

**Determining the minimum sample size required to obtain sufficient progeny with a desired genotype
at two quantitative trait loci***

S.-F. Hsu Schmitz^{1†}, S. J. Schwager¹ and E. J. Pollak²

¹ Biometrics Unit, 337 Warren Hall, Cornell University, Ithaca, NY 14853, USA.

² Department of Animal Science, B22 Morrison Hall, Cornell University, Ithaca, NY 14853, USA.

Please send proofs to: Prof. S. J. Schwager
Biometrics Unit
337 Warren Hall
Cornell University
Ithaca, NY 14853
U.S.A.

* BU-1168-MB in the Biometrics Unit Technical Report Series, 337 Warren Hall, Cornell University, Ithaca, NY 14853.

[†] Formerly known as S.-F. Shyu.

Summary. In this paper we determine minimum progeny sample size n needed to obtain, with probability α , at least m individuals of desired two-locus genotype g affecting quantitative traits. The two quantitative trait loci (QTL's) of interest may be linked or independent, with or without epistatic interaction between them. Parental genotypes may be known or unknown, and gene action at either locus may range from additive to overdominance. To reduce the required sample size, mating patterns that will produce a high proportion of desired progeny are suggested for different desired progeny genotypes and dominance levels. Based on the assumption of normally distributed quantitative trait expression, individuals can be classified into a genotype or genotypic group according to their phenotypic expressions. This technique is used to select both parents and progeny with unknown genotypes. Choice of parental classification criteria for a given quantitative trait affects classification accuracy, hence the probability of obtaining progeny of desired genotype. The complexity of this probability depends on dominance level at each locus, recombination fraction, and awareness of parental genotypes. The procedure can be expanded to deal with more than two loci.

Key words: Sample size - Quantitative trait loci - Genotype - Phenotypic classification

Introduction

Choosing the minimum required sample size is an important practical and economic problem. For a broad range of genetic experiments, Hanson (1959) used the binomial distribution to find minimum sample size required to obtain at least one desired individual with a specified probability or confidence level (α). His first underlying assumption is that genes at each locus follow the Mendelian law of independent assortment. Hanson's second assumption is that all alleles at each locus have the same chance of being transmitted to progeny and have equal fitness, *i.e.*, equal survival rate. Rick (1965), Pelham (1968, 1970), and Laterrot (1975), however, reported differential allelic survival rates in tomatoes, which shows a modified binomial model for differential fitness is needed.

For genetic conservation, Crossa (1989) used the binomial model to find the sample size required to obtain, with a specified confidence level, at least one individual having a desired genotype in a large population, but only for a single locus. His implicit assumptions are similar to Hanson's.

Sedcole (1977) and Scully and Federer (1991) generalized Hanson's binomial model analysis to treat the problem in which more than one desired individual in the sample is required. They assumed independence among loci and equal allelic survival rate. Schwager *et al.* (1993) incorporated recombination rate and differential allelic survival rates into the binomial model for categorical traits controlled by one or several loci. Similar models for more than two loci are considered as well.

To extend the binomial model and the analysis of Schwager *et al.* to QTL's, this paper introduces two concepts. The first is mating pattern, which can be used to increase the probability of obtaining progeny with the desired genotype when compared with random mating, hence reducing the required sample size. This idea is also applicable to qualitative traits. The second is phenotypic classification, which allows selection of parents with unknown genotypes or genotypic groups to form a suggested mating pattern with certain classification accuracies. A *genotypic group* is defined here as a collection of all genotypes that produce the same phenotype. Continuous variation of phenotypic values of quantitative traits is incorporated in the model by assuming a normal distribution of trait values within genotypic group. Different allelic frequencies and dominance levels at the QTL are allowed, as well as differential allelic survival rates. Thus, this modified model is applicable in many situations involving two loci. For situations involving more than two loci, generalizations can be developed.

Mating pattern

A mating pattern is a pair of parental genotypic groups. Some mating patterns produce a relatively high proportion of progeny with the desired genotype, but others do not. Tables 1 and 2 present suggested mating patterns for different desired progeny genotypes and different levels of dominance at one and two diallelic loci, respectively. Alleles A and B are assumed to be dominant over a and b when dominance exists. Notation A- represents the combination of AA and Aa for complete dominance, and similarly for B-. Incomplete dominance includes codominance, partial dominance and overdominance. We assume that for autosomal loci, reciprocal mating patterns will give the same result. If genotypes of individual parents are known, to obtain the highest proportion

of desired progeny, we find the mating pattern in the column for incomplete dominance and select pairs of parents with this mating pattern. For individual parents with unknown genotypes, first find the suggested mating pattern shown in Table 1 or 2 for the appropriate dominance level, and then select pairs of parents based on their phenotypic expressions. Similar tables could be established for cases of multiallelic loci. The accuracy of classification and misclassification rate from phenotypic expressions to genotypic groups are to be discussed.

Phenotypic classification

The correspondence between phenotypes and genotypes is not always one-to-one, *e.g.*, the case of dominance and quantitative traits with continuous variation. Elsen *et al.* (1988) and Goffinet *et al.* (1990) propose some statistical methods to identify the genotype of sires at a major locus by using phenotypic expression of progeny. Here we use only the phenotypic expressions of individuals to predict genotypes or genotypic groups, *i.e.*, mass selection. For phenotypic classification of parents, Hardy-Weinberg equilibrium and linkage equilibrium are assumed. In this section, we discuss first the case of one locus, and then the case of two or more loci without or with epistatic interaction. Examples of diallelic loci are presented. Generalization for multiallelic loci requires little extra effort.

