be that some desired combination of alleles occurs at two or more loci, which may or may not be independent, or that a zygote with a desired phenotype is formed by a random pair of gametes. Throughout this paper, success will be defined along these lines. For example, for two independent loci, the probability that a given gamete possesses both desired alleles is $p = p_a p_b$, where p_a and p_b are the relative frequencies of the desired alleles a_1 and b_1, respectively. The probability that there are at least m successes, that is, at least m gametes out of n that possess both desired alleles, is given by equation (2) with $p = p_a p_b$, that is, with p replaced by the expression $p_a p_b$. The probability that there is at least one success is given by equation (3) with $p = p_a p_b$.

If two loci are linked, their segregation will be correlated rather than independent. Let desired alleles be denoted by the subscript 1, e.g., a_1, b_1, and the corresponding undesired alleles by the subscript 2, e.g., a_2, b_2. If a_1 and b_1 are in repulsion, there must be a crossover between the loci to get the two desired alleles onto the same chromosome.

If r denotes the percent recombination between the two loci, then each of the two parental genotypes occurs with probability $0.5(1-r)$, and each of the recombinant genotypes occurs with probability $0.5r$, as shown in Fig. 1.

Let the recombination parameter λ be the function of the percent recombination shown in Fig. 2 and specified by

$$\lambda = f(r) = \begin{cases} r \text{ for } 0 \le r \le 0.5 \text{ under coupling} \\ 1 - r \text{ for } 0 \le r \le 0.5 \text{ under repulsion} \end{cases} \quad (4)$$

The parameter λ varies between 0 and 1. Using the parameter λ instead of r allows us to consider both coupling and repulsion within a common, unified formulation. When λ is 0.5, the loci segregate independently, and there is no linkage. When λ is 0, linkage is complete in coupling; when λ is 1, linkage is complete in repulsion. In both of these cases of complete linkage, the recombination rate r equals 0. If the distance between the two loci exceeds zero, then λ can approach 0 or 1 but cannot equal either of these values.

The probabilities of the possible combinations of alleles at the two loci are given in the Punnett squares in Table 1. In the case of coupling, $\lambda = r$, so $0 \le \lambda < 0.5$, giving the Punnett square in Table 1A. In the case of repulsion, $\lambda = 1 - r$, so $0.5 < \lambda \le 1$, leading to the Punnett square in Table 1B. In the case of independence, $\lambda = 0.5$, and the Punnett square is the one shown in Table 1C, commonly found in textbooks. All three of these Punnett squares are special cases of the square in Table 1D.

The probability of observing $a_1 b_1$ in a randomly selected gamete is $0.5(1-\lambda)$, where the degree and kind of linkage determine the value of the parameter λ. For a random sample of n gametes, the probability that at least one $a_1 b_1$ gamete occurs is given, based on the

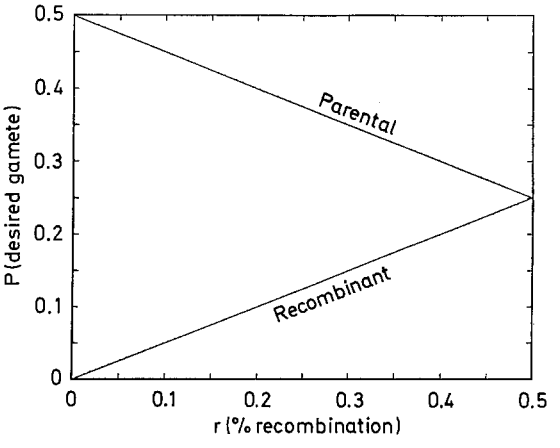

Fig. 1. The probability of each type of gamete as a function of percent recombination

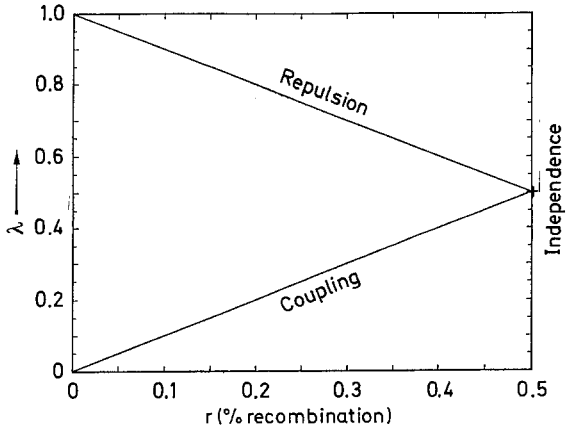

Fig. 2. λ plotted as a function of percent recombination

Table 1. Probabilities of the possible combinations of alleles at two linked loci

Locus B	Locus A	
	a_1	a_2
b_1	$0.5(1-r)$	$0.5r$
b_2	$0.5r$	$0.5(1-r)$

A. Coupling

Locus B	Locus A	
	a_1	a_2
b_1	0.25	0.25
b_2	0.25	0.25

C. Independence

Locus B	Locus A	
	a_1	a_2
b_1	$0.5r$	$0.5(1-r)$
b_2	$0.5(1-r)$	$0.5r$

B. Repulsion

Locus B	Locus A	
	a_1	a_2
b_1	$0.5(1-\lambda)$	0.5λ
b_2	0.5λ	$0.5(1-\lambda)$

D. Generalized probabilities for coupling, repulsion, and independence

binomial distribution $f[x; n, 0.5(1-\lambda)]$, by equation (3) with $p = 0.5(1-\lambda)$. The probability that at least m out of the n randomly selected gametes possess the genotype a_1b_1 is given by the sum of terms from this binomial distribution produced by m or more gametes with genotype a_1b_1, that is, by equation (2) with $p = 0.5(1-\lambda)$.

We now incorporate differential survival rates among genotypes into the analysis. Let $u(g)$ denote the recombination frequency of genotype g, that is, the relative frequency with which this genotype is produced. For example, in Table 1D, $u(a_1b_1) = u(a_2b_2) = 0.5(1-\lambda)$ and $u(a_1b_2) = u(a_2b_1) = 0.5\lambda$. Let $s(g)$ denote the survival rate of genotype g, which may differ from one genotype to another. For example, if a gamete carrying the b_2 allele were nonviable at high temperatures, then the survival rates of genotypes involving b_2, e.g., $s(a_1b_2)$, would be 0 under these conditions. For each genotype g, the fraction of gametes possessing this genotype and surviving is given by $u(g)s(g)$, the product of the relative frequency of g and the survival rate of g. In a pool of gametes, let Q denote the total fraction of surviving gametes; it is obtained by summing the term $u(g)s(g)$ over all possible genotypes g:

$$Q = \sum_g u(g)s(g). \tag{5}$$

The proportion of these surviving gametes with genotype g is denoted by $v(g)$; it is equal to the fraction of gametes possessing genotype g and surviving, expressed as a proportion of Q, the fraction of all surviving gametes:

$$v(g) = u(g)s(g)/Q. \tag{6}$$

Under coupling, each of the parental genotypes, $g = a_1b_1$ and $g = a_2b_2$, has a relative frequency $u(g) = 0.5(1-r)$, while each of the recombinant genotypes, $g = a_1b_2$ and $g = a_2b_1$, has a relative frequency $u(g) = 0.5r$. Under repulsion, each of the parental genotypes, $g = a_1b_2$ and $g = a_2b_1$, has a relative frequency $u(g) = 0.5(1-r)$, while each of the recombinant genotypes, $g = a_1b_1$ and $g = a_2b_2$, has a relative frequency $u(g) = 0.5r$. In both cases, as well as for independence, it is routine to verify from Tables 1A and 1B that $u(g) = 0.5(1-\lambda)$ for $g = a_1b_1$ and $g = a_2b_2$, $u(g) = 0.5\lambda$ for $g = a_1b_2$ and $g = a_2b_1$.

Table 2. Punnett square: two alleles, with differential survival among genotypes. Entry for each genotype g is $v(g) = u(g)s(g)/Q$

Locus B	Locus A	
	a_1	a_2
b_1	$0.5(1-\lambda)s(a_1b_1)/Q$	$0.5\lambda s(a_2b_1)/Q$
b_2	$0.5\lambda s(a_1b_2)/Q$	$0.5(1-\lambda)s(a_2b_2)/Q$

The Punnett square under differential survival is then given by Table 2. Its entries are the terms $v(g)$ of (6), expressed in terms of λ, $s(g)$, and Q. For example, in Table 2, the term $v(a_1b_1)$ appears as $0.5(1-\lambda)s(a_1b_1)/Q$. The entries in Table 1D have been used to express the relative frequency $u(g)$ from (6) in terms of λ for each g; for instance, $u(a_1b_1)$ is expressed as $0.5(1-\lambda)$. It is clear from (5) and (6) that Q, the total fraction of survivors, is the sum of the four numerators in the entries of this square. Consequently, the sum of the entries of the Punnett square in Table 2 is 1, and the entries are the relative frequencies of the genotypes among the surviving gametes. If the survival rates $s(g)$ are all equal, the Punnett square of Table 2 reduces to the square of Table 1D by routine algebra. If the survival rate of the gamete is determined by the haploid genotype, then the survival rate $s(g)$ will be a product of

survival rates associated with individual alleles at each of the loci under consideration: $s(g) = s_a(g_a)s_b(g_b)$, where s_a is the survival rate of allele g_a, which may be either a_1 or a_2, and similarly for b.

To compute the terms $v(g)$, the proportions of all genotypes g found among the surviving gametes, we must know the kind of linkage (coupling or repulsion) and the percent recombination r, from which we can find λ, and the survival rates $s(g)$. Because relative survival rates, not absolute rates, are important, it suffices to know the three ratios $s(g)/s(a_2b_2)$ for $g \neq a_2b_2$. Similarly, if a gamete's survival is determined by its genotype, it suffices to know the two ratios $s_a(a_1)/s_a(a_2)$ and $s_b(b_1)/s_b(b_2)$. The probability of at least m successes is then given by equation (2) with $p = v(a_1b_1)$. For $m = 1$, this reduces to equation (3) with $p = v(a_1b_1)$.

Determination of required sample size

We now derive the sample size necessary to ensure that the probability of obtaining at least m successes attains any desired value α. Applying the general formulas (2) and (3) for the binomial probability model to the case of two linked loci, let p denote the probability $v(a_1b_1)$ from equation (6) and Table 2. Let m be the number of successes (occurrences of this desired genotype) that we want to obtain in a random sample of n gametes from a heterozygous plant. The desired number of successes, m, can take on a variety of values, depending on the needs of the breeder. Let α be the desired probability of obtaining at least m successes; this will be specified in advance, typically as a high value (e.g., 0.99, 0.95, or 0.90) so the chance of failing to observe at least m successes will be small. For given values of p, m, and α, the goal is to determine the smallest sample size n for which

$$P(\text{at least } m \text{ successes}) \geq \alpha, \tag{7}$$

where the left-hand side is calculated from (2) for $m > 1$ or from (3) for $m = 1$.