(A) One locus

Variation of trait expression due to genotypes at other loci that influence the trait of interest, environmental variance, and errors in measurement must be considered for classifying quantitative traits on phenotypic expression. To deal with these sources of variation, we combine them and call the result of this combination the residual variance for the QTL of interest.

All genotypic groups of a QTL are assumed here to have normal distributions in trait expressions with different means and residual variances, so that the trait expression of the entire population is a mixture of normal distributions (Elsen *et al.* 1988). The number of genotypic groups depends on the number of possible alleles and on dominance level. For a diallelic locus, the expected fractions of the genotypic groups in the population under Hardy-Weinberg equilibrium are $f_{AA} = f_A^2$, $f_{Aa} = 2f_A f_a$, $f_{aa} = f_a^2$, and $f_{A-} = f_{AA} + f_{Aa}$, where f_A and f_a are the frequencies of alleles A

and \mathbf{a} ; f_{AA} , f_{Aa} , and f_{aa} are the frequencies of genotypes **AA**, **Aa**, and **aa**; and f_{A-} is the frequency of phenotype **A-**. Similar results could be obtained for multiallelic loci.

Let X denote the level of trait expression for a given quantitative trait, and assume that there are three genotypic groups, which are labeled **R** (right), **C** (center), and **L** (left), with expected fractions f_R , f_C , and f_L at the locus of interest (see Figure 1). Let μ_R , μ_C , μ_L be the means of X and σ_R , σ_C , σ_L be the square roots of residual variances for the right, center, and left distributions, respectively. Usually the three distributions overlap in some parts. With known means and residual variances of these normal distributions, and known expected fraction of each genotypic group, we can calculate the expected proportion of each distribution that lies beyond any fixed point or between any two fixed points. Ducrocq and Quaas (1988) described a method to select animals from a mixture of different age group distributions. Their method is applied to this problem by replacing age groups with genotypic groups. The following formulas are also applicable to other cases with some minor modification. For example, with two genotypic groups, *e.g.*, complete dominance, the center distribution is omitted; with more than three genotypic groups, say, n groups, then there are $n-2$ adjacent center distributions.

For convenience, the first genotypic group gg_1 is assumed to be the target for a given classification criterion. Let $\theta(gg_1|X \in I)$ denote the classification accuracy of assigning an individual with trait expression X satisfying the classification criterion I , *i.e.*, $X \in I$, into the target genotypic group gg_1 . In other words, $\theta(gg_1|X \in I)$ is the probability that an individual whose trait expression X satisfies I belongs to group gg_1 . Then the misclassification rate is defined as

$$\delta(gg_1|X \in I) = 1 - \theta(gg_1|X \in I) = \sum_{i \neq 1} \theta(gg_i|X \in I), \quad (1)$$

where $\sum_{i \neq 1}$ denotes summation over all non-target genotypic groups, and the terms $\theta(gg_i|X \in I)$, $i \neq 1$, are called misclassification components. Another approach to calculating classification accuracy and misclassification components is presented in Hoeschele (1988).

As an example, with partial dominance, suppose that **AA** = **R**, **Aa** = **C**, **aa** = **L**, σ_{AA} , σ_{Aa} and σ_{aa} have the common value σ , $f_A = 0.6$, $f_a = 0.4$, $\mu_{AA} = 2\sigma$ and $\mu_{aa} = -2\sigma$; then classification accuracy and one of the two misclassification components for different boundary values and different

μ_{Aa} values are given in Tables 3, 4, and 5, for the right, left, and center classifications, respectively, which will now be presented. The other misclassification component can be calculated by subtraction.

(a) Right classification (*e.g.*, for dominant homozygote **AA**):

For an appropriate boundary value t_1 , an individual with $X \geq t_1$ is assigned into the **R** genotypic group. Then the three distributions of the three genotypic groups are truncated at $X = t_1$, and the areas of the three truncated regions to the right of t_1 , each weighted by the corresponding expected fraction f_i , are

$$A_i = f_i \left\{ 1 - \Phi \left(\frac{t_1 - \mu_i}{\sigma_i} \right) \right\} \text{ for the } i\text{th genotypic group, for } i = \text{R, C, L,} \quad (2)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Let $A_+ \equiv \sum_i A_i$; then the classification accuracy based on the interval $I = \{X: X \geq t_1\}$ is

$$\theta(\text{R}|X \geq t_1) = A_R/A_+, \quad (3)$$

and the misclassification components are

$$\theta(i|X \geq t_1) = A_i/A_+, \text{ for } i = \text{C, L.} \quad (4)$$

According to equation (1), the misclassification rate is

$$\delta(\text{R}|X \geq t_1) = 1 - \theta(\text{R}|X \geq t_1) = \sum_{i \neq \text{R}} \theta(i|X \geq t_1). \quad (5)$$

(b) Left classification (*e.g.*, for recessive homozygote **aa**):