The probability p can take on a wide range of values. In the case of two loci with no linkage, no differential survival, and no mutation, each of the four possible gametes has an equal probability (0.25). Considering varying degrees of linkage resulting in a recombination rate ranging from 50 percent (no linkage) to nearly zero percent (total linkage), the probability of a desired gamete of a parental genotype varies from 0.25 to nearly 0.50, and the probability of a desired gamete of a recombinant genotype varies from 0.25 to nearly zero. (These possibilities are illustrated in Fig. 1.)

To determine the smallest sample size n satisfying (7), P(at least m successes) was calculated for given values of p, m and α and a trial value of n. If this probability was less than α, a larger trial value was substituted for n. If (7) was satisfied, and if this remained true when n was replaced by $n - 1$, a smaller trial value was substituted for n. This process continued until (7) was satisfied for n but not when n was replaced by $n - 1$. A computer program for evaluating P(at least m successes) was written in FORTRAN, using the IMSL subroutine DBINDF (IMSL 1989). The program runs quickly and accurately, with all calculations performed in double precision mode. A different programming approach was taken by Mansur et al. (1990); their program for the IBM PC executes slowly as m increases and/or p decreases.

For $\alpha = 0.95$, the smallest adequate sample size n was found for selected pairs m and p. The values of m were 1, 5, 10, 20, 30, 40, and 50. The values of p ranged from 0.0001 to 0.001 in increments of 0.0001, from 0.001 to 0.01 in increments of 0.001, and from 0.01 to 0.50 in increments of 0.01. For each pair m and p of these values, the smallest adequate sample size was computed. These results are graphed in Fig. 3, with both p and n on logarithmic scales, and the values of n for selected pairs m and p are shown in Table 3. The linearity of the curves in Fig. 3 is striking. For each value of m, the required sample size n in Table 3 is inversely proportional to p when p is small. The mathematical basis for this behavior is explained in the Appendix. Complete tables for $\alpha = 0.99$, 0.95, and 0.90, and the computer program that produced them, are available from the first author. Other values of α can easily be accommodated in this program if desired.

Example: For two loci with no linkage, no differential survival, and no mutation, the relative frequency of

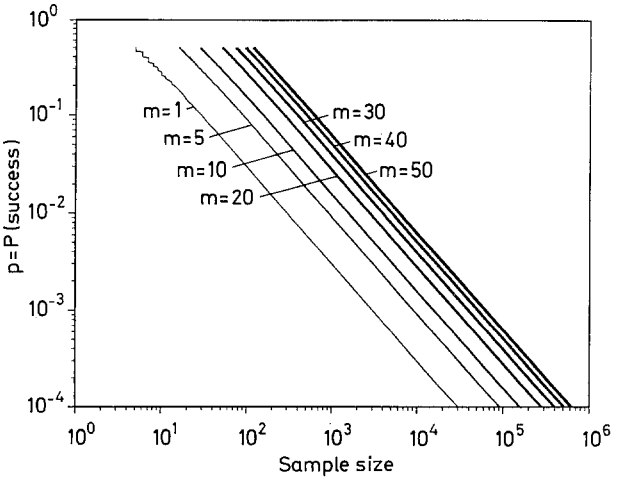

Fig. 3. Sample size n required to produce with probability 0.95 at least m individuals of a desired genotype or phenotype, as a function of m and the probability p of a randomly selected individual being of the desired type

Table 3. Table of sample size n required to produce with probability 0.95 at least m individuals of a desired genotype or phenotype, as a function of m and the probability p of a randomly selected individual being of the desired type

p	m				
	1	5	10	20	40
0.0001	29956	91533	157049	278788	509392
0.0002	14978	45766	78523	139392	254693
0.0004	7488	22882	39260	69694	127344
0.0006	4992	15254	26173	46461	84894
0.0008	3744	11440	19629	34845	63669
0.001	2995	9151	15702	27875	50934
0.002	1497	4575	7850	13936	25464
0.004	748	2286	3923	6966	12729
0.006	498	1524	2615	4643	8484
0.008	373	1142	1960	3481	6362
0.01	299	913	1568	2784	5088
0.02	149	456	782	1390	2541
0.04	74	227	390	693	1268
0.06	49	150	259	461	843
0.08	36	112	193	344	631
0.10	29	89	154	275	504
0.15	19	59	102	183	334
0.20	14	44	76	135	249
0.25	11	34	60	107	198
0.30	9	28	49	89	164
0.35	7	24	42	75	140
0.40	6	21	36	65	121
0.45	6	18	32	57	107
0.50	5	16	28	51	96

each of the four possible gametes is 0.25. A sample of 34 gametes is necessary to ensure 95% certainty of obtaining at least five gametes of the desired form a_1b_1. If the desired alleles, a_1 and b_1, are linked in repulsion and the relative frequency of recombination between the loci is 20%, then the relative frequency of the desired gamete is 0.10, and a sample of 89 gametes is necessary to ensure 95% certainty of having at least five gametes of the desired genotype a_1b_1.

Example: For two loci possibly linked but with no differential survival and no mutation, the value of $p = v(a_1b_1)$ is $0.5(1-\lambda)$. Take $m = 1$; then equation (3) becomes

$$\alpha = 1 - [1 - 0.5(1-\lambda)]^n. \tag{8}$$

Solving this gives an explicit formula for n in terms of α and λ,

$$n = \log(1-\alpha)/\log[1 - 0.5(1-\lambda)]$$
$$= \log(1-\alpha)/\log[0.5(1+\lambda)]. \tag{9}$$

When the loci segregate independently ($\lambda = 0.5$), this reduces to $n = \log(1-\alpha)/\log(0.75)$, which for $\alpha = 0.95$ gives $n = 10.4$, so the minimum sample size is 11. When coupling is present between the loci ($\lambda < 0.5$), sample sizes will be smaller than for independence between the loci; e.g., $\lambda = 0.4$ gives $p = v(a_1b_1) = 0.30$, and for $\alpha = 0.95$ equation (9) gives $n = \log(0.05)/\log(0.70) = 8.4$, so the minimum sample size is 9. When repulsion is present ($0.5 < \lambda < 1$), sample sizes will be larger than for independence; e.g., $\lambda = 0.6$ gives $p = v(a_1b_1) = 0.20$, and for $\alpha = 0.95$ equation (9) gives $n = \log(0.05)/\log(0.80) = 13.4$, so the minimum sample size is 14. These values (9, 11, and 14) appear in the $m = 1$ column of Table 3. As λ approaches the value 1, corresponding to complete repulsion, $v(a_1b_1)$ approaches 0, and the sample size needed to achieve a given level of probability increases without bound. In practice, the exact value of λ may not be known, but should be estimated by the observation of appropriate segregating populations.

We end this section by pointing out the extremely wide applicability of its results. To apply formulas (2) and (3) to the situation of two linked loci, one need only replace the general term p by the quantity $v(a_ib_j)$, which was treated in the section on two linked loci. The method for determining sample size is equally applicable to three or more linked (or unlinked) loci, where p will be determined by quantities of the form $v(a_ib_jc_k)$, and to zygotes of a desired phenotype formed by a random pair of gametes, where p will be of a form $P(\mathscr{C})$ to be discussed in the section on the sporophyte level. In all of these cases, once the proper term to use for the probability p has been worked out, the values of m, p, and α are known and the analysis can proceed according to the results of Fig. 3 and Table 3.

Three linked loci

In the case of three loci with no linkage, each of the eight possible gametes has an equal probability (0.125). Considering varying degrees of linkage on either side of the central locus from 50 percent recombination (no linkage) to nearly zero percent recombination (total linkage), the probability of a desired gamete of a parental genotype varies from 0.125 to 0.50, the probability of a desired gamete of a recombinant genotype involving a single crossover varies from 0.25 to nearly zero, and the probability of a desired gamete of a recombinant genotype involving a double crossover varies from 0.125 to nearly zero. The speed with which these values approach their limits is shown in Fig. 4A–D. Let r_1 denote the percent recombination between loci 1 and 2, and r_2 the percent recombination between loci 2 and 3. Then each of the two parental genotypes occurs with probability $0.5(1-r_1)(1-r_2)$, each recombinant genotype with a single crossover between loci 1 and 2 occurs with probability $0.5r_1(1-r_2)$, each recombinant genotype with a single crossover between loci 2 and 3 occurs with probability $0.5(1-r_1)r_2$, and each of the recombinant genotypes

with a double crossover occurs with probability $0.5r_1r_2$. These probabilities are depicted graphically by the surfaces in Fig. 4A–D, which show the values of the probabilities as functions of the pair (r_1, r_2). Their algebraic expressions appear in Table 4. The parental probability in Fig. 4A changes more rapidly than the other probabilities as a function of (r_1r_2); the single crossover probabilities in Fig. 4B, C are the same except that the roles of r_1 and r_2 are interchanged.

Consider three loci arranged on the chromosome. Let the desired alleles at these loci be denoted by a_1, b_1, and c_1 and the three corresponding undesirable alleles by a_2, b_2, and c_2. Three cases, shown in Fig. 5A–C, must be considered: coupling of a_1, b_1, and c_1; coupling between a_1 and b_1 and repulsion between b_1 and c_1; and repulsion between a_1 and b_1 and also between b_1 and c_1. A fourth case, repulsion between a_1 and b_1 and coupling between b_1 and c_1, is shown in Fig. 5D; it is equivalent to the case of Fig. 5B by symmetry. Let the recombination parameters λ_1 and λ_2 be the functions of the percent recombinations r_1 and r_2 specified by

$$\lambda_1 = f(r_1) = \begin{cases} r_1 \text{ for } 0 \le r_1 \le 0.5 \text{ under coupling} \\ 1 - r_1 \text{ for } 0 \le r_1 \le 0.5 \text{ under repulsion,} \end{cases} \quad (10)$$

$$\lambda_2 = f(r_2) = \begin{cases} r_2 \text{ for } 0 \le r_2 \le 0.5 \text{ under coupling} \\ 1 - r_2 \text{ for } 0 \le r_2 \le 0.5 \text{ under repulsion.} \end{cases} \quad (11)$$

For the case of coupling among all pairs, $\lambda_1 = r_1$ and $\lambda_2 = r_2$. Table 4 lists the genotypes g and their probabilities $u(g)$. Survival rates $s(g)$ behave exactly as for two linked loci. When there are three linked alleles, equations (5) and (6) remain valid, but summation is over the eight genotypes involving the three loci.

To construct the Punnett square for this situation, incorporating differential survival, use equation (6) and replace r_1 and r_2 by the equivalent expressions involving λ_1 and λ_2. The result is given in Table 5. Its entries are the terms $v(g)$ expressed in terms of λ_1, λ_2, $s(g)$, and Q.