For an appropriate boundary value t_2 , an individual with $X \leq t_2$ is assigned into the **L** genotypic group. Then the three truncated regions to the left of t_2 have areas, again weighted by genotypic group expected fraction, given by

$$B_i = f_i \Phi \left(\frac{t_2 - \mu_i}{\sigma_i} \right) \text{ for the } i\text{th genotypic group, for } i = \text{R, C, L.} \quad (6)$$

Let $B_+ \equiv \sum_i B_i$; then the classification accuracy based on the interval $I = \{X: X \leq t_2\}$ is

$$\theta(\text{L}|X \leq t_2) = B_L/B_+, \quad (7)$$

and the misclassification components are

$$\theta(i|X \leq t_2) = B_i/B_+, \text{ for } i = R, C. \quad (8)$$

Similarly, the misclassification rate is

$$\delta(L|X \leq t_2) = 1 - \theta(L|X \leq t_2) = \sum_{i \neq L} \theta(i|X \leq t_2). \quad (9)$$

(c) Center classification (*e.g.*, for heterozygote **Aa**):

The way to classify individuals into the **C** genotypic group is a little different. For appropriate boundary values t_1 and t_2 , an individual with trait expression X such that $t_2 < X < t_1$ is assigned into the **C** genotypic group. The three truncated regions in the center of the mixed distribution, between t_2 and t_1 , have the weighted areas

$$C_i = f_i \left\{ \Phi\left(\frac{t_1 - \mu_i}{\sigma_i}\right) - \Phi\left(\frac{t_2 - \mu_i}{\sigma_i}\right) \right\} \text{ for the } i\text{th genotypic group, for } i = R, C, L. \quad (10)$$

Let $C_+ \equiv \sum_i C_i$; then the classification accuracy based on the interval $I = \{X: t_2 < X < t_1\}$ is

$$\theta(C|t_2 < X < t_1) = C_C/C_+, \quad (11)$$

the misclassification components are

$$\theta(i|t_2 < X < t_1) = C_i/C_+, \text{ for } i = R, L, \quad (12)$$

and the misclassification rate is

$$\delta(C|t_2 < X < t_1) = 1 - \theta(C|t_2 < X < t_1) = \sum_{i \neq C} \theta(i|t_2 < X < t_1). \quad (13)$$

In Table 3, for a fixed μ_{Aa} , the larger t_1 is, the greater is the classification accuracy; for a fixed t_1 , the larger μ_{Aa} is, the lower is the classification accuracy. In Table 4, the pattern of the numerical entries is exactly the opposite. In Table 5, for fixed μ_{Aa} and t_2 , the larger t_1 is, the lower the accuracy; for fixed μ_{Aa} and t_1 , the classification accuracy first increases and then decreases as t_2 increases. The pattern of entries in these tables is somewhat complex because many factors are involved.

Any change in allelic frequencies and dominance levels will clearly affect the result. For example, with overdominance, the situation might become **Aa = R**, **AA = C**, **aa = L**; with complete dominance, it might become **A- = R**, **aa = L**. Similar tables for different allelic frequencies and

dominance levels are provided in Shyu (1992). In addition, the difference between μ_{AA} and μ_{aa} is also influential, that is, the larger $(\mu_{AA} - \mu_{aa})/\sigma$ is, the greater the classification accuracy (for brevity, results are not presented here).

(B) M ($M \geq 2$) loci with no interaction

Let $\theta(gg_{k1}|X_k \in I_k)$ and $\delta(gg_{k1}|X_k \in I_k)$ denote the classification accuracy and misclassification rate of assigning an individual with the k th locus trait expression X_k satisfying the classification criterion I_k into the target genotypic group gg_{k1} , for $k = 1, 2, \dots, M$. Then $\theta(gg_{ki}|X_k \in I_k)$ with $i \neq 1$ are misclassification components. Similarly, $\theta(gg_{11}gg_{21} \cdots gg_{M1}|X_1 \in I_1, X_2 \in I_2, \dots, X_M \in I_M)$ and $\delta(gg_{11}gg_{21} \cdots gg_{M1}|X_1 \in I_1, X_2 \in I_2, \dots, X_M \in I_M)$ denote the joint classification accuracy and the joint misclassification rate for M loci, respectively.

Because interaction among the M loci is absent, trait expression of one of these loci is not influenced by the other loci. Thus joint classification accuracy is equal to the product of the individual single-locus classification accuracies, and joint misclassification rate consists of the sum of products of all other possible combinations of classification accuracies and misclassification components at these M loci. That is, joint classification accuracy is

$$\theta(gg_{11}gg_{21} \cdots gg_{M1}|X_1 \in I_1, X_2 \in I_2, \dots, X_M \in I_M) = \prod_{k=1}^M \theta(gg_{k1}|X_k \in I_k), \quad (14)$$

and joint misclassification rate is

$$\begin{aligned} \delta(gg_{11}gg_{21} \cdots gg_{M1}|X_1 \in I_1, X_2 \in I_2, \dots, X_M \in I_M) &= 1 - \prod_{k=1}^M \theta(gg_{k1}|X_k \in I_k) \\ &= \sum_{(i_1, i_2, \dots, i_M) \neq (1, 1, \dots, 1)} \prod_{k=1}^M \theta(gg_{ki}|X_k \in I_k); \quad (15) \end{aligned}$$

where the summation excludes the case of all i_k equal to 1.

(C) M ($M \geq 2$) loci with epistatic interaction

If M QTL's of interest show epistatic interaction in trait expressions, then trait expression of a locus would be influenced by the other $M - 1$ loci. In this case, it is impossible to calculate individual classification accuracy and misclassification components for each locus. We solve this problem by

considering these loci as a joint locus. Individuals with certain joint trait expressions have to be classified into corresponding joint genotypic groups directly, instead of into separate one-locus genotypic groups.

For the case of two loci, let W denote the total number of distinct two-locus joint genotypic groups, and $w = 1, 2, \dots, W$. Let XX represent the joint trait expression of an individual, Π be the joint classification criterion for the target two-locus genotypic group gg_1 . Then the classification accuracy is $\theta(gg_1|XX \in \Pi)$, and the misclassification rate is

$$\delta(gg_1|XX \in \Pi) = 1 - \theta(gg_1|XX \in \Pi) = \sum_{w=2}^W \theta(gg_w|XX \in \Pi). \quad (16)$$

All calculations are the same as for the one-locus case, except that the expected fraction of each joint genotypic group is involved in allelic frequencies at these two loci. For example, if linkage is in equilibrium, then the expected fraction of the genotypic group **AABB** is equal to the product of the expected fraction of **AA** and the expected fraction of **BB**. Details are in Crow and Kimura (1970) or in Shyu (1992, Table 3.3.1); the first reference also provides information for the case of linkage disequilibrium. Similar formulas could be generalized for more than two loci.

(D) Some loci with no interaction and some loci with epistatic interaction

If N ($1 \leq N \leq M - 2$) out of M loci have no interaction and $M - N$ (≥ 2) loci have epistatic interaction, then (D) is a combination of (B) and (C). We first consider these $M - N$ loci as a joint locus as in (C), then treat those N single loci and the joint locus as in (B).

The probability of success

The probability of success $P(\mathfrak{G})$ is the probability that a particular progeny from specified parents has the desired genotype \mathfrak{G} , which is the fusion of a random pair of gametes. Under random mating,

$$\begin{aligned} P(\mathfrak{G}) &= P(\text{random pair of gametes results in genotype } \mathfrak{G}) \\ &= \sum_{\mathfrak{g}} v(\mathfrak{g})v(\mathfrak{g}'), \end{aligned} \quad (17)$$

where $\sum_{\mathfrak{g}}$ denotes summation over all pairs of gametes ($\mathfrak{g}, \mathfrak{g}'$) resulting in genotype \mathfrak{G} , $v(\mathfrak{g})$ is the

fraction of the first gamete g among all surviving gametes, and $v'(g')$ is the corresponding fraction for the second gamete g' .

If parents are specified with known genotypes gt and gt' , then the probability that a particular progeny from these parents has the desired genotype \mathfrak{G} is

$$P(\mathfrak{G}) = P(\mathfrak{G}|gt \times gt') = \sum_{\mathfrak{G}} v(g|gt)v'(g'|gt'), \quad (18)$$

where $v(g|gt)$ is the fraction of gamete g among all surviving gametes from the parent with genotype gt , and $v'(g'|gt')$ is the corresponding fraction of gamete g' from the other parent.

Let $u(g|gt)$ be the fraction of gamete g among all gametes from the parent with genotype gt before survival adjustment, and $s(g)$ be the survival rate of gamete g ; then

$$v(g|gt) = \frac{s(g)u(g|gt)}{\sum_{\mathfrak{g}} s(\mathfrak{g})u(\mathfrak{g}|gt)} = \frac{s(g)u(g|gt)}{Q}, \quad (19)$$

where $\sum_{\mathfrak{g}}$ denotes summation over all gametes possible from a parent and $Q \equiv \sum_{\mathfrak{g}} s(\mathfrak{g})u(\mathfrak{g}|gt)$ (Schwager *et al.* 1993). For two-locus cases, $v(g|gt)$ is equal to the product of the fraction of the required allele at the first locus and the fraction of the required allele at the second locus, except the $v(g|gt)$ from doubly heterozygous parents, *e.g.*, **AaBb**, which is affected by the recombination rate between these two loci. Let r denote the recombination rate between the two loci of interest, so $0 < r \leq 0.5$. Then define λ as

$$\lambda = \begin{cases} r & \text{under coupling} \\ 1-r & \text{under repulsion} \end{cases}, \quad (20)$$

such that $\lambda = r = 0.5$ when the two loci are independent (Schwager *et al.* 1993). Thus the $v(g|gt)$ from doubly heterozygous parents can be expressed as a function of λ , allele fractions in genotype gt and gamete survival rates. The $v(g|gt)$ from all possible parental diallelic two-locus genotypes can be found in Table 5.1.1 in Shyu (1992). For cases of more than two loci, Schwager *et al.* (1993) present some generalizations.

If parental genotypes are not known, and parents are selected on phenotypic expression from the population, then calculation of $P(\mathfrak{G})$ becomes more complicated due to phenotypic classification. Also parents can be selected by using two different classification criteria to form a favored mating pattern.