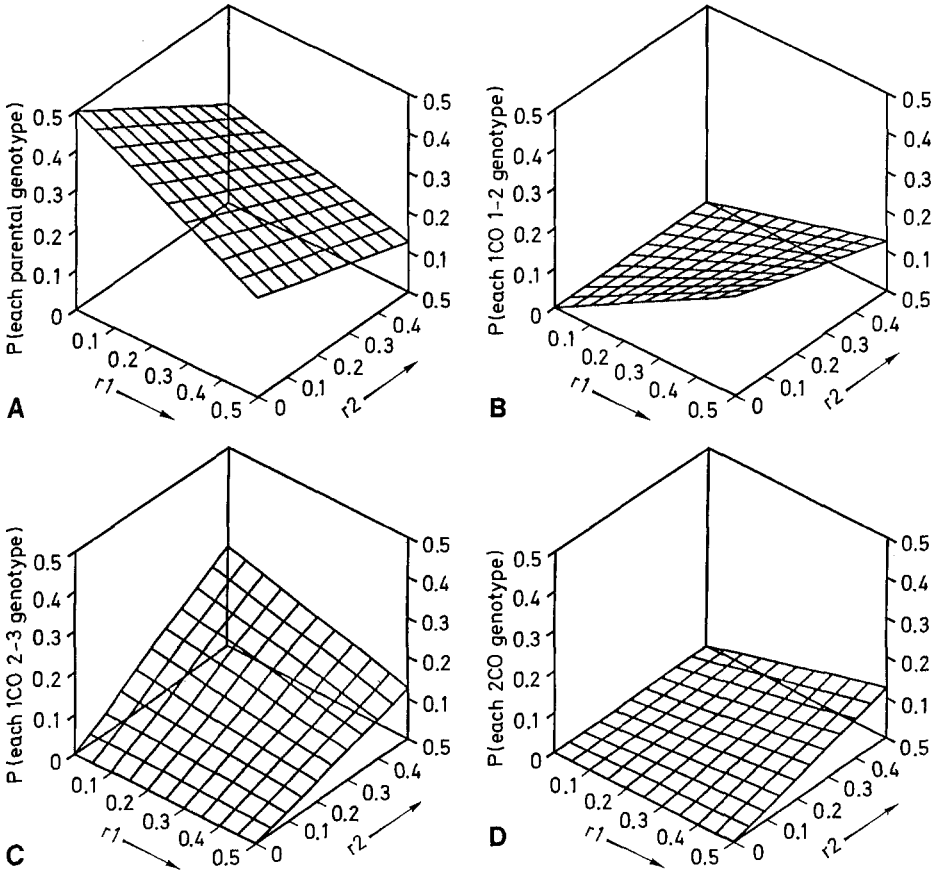

Fig. 4A–D. Probabilities of parental, single crossover, and double crossover genotypes as functions of percent recombinations r_1, r_2. **A** Parental genotype: $P(\text{each parental genotype}) = 0.5(1 - r_1)(1 - r_2)$. **B** Single crossover 1-2 genotype: $P(\text{each 1CO 1-2 genotype}) = 0.5r_1(1 - r_2)$. **C** Single crossover 2-3 genotype: $P(\text{each 1CO 2-3 genotype}) = 0.5(1 - r_1)r_2$. **D** Double crossover genotype: $P(\text{each 2CO genotype}) = 0.5r_1r_2$

Table 4. Coupling among all pairs: genotypes and their probabilities

Description	Probability $u(g)$	Genotypes g
Parental	$0.5(1-r_1)(1-r_2)$	$a_1b_1c_1$, $a_2b_2c_2$
Single crossover, alleles 1-2	$0.5r_1(1-r_2)$	$a_1b_2c_2$, $a_2b_1c_1$
Single crossover, alleles 2-3	$0.5(1-r_1)r_2$	$a_1b_1c_2$, $a_2b_2c_1$
Double crossover	$0.5r_1r_2$	$a_1b_2c_1$, $a_2b_1c_2$

It is useful to examine some special cases. When all of the survival rates $s(g)$ equal 1, then $Q=1$, so the $s(g)$ terms and Q drop out of all entries in Table 5, e.g., the entry $v(a_1b_1c_1)$ reduces to $0.5(1-\lambda_1)(1-\lambda_2)$. In fact, when all of the survival rates $s(g)$ equal any common value, then Q also equals that value, leading to the same result. When all of the survival rates are equal and the three loci are independent, so $\lambda_1=\lambda_2=0.5$, each of the entries in Table 5 reduces to 1/8. When all of the survival rates are equal, loci 1 and 2 are independent, and loci 2 and 3 exhibit complete coupling, so $\lambda_1=0.5$ and $\lambda_2=0$, each of the four entries in the top and bottom rows of Table 5 equals 1/4, while the remaining four entries all equal 0.

For the case of coupling between a_1 and b_1 and repulsion between b_1 and c_1, $\lambda_1=r_1$ and $\lambda_2=1-r_2$. Genotypes $a_1b_1c_2$ and $a_2b_2c_1$ are parentals, $a_1b_2c_1$ and $a_2b_1c_2$ are single crossovers at 1-2, $a_1b_1c_1$ and $a_2b_2c_2$ are single crossovers at 2-3, and $a_1b_2c_2$ and $a_2b_1c_1$ are double crossovers. Again the survival rates $s(g)$ behave exactly as for two linked alleles and equations (5) and (6) remain valid. To construct the Punnett square for this situation, incorporating differential survival, use equation (6) and replace r_1 and r_2 by λ_1 and $1-\lambda_2$, their equivalent expressions involving λ_1 and λ_2, respectively. The result is again the square in Table 5.