Let s_k be the number of possible genotypic groups at locus k . For diallelic QTL's, s_k is 2 for complete dominance, and 3 for incomplete dominance. For two loci with no interaction, the additional information required for calculating the probability of success $P(\mathfrak{G})$ is the probability that an individual with trait expressions X_1 and X_2 satisfying classification criteria I_1 and I_2 , respectively, belongs to genotypic group gg_{1i} at the first locus and to genotypic group gg_{2j} at the second locus:

$$\theta(gg_{1i};gg_{2j}|X_1 \in I_1, X_2 \in I_2) = \theta(gg_{1i}|X_1 \in I_1) \times \theta(gg_{2j}|X_2 \in I_2) \quad i = 1, \dots, s_1 \text{ and } j = 1, \dots, s_2. \quad (21)$$

Because genotypic groups gg_{11} and gg_{21} are assumed to be the targets, $\theta(gg_{11}gg_{21}|X_1 \in I_1, X_2 \in I_2)$ is joint classification accuracy. When i and j are not both equal to 1, $\theta(gg_{1i};gg_{2j}|X_1 \in I_1, X_2 \in I_2)$ is a joint misclassification component. For a single individual with $X_1 \in I_1$ and $X_2 \in I_2$, there are $W=s_1s_2$ possible genotypic group combinations at these two loci to which the individual could belong. For a pair of individuals selected as parents, there are $W^2 = (s_1s_2)^2$ possible genotypic group combinations. To simplify notation, let gg_w denote two-locus genotypic group combination $gg_{1i};gg_{2j}$ for specified i and j of parent 1, $w = 1, 2, \dots, W$. Similarly, $gg_{w'}$ denotes a specified genotypic group combination of parent 2, $w' = 1, 2, \dots, W$. Then the partial probability of success $P_{ww'}(\mathfrak{G})$, which is the probability that a particular progeny from parents of the genotypic group combinations gg_w and $gg_{w'}$ has genotype \mathfrak{G} , is

$$P_{ww'}(\mathfrak{G}) = P(\mathfrak{G}|gg_w \times gg_{w'}) \\ = \sum_{i=1}^{N_w} \sum_{j=1}^{N_{w'}} f(gt_i|gg_w) f(gt_j|gg_{w'}) [\sum_{\mathfrak{G}} v(\mathfrak{G}|gt_i) v(\mathfrak{G}|gt_j)], \quad (22)$$

where N_w and $N_{w'}$ are numbers of possible two-locus genotype combinations for the genotypic group combinations gg_w and $gg_{w'}$, and $f(gt_i|gg_w)$ and $f(gt_j|gg_{w'})$ are the expected fractions of the two-locus genotype combinations gt_i and gt_j in genotypic group combinations gg_w and $gg_{w'}$, respectively. Table 6 presents $f(gt_i|gg)$ for some genotypic group combinations gg .

Because of misclassification, selected parents may belong to non-target genotypic groups. Therefore, overall probability of success $P(\mathfrak{G})$ is the sum of products of partial probabilities $P_{ww'}(\mathfrak{G})$ and corresponding probabilities of classification (classification accuracies or misclassification components). To simplify notation, define

$$\theta_w \equiv \theta(gg_w | X_1 \in I_1, X_2 \in I_2), \quad (23)$$

$$\theta'_w \equiv \theta'(gg_w' | X_1' \in I_1', X_2' \in I_2'). \quad (24)$$

Then

$$\begin{aligned} P(\mathcal{G}) &= P(\mathcal{G} | (X_1 \in I_1, X_2 \in I_2) \times (X_1' \in I_1', X_2' \in I_2')) \\ &= \sum_{w=1}^W \sum_{w'=1}^W \theta(gg_w | X_1 \in I_1, X_2 \in I_2) \theta'(gg_w' | X_1' \in I_1', X_2' \in I_2') \\ &\quad \times \left\{ \sum_{i=1}^{N_w} \sum_{j=1}^{N_{w'}} f(gt_i | gg_w) f(gt_j | gg_w') \left[\sum_{\mathcal{G}} v(g|gt_i) v'(g'|gt_j) \right] \right\} \\ &= \sum_{w=1}^W \sum_{w'=1}^W \theta_w \times \theta'_w \times P_{ww'}(\mathcal{G}). \end{aligned} \quad (25)$$

With epistatic interaction between the two loci, classification accuracies would differ, but calculation of probability of success is the same. In this case, gg_w and gg_w' represent two-locus joint genotypic groups as in equation (16), W might not be equal to $s_1 s_2$, and joint trait expressions, XX and XX' , and joint classification criteria, Π and Π' , are applied, *i.e.*,

$$\theta_w \equiv \theta(gg_w | XX \in \Pi), \quad (26)$$

$$\theta'_w \equiv \theta'(gg_w' | XX' \in \Pi'). \quad (27)$$

Substituting these probabilities of classification θ_w and θ'_w into equation (25) produces the desired result.

For cases of more than two loci, the calculation is similar with some modification of $v(g|gt_i)$ and $f(gt_i | gg_w)$, while the required θ_w has been discussed in the previous section.

Minimum required sample size

The genotype of each progeny from any specified parents is either desired (success) or undesired (failure), and is independent of genotypes of other progeny, even those from the same parents. Thus the probability of getting at least m desired progeny in a sample of size n from specified parents can be calculated by the binomial formula:

$$\begin{aligned}
 P(\text{at least } m \text{ successes}) &= \sum_{x=m}^n f[x; n, P(\mathcal{G})] \\
 &= \sum_{x=m}^n \frac{n!}{x!(n-x)!} [P(\mathcal{G})]^x [1 - P(\mathcal{G})]^{n-x}
 \end{aligned} \tag{28}$$

as in Schwager *et al.* (1993). When $m = 1$, this becomes

$$\begin{aligned}
 P(\text{at least one success}) &= 1 - f[0; n, P(\mathcal{G})] \\
 &= 1 - [1 - P(\mathcal{G})]^n.
 \end{aligned} \tag{29}$$

With known $P(\mathcal{G})$, specified m and confidence level α , we can find the minimum sample size n satisfying the condition that $P(\text{at least } m \text{ successes}) \geq \alpha$. Table 7 gives minimum sample size n for $\alpha = 0.90$ for different $P(\mathcal{G})$ and m . Table 8 gives similar results for $\alpha = 0.95$. For fixed m and α , the larger $P(\mathcal{G})$ is, the smaller n is; while for fixed $P(\mathcal{G})$ and m , the larger α is, the larger n is. For other values of $P(\mathcal{G})$, m , and α , the minimum required sample size can be calculated by simple programming or by using the FORTRAN subroutine described by Schwager *et al.* (1993).

Of course, n will increase with m if $P(\mathcal{G})$ and α are fixed. A larger ratio of m/n is preferable, which could be achieved by increasing m , but n also increases. Similarly, n will increase with α if $P(\mathcal{G})$ and m are fixed. Although $P(\mathcal{G})$ is determined by the genetic situation, not controlled by the experimenter, a larger value of $P(\mathcal{G})$ is welcome to reduce the required sample size. For quantitative traits, different classification criteria will result in different $P(\mathcal{G})$. Some illustrative examples are given in Shyu (1992).

Choosing appropriate values of m and α involves compromise: m must be large enough for the experimenter's purpose, *e.g.*, further work involving the offspring having the desired genotype, and α must be close enough to 1 to make failure to obtain the needed number of desired progeny very unlikely, *e.g.*, $\alpha = 0.95$ or 0.99 . However, if n required by the desired m and α is so large that the experiment is not feasible, then the experimenter has no choice but to reduce m or α (or both).

Geneticists and animal or plant breeders might have another question: how to recognize those m (or more) desired progeny among the n individuals in the sample? Test crosses and DNA examination are certainly applicable. However, the method of classifying individuals based on phenotypic expression could be applied to the progeny sample first, followed by test crosses or DNA

examination to screen uncertain progeny. This approach would save time, energy, and money. Such additional phenotypic classification would result in a lower confidence level.

Conclusions

Some mating patterns yield large proportions of desired progeny, but others yield small proportions. The dominance level at each locus is important in choosing mating patterns. For quantitative traits controlled by many loci, their phenotypic expressions have continuous variation because of different alleles at those loci, environmental effects, and measurement errors. Under the assumption of a normal distribution for each quantitative trait expression of a genotypic group at a locus, the expected proportion of each genotypic group can be obtained, based on certain chosen boundary value(s). These expected proportions provide information about classification accuracy and misclassification components. Appropriate boundary values would improve classification accuracy. Within each genotypic group, expected proportions of different genotypes can be calculated from allelic frequencies. With no interaction in trait expressions, two-locus joint classification accuracy is the product of individual classification accuracies of the two loci. With epistatic interaction, individuals with certain joint trait expressions are directly classified into corresponding two-locus joint genotypic groups by using a joint classification criterion.

Once the mating pattern has been decided, if parental genotypes or genotypic groups are unknown, then prospective parents are selected based on phenotypic expressions with certain classification accuracies to produce desired progeny. Finally, the binomial formula derived by Schwager *et al.* (1993) is used to calculate the minimum required progeny sample size n for which the probability of obtaining at least m individuals with desired genotype equals or exceeds the confidence level α . The probability of success (obtaining a desired progeny) can be modified to apply the formula to different mating patterns and quantitative traits. The higher the probability of success, the smaller is the sample size required to satisfy specified α and m . For fixed m and probability of success, the required sample size increases with α . Finally, desired individuals in the progeny sample can be screened by test crosses, DNA examination, or a secondary phenotypic classification as applied to parent selection.

The above procedures are general and can be applied to many situations. With some specific assumptions, the procedure is applicable to qualitative trait loci as well (Shyu, 1992), such as the case of a marker locus linked with a QTL.

Although in this paper we have used examples of two diallelic loci under the assumptions that linkage is in equilibrium before parent selection and that zygotes have equal survival rate, these constraints can be removed easily by some modification. For cases of multiallelic loci on autosomes or sex-linked chromosome(s), Karlin (1978) reviewed expected fractions of different genotypes. For the case of linkage disequilibrium, see Crow and Kimura (1970), Weir (1979), and Weir and Cockerham (1989). Differential zygote survival rates can be incorporated easily into corresponding formulas. Classification accuracy from mass selection of parents based on phenotypic expression could be improved by incorporating any available information from their relatives (Elsen *et al.* 1988 and Goffinet *et al.* 1990).