For the case of repulsion between a_1 and b_1 and also between b_1 and c_1, $\lambda_1=1-r_1$ and $\lambda_2=1-r_2$. The method of the previous two cases leads again to the Punnett square in Table 5. The case of three linked loci, with the desired gamete having a recombinant genotype involving a double crossover, occurs commonly in the introduction of a gene from a wild to a domestic species. In this situation, one desires to transfer the gene in as small a segment of flanking DNA as possible.

If success is defined to be the occurrence of a gamete with genotype $a_1b_1c_1$, which happens with probability $v(a_1b_1c_1)$, then the probability of obtaining at least m successes in a group of n randomly chosen gametes is given by the binomial probability formula (2), with the general success probability p replaced by its specific expression $v(a_1b_1c_1)$ for the situation of three linked loci. Thus, all the results on determination of required sample size are immediately applicable here, as was noted at the end of that section. For $m=1$, this reduces to the probability of obtaining at least one success out of n random gametes, which is given by equation (3) with $p=v(a_1b_1c_1)$.

To determine the sample size necessary to make the probability of obtaining at least m successes attain any specified value α, begin by determining $v(a_1b_1c_1)$. Then, using equation (2) if $m>1$ and equation (3) if $m=1$, evaluate the expression on the equation's right-hand side for a sequence of values of n. As n increases, the right-hand side of (2) or (3) increases, so it is not difficult to find the smallest value of n that makes the right-

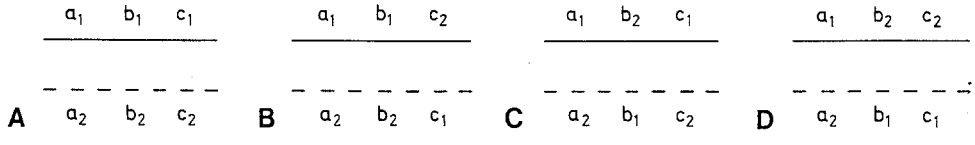

Fig. 5A–D. Arrangements of three linked loci. **A** Coupling a_1–b_1 and b_1–c_1. **B** Coupling a_1–b_1, repulsion b_1–c_1. **C** Repulsion a_1–b_1 and b_1–c_1. **D** Repulsion a_1–b_1, coupling b_1–c_1

Table 5. Punnett square: three alleles, with differential survival among genotypes

Locus B	Locus C	Locus A	
		a_1	a_2
b_1	c_1	$0.5(1-\lambda_1)(1-\lambda_2)s(a_1b_1c_1)/Q$	$0.5\lambda_1(1-\lambda_2)s(a_2b_1c_1)/Q$
	c_2	$0.5(1-\lambda_1)\lambda_2 s(a_1b_1c_2)/Q$	$0.5\lambda_1\lambda_2 s(a_2b_1c_2)/Q$
b_2	c_1	$0.5\lambda_1\lambda_2 s(a_1b_2c_1)/Q$	$0.5(1-\lambda_1)\lambda_2 s(a_2b_2c_1)/Q$
	c_2	$0.5\lambda_1(1-\lambda_2)s(a_1b_2c_2)/Q$	$0.5(1-\lambda_1)(1-\lambda_2)s(a_2b_2c_2)/Q$

hand expression equal to or greater than α. Once the value of $v(a_1b_1c_1)$ is known, Table 3 can be used with $p = v(a_1b_1c_1)$ to calculate the required sample size.

Example: For three linked loci, assume equal survival rates and no mutation, and take $m = 1$. Then the value of $p = v(a_1b_1c_1)$ is $0.5\,(1-\lambda_1)(1-\lambda_2)$, and equation (3) becomes

$$\alpha = 1 - [1 - 0.5(1-\lambda_1)(1-\lambda_2)]^n. \tag{12}$$

Solving this gives an explicit formula for n in terms of α, λ_1, and λ_2,

$$n = \log(1-\alpha)/\log[1 - 0.5(1-\lambda_1)(1-\lambda_2)]. \tag{13}$$

When coupling is present between both adjacent pairs of loci, sample sizes will be smaller than for independence between loci. Similarly, when repulsion is present, sample sizes will be larger than for independence. As either λ_1 or λ_2 approaches the value 1, corresponding to complete repulsion, the sample size needed to achieve a given level of probability increases without bound.

Four or more linked loci

In multiple gene selection, the t loci on a chromosome can be linked unless the map distance between two adjacent loci on the chromosome is greater than 50 centiMorgans. The procedure described for three loci can be generalized easily, quantified by the parameters λ_i, $i = 1, \ldots, t-1$, defined as in (10) and (11). The number of possible patterns of coupling and repulsion increases with t. For example, for $t = 4$, the following cases all require different analyses:

1. Coupling of a_1, b_1, c_1, and d_1,
2. Coupling between a_1 and b_1 and also between b_1 and c_1, and repulsion between c_1 and d_1,
3. Coupling between a_1 and b_1, and repulsion between b_1 and c_1 and also between c_1 and d_1,
4. Coupling between a_1 and b_1 and also between c_1 and d_1, and repulsion between b_1 and c_1,
5. Repulsion between a_1 and b_1, between b_1 and c_1, and also between c_1 and d_1,
6. Coupling between b_1 and c_1, and repulsion between a_1 and b_1 and also between c_1 and d_1.

Each of the other two possible situations (repulsion between a_1 and b_1, and coupling between b_1 and c_1 and also between c_1 and d_1; and repulsion between a_1 and b_1 and also between b_1 and c_1, and coupling between c_1 and d_1) is equivalent to one of the above by symmetry. As with three linked loci, these situations can be integrated through a common Punnett square whose entries are expressions involving the λ_is, survival rates $s(g)$ and Q. Success is defined to be the occurrence of a gamete with genotype $a_1b_1c_1d_1$, and the binomial probability formulas (2) and (3) apply with $p = v(a_1b_1c_1d_1)$. Again, Table 3 can be used with $p = v(a_1b_1c_1d_1)$ to calculate the required sample size.

More than four linked loci can be analyzed similarly. No new concepts are involved, but the computations grow increasingly laborious with more loci.

The sporophyte/diploid level

Let $\mathscr{C}$ denote the conditions under which fusion of a random pair of gametes results in a zygote with the desired phenotype. The specifics of $\mathscr{C}$ are determined by the number of loci involved, the linkage relationships among these loci, and the desired combination of alleles (homozygous for one allele, heterozygous, or either) at each locus. The probability of obtaining the desired phenotype at the diploid level from a randomly chosen pair of gametes, denoted by $P(\mathscr{C})$, is then given by

$$P(\mathscr{C}) = P(\text{random pair of gametes satisfy } \mathscr{C}) = \sum_{\mathscr{C}} v(g)v(g') \tag{14}$$

where $\sum_{\mathscr{C}}$ denotes summation over all pairs of genotypes (g, g') satisfying $\mathscr{C}$. It then follows that the probability of at least m successes and the probability of at least one success are given again by the binomial probability formulas (2) and (3), with the general success probability p replaced by its specific expression $P(\mathscr{C})$ for the situation at the diploid level. Thus, Table 3 can be used with $p = P(\mathscr{C})$ to calculate the required sample size.

In the examples that now follow, the expressions $v(a_ib_j)$ will be represented by the more compact notation v_{ij} for all i, j, and $v(a_ib_jc_k)$ by v_{ijk} for all i, j, k. Because we are now considering desirability at the diploid level, the numeric subscripts of the alleles are now used solely for identification, and do not imply the desirability or undesirability of the allele. All possible combinations of gametes involving two loci in a diploid genome and the probabilities of these combinations are given in Table 6. Analogous tables for three or more loci are straightforward to produce.

Example 1. Consider the case of two linked loci, in which the desired alleles are dominant at both loci A and B and the desired combination of alleles in the offspring is homozygous dominant at both loci. Then to obtain the desired phenotype, both randomly selected gametes must be a_1b_1, so $P(\mathscr{C}) = v_{11}^2$. Similarly, when the desired alleles are recessive at both loci, both randomly selected gametes must be a_2b_2, so $P(\mathscr{C}) = v_{22}^2$. Furthermore, when neither allele is dominant at either or both loci, if one desires homozygosity for a given

allele (for example, a_1 and b_2) at each locus, then $P(\mathscr{C})$ again equals the corresponding v_{ij}^2 (e.g., $P(\mathscr{C}) = v_{12}^2$). This demonstrates that a common structure prevails regardless of the dominance relationship at each of the loci.

Example 2. Consider the case of two linked loci, in which the desired alleles are again dominant at both A and B, but now the desired combination of alleles in the offspring is either homozygous dominant or heterozygous. Then the possibilities for a pair of gametes to satisfy conditions $\mathscr{C}$ are a_1b_1 with any other gametes, and a_1b_2 with a_2b_1. Recalling that $v_{11} + v_{12} + v_{21} + v_{22} = 1$, we have

$$P(\mathscr{C}) = v_{11}^2 + 2v_{11}v_{12} + 2v_{11}v_{21} + 2v_{11}v_{22} + 2v_{12}v_{21}$$
$$= (2 - v_{11})v_{11} + 2v_{12}v_{21}. \quad (15)$$

Example 3. Consider the case of two linked loci, in which the desired alleles are now dominant at locus A and recessive at B, and the desired combination of alleles is either homozygous dominant or heterozygous at locus A. Then there are two possibilities for a pair of gametes to satisfy conditions $\mathscr{C}$: both a_1b_1, and a_1b_1 with a_2b_1. Consequently,

$$P(\mathscr{C}) = v_{11}^2 + 2v_{11}v_{21}. \quad (16)$$

Example 4. Consider the case of three linked loci, in which the desired alleles are dominant at all three loci A, B, and C, and the desired combination of alleles is homozygous dominant at all three loci. Then both randomly selected gametes must be $a_1b_1c_1$, so $P(\mathscr{C}) = v_{111}^2$. (Similarly, when the desired phenotype is homozygous recessive at all three loci, both randomly selected gametes must be $a_2b_2c_2$, so $P(\mathscr{C}) = v_{222}^2$. Additional cases of desired genotypes at the A, B, and C loci can be constructed in a similar manner.)