Acknowledgment. The valuable suggestions of three reviewers are gratefully acknowledged.

References

- Crossa J (1989) Methodologies for estimating the sample size required for genetic conservation of outbreeding crops. *Theor Appl Genet* 77: 153-161.
- Crow JF, Kimura M (1970) *An Introduction To Population Genetics Theory*. Burgess Publishing Company. Minneapolis, Minnesota.
- Ducrocq V, Quaas RL (1988) Prediction of genetic response to truncation selection across generations. *J Dairy Science* 71: 2543-2553.
- Elsen JM, Vu Tien Khang J, Le Roy P (1988) A statistical model for genotype determination at a major locus in a progeny test design. *Génét Sél Evol* 20(2): 211-226.
- Goffinet B, Elsen JM, Le Roy P (1990) Statistical tests for identification of the genotype at a major locus of progeny-tested sires. *Biometrics* 46: 583-594.
- Hanson WD (1959) Minimum family sizes for the planning of genetic experiments. *Agronomy J* 51: 711-715.

- Hoeschele I (1988) Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. *Theor Appl Genet* 76: 81-92.
- Karlin S (1978) Theoretical aspects of multi-locus selection balance I. In: Levin SA (ed) *Studies in Mathematical Biology, Volume 16, Part II: Populations and Communities, Studies in Mathematics*. The Mathematical Association of America, Washington D. C.
- Laterrot H (1975) Localisation chromosomique de I_2 chez la tomate contrôlant la résistance au pathotype 2 de *Fusarium oxysporum f. Lycopersici*. *Ann Amélior Plantes* 26: 485-491.
- Pelham J (1968) Disturbed segregation of genes on chromosome 9 – Gamete promoter, Gp, a new gene. *Rept Tomato Genet Coop* 18: 27-28.
- Pelham J (1970) More information on Gp. *Rept 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.
- Schwager SJ, Mutschler MA, Federer WT, Scully BT (1993) The effect of linkage on sample size determination for multiple trait selection. *Theor Appl Genet* (in press).
- Scully BT, Federer WT (1991) 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, OR.
- Sedcole JR (1977) Number of plants necessary to recover a trait. *Crop Science* 17: 667-668.
- Shyu SF (1992) Minimum progeny sample size determination for a specified progeny two-locus genotype. MS thesis. Cornell University, Ithaca, NY 14853.
- Weir BS (1979) Inferences about linkage disequilibrium. *Biometrics* 35: 235-254.
- Weir BS, Cockerham CC (1989) Complete characterization of disequilibrium at two loci. In: Feldman MW (ed) *Mathematical Evolutionary Theory*. Princeton University Press, Princeton, New Jersey.

Table 1. Suggested parent mating patterns of genotypic groups to obtain different desired progeny genotypes at one locus

Desired Progeny Genotype	Dominance Level	
	Complete	Incomplete
AA	A- × A-	AA × AA
Aa	A- × aa	AA × aa
aa	aa × aa	aa × aa

Table 2. Suggested parent mating patterns of genotypic groups to obtain different desired progeny genotypes at two loci

Desired Progeny Genotype	Dominance Level		
	A completely dominant B completely dominant	A incompletely dominant B completely dominant	A incompletely dominant B incompletely dominant
AABB	A-B- × A-B-	AAB- × AAB-	AABB × AABB
AaBB	A-B- × aaB-	AAB- × aaB-	AABB × aaBB
aaBB	aaB- × aaB-	aaB- × aaB-	aaBB × aaBB
AABb	A-B- × A-bb	AAB- × AAbb	AABB × AAbb
AaBb	A-B- × aabb A-bb × aaB-	AAB- × aabb AAbb × aaB-	AABB × aabb AAbb × aaBB
aaBb	aaB- × aabb	aaB- × aabb	aaBB × aabb
AAbb	A-bb × A-bb	AAbb × AAbb	AAbb × AAbb
Aabb	A-bb × aabb	AAbb × aabb	AAbb × aabb
aabb	aabb × aabb	aabb × aabb	aabb × aabb

Table 3. Right classification accuracy and a misclassification component of assigning an individual with trait expression $X \geq t_1$ into the AA genotypic group at a QTL of partial dominance, with equal residual variance σ^2 , $\mu_{AA} = 2\sigma$, $\mu_{aa} = -2\sigma$, $f_A = 0.6$ and $f_a = 0.4$.