To take a specific situation, assume that $m = 1$ and all survival rates $s(g)$ are equal. For specified values of λ in Examples 1–3 and λ_1 and λ_2 in Example 4, routine substitution shows that

Example 1: $P(\mathscr{C}) = v_{11}^2 = 0.25(1 - \lambda)^2$

Table 6. Probabilities of diploid combinations of gametes involving two loci

Loci	a_1b_1	a_1b_2	a_2b_1	a_2b_2
a_1b_1	v_{11}^2	$v_{11}v_{12}$	$v_{11}v_{21}$	$v_{11}v_{22}$
a_1b_2	$v_{11}v_{12}$	v_{12}^2	$v_{12}v_{21}$	$v_{12}v_{22}$
a_2b_1	$v_{11}v_{21}$	$v_{12}v_{21}$	v_{21}^2	$v_{21}v_{22}$
a_2b_2	$v_{11}v_{22}$	$v_{12}v_{22}$	$v_{21}v_{22}$	v_{22}^2

Example 2: $P(\mathscr{C}) = [2 - 0.5(1 - \lambda)](0.5)(1 - \lambda) + 2(0.5\lambda)^2$
$= 0.25(3 - 2\lambda + \lambda^2)$

Example 3: $P(\mathscr{C}) = 0.25(1 - \lambda)^2 + 2[0.5(1 - \lambda)](0.5\lambda)$
$= 0.25(1 - \lambda^2)$

Example 4: $P(\mathscr{C}) = v_{111}^2 = 0.25(1 - \lambda_1)^2(1 - \lambda_2)^2$.

Conclusions

Breeding programs must reconcile two conflicting pressures: to ensure a high probability of success and to use resources efficiently. Determining the minimum sample size needed becomes increasingly difficult when more genes are considered and when there is linkage among these genes. We have modelled genetic situations involving gametophytic and sporophytic systems of two or more loci. We have developed a method of determining the minimum sample size needed to produce with a specified probability at least m individuals of a desired genotype. Since the determination of sample size depends only on the probability p of observing the desired event and the number m of desired individuals, our method is applicable at both gametophytic and sporophytic levels. Our method is also applicable not only to genotypes controlled in simple Mendelian fashion, but also to genotypes with a structure made more complex by linkage in coupling or repulsion of some of the genes of interest, or by factors affecting the survival or success of a gamete or an individual of a specific genotype, or by both of these departures from Mendelian inheritance.

This method is valid for any diploid genetic system. With suitable modifications, the method can also be applied to polyploid systems. For ease of discussion, the assumption that each gene of interest governs a separate trait was made in describing the method; however, the method is equally valid for systems in which the genes of interest interact to govern a single oligo- or multi-genic trait or several such traits. Because many traits important for future variety development, such as considerations of yields, food/feed quality, and pest or stress resistance, are oligo- or multi-genic, the determination of sample size for selection of such traits is of particular interest.

Appendix

Table 3 shows that for each fixed value of m, the required sample size n is inversely proportional to p when p is small. The linearity of the curves in Fig. 4 over the range where p is small also reveals

this. The inverse variation can be explained by using the fact that for given values of m, n, and p, the probability of obtaining less than m successes out of n trials is well approximated by the cumulative normal probability $\Phi[(m-0.5-np)/(np(1-p))^{1/2}]$. When p is near 0, the term $1-p$ is virtually 1 and hence can be ignored. The required sample size is then well approximated by solving for n the equation

$$1-\Phi[(m-0.5-np)/(np)^{1/2}]=\alpha, \quad \text{(A1)}$$

where m, p, and α are given. Notice that n and p enter this equation only in the product term np. Now hold m and α fixed and consider a different value of p, $p'=cp$. The required sample size corresponding to p' is well approximated by solving equation (A1) with n and p replaced by n' and p', respectively. It is clear from $p'=cp$ that $n'=n/c$. Thus, if p' equals $0.1p$, n' will equal $10n$. The solutions to (A1) are very good approximations to the required sample size for values of p very near 0, so the entries in each column of Table 3 display variation inversely proportional to p. Consequently, with p and n both graphed using logarithmically scaled axes, the curves in Fig. 4 display linearity when p is very small. The approximation (A1) becomes less accurate as p increases, resulting in the gradual loss for larger p of both inverse proportionality in the columns of Table 3 and linearity in the curves of Fig. 4.

Acknowledgments. We thank Sam Fridman for producing Figs. 3 and 4.

References

Hanson WD (1959) Minimum family sizes for the planning of genetic experiments. Agron J 51:711–715

IMSL, Inc. (1989) IMSL Statistics/Library: *Fortran Subroutines for Mathematical Applications – User's Manual.* Ver. 1.1, vol. 3, 891–892

Laterrot H (1975) Localisation chromosomique de I_2 chez la tomate controlant la resistance au pathotype 2 de *Fusarium oxysporum* f. *Lycopersici.* Ann Amélior Plantes 26:485–491

Mansur LM, Hadder KM, Suarez JC (1990) A computer program for calculating the population size necessary to recover any number of individuals exhibiting a trait. J Hered 81:407–408

Pelham J (1968) Disturbed segregation of genes on chromosome 9 – gamete promoter, *Gp*, a new gene. Rep Tomato Genet Coop 18:27–28

Pelham J (1970) More information on *Gp*. Rep Tomato Genet Coop 20:38–39

Rick CM (1965) Abortion of male and female gametes in the tomato determined by allelic interaction. Genetics 53: 85–96

Scully BT, Federer WT (1993) Application of genetic theory in breeding for multiple virus resistance. In: Kyle, MM (ed) Resistance to viral disease of vegetables: genetics and breeding. Timber Press, Portland, Oregon

Sedcole JR (1977) Number of plants necessary to recover a trait. Crop Sci 17:667–668

Zamir D, Tadmor Y (1986) Unequal segregation of nuclear genes in plants. Bot Gaz 147:355–358