$\theta(\text{AA} X \geq t_1)$ $\theta(\text{Aa} X \geq t_1)$	μ_{Aa}			
	0.0 σ	0.5 σ	1.0 σ	1.5 σ
-1.0 σ	0.4558	0.4317	0.4210	0.4171
	0.5120	0.5378	0.5493	0.5534
-0.5 σ	0.5108	0.4632	0.4382	0.4272
	0.4739	0.5229	0.5487	0.5601
0.0 σ	0.5908	0.5118	0.4633	0.4379
	0.4031	0.4829	0.5319	0.5576
0.5 σ	0.6926	0.5823	0.5023	0.4535
	0.3053	0.4160	0.4962	0.5452
1.0 σ	0.7986	0.6713	0.5577	0.4770
	0.2008	0.3282	0.4419	0.5227
1.5 σ	0.8858	0.7656	0.6269	0.5091
	0.1141	0.2342	0.3730	0.4908
t_1 2.0 σ	0.9428	0.8488	0.7027	0.5486
	0.0572	0.1512	0.2973	0.4514
2.5 σ	0.9739	0.9105	0.7760	0.5933
	0.0261	0.0895	0.2240	0.4067
3.0 σ	0.9888	0.9504	0.8395	0.6404
	0.0112	0.0496	0.1605	0.3596
3.5 σ	0.9954	0.9738	0.8897	0.6877
	0.0046	0.0262	0.1103	0.3123
4.0 σ	0.9981	0.9865	0.9267	0.7332
	0.0019	0.0135	0.0733	0.2668
4.5 σ	0.9993	0.9932	0.9524	0.7753
	0.0007	0.0068	0.0476	0.2247
5.0 σ	0.9997	0.9967	0.9697	0.8132
	0.0003	0.0033	0.0303	0.1868

Table 4. Left classification accuracy and a misclassification component of assigning an individual with trait expression $X \leq t_2$ into the aa genotypic group at a QTL of partial dominance, with equal residual variance σ^2 , $\mu_{AA} = 2\sigma$, $\mu_{aa} = -2\sigma$, $f_A = 0.6$ and $f_a = 0.4$.

$\theta(\text{aa} X \leq t_2)$	μ_{Aa}			
	0.0σ	0.5σ	1.0σ	1.5σ
$\theta(\text{Aa} X \leq t_2)$	0.9994	1.0000	1.0000	1.0000
	0.0006	0.0000	0.0000	0.0000
-5.0σ	0.9984	0.9999	1.0000	1.0000
	0.0016	0.0001	0.0000	0.0000
-4.5σ	0.9958	0.9996	1.0000	1.0000
	0.0042	0.0004	0.0000	0.0000
-4.0σ	0.9897	0.9986	0.9998	1.0000
	0.0103	0.0014	0.0002	0.0000
-3.5σ	0.9751	0.9956	0.9994	0.9999
	0.0249	0.0044	0.0006	0.0001
-3.0σ	0.9430	0.9870	0.9977	0.9997
	0.0569	0.0130	0.0023	0.0003
t_2 -2.0σ	0.8798	0.9639	0.9918	0.9985
	0.1201	0.0359	0.0080	0.0014
-1.5σ	0.7748	0.9095	0.9730	0.9934
	0.2246	0.0898	0.0262	0.0058
-1.0σ	0.6372	0.8053	0.9219	0.9749
	0.3605	0.1918	0.0748	0.0216
-0.5σ	0.4983	0.6557	0.8132	0.9190
	0.4942	0.3344	0.1746	0.0672
0.0σ	0.3865	0.5001	0.6496	0.7952
	0.5933	0.4737	0.3164	0.1631
0.5σ	0.3088	0.3759	0.4802	0.6134
	0.6445	0.5673	0.4472	0.2938
1.0σ	0.2574	0.2912	0.3497	0.4378
	0.6506	0.6048	0.5253	0.4058

Table 5. Center classification accuracy and a misclassification component of assigning an individual with trait expression X such that $t_2 < X < t_1$ into the Aa genotypic group at a QTL of partial dominance, with equal residual variance σ^2 , $\mu_{AA} = 2\sigma$, $\mu_{aa} = -2\sigma$, μ_{Aa} is expressed as μ , $f_A = 0.6$ and $f_a = 0.4$.

$\theta(Aa t_2 < X < t_1)$		t_1						
$\theta(AA t_2 < X < t_1)$		$\mu+0.5\sigma$	$\mu+1.0\sigma$	$\mu+1.5\sigma$	$\mu+2.0\sigma$	$\mu+2.5\sigma$	$\mu+3.0\sigma$	
0.0σ	μ							
	t_2							
	$\mu-3.0\sigma$	0.6775	0.6780	0.6455	0.5982	0.5540	0.5225	
		0.0492	0.0960	0.1603	0.2299	0.2895	0.3306	
	$\mu-2.5\sigma$	0.7110	0.7053	0.6675	0.6159	0.5687	0.5353	
		0.0520	0.1005	0.1666	0.2379	0.2986	0.3404	
	$\mu-2.0\sigma$	0.7570	0.7416	0.6958	0.6380	0.5863	0.5503	
		0.0567	0.1078	0.1768	0.2506	0.3131	0.3558	
	$\mu-1.5\sigma$	0.8056	0.7778	0.7218	0.6559	0.5987	0.5595	
		0.0644	0.1193	0.1926	0.2700	0.3348	0.3788	
	$\mu-1.0\sigma$	0.8421	0.8002	0.7323	0.6573	0.5941	0.5516	
		0.0776	0.1383	0.2178	0.3003	0.3682	0.4137	
	$\mu-0.5\sigma$	0.8537	0.7965	0.7150	0.6301	0.5610	0.5155	
		0.1013	0.1709	0.2596	0.3489	0.4208	0.4679	
	0.5σ	μ						
		t_2						
$\mu-3.0\sigma$		0.6641	0.6453	0.6061	0.5657	0.5353	0.5174	
		0.1145	0.1778	0.2439	0.3006	0.3404	0.3631	
$\mu-2.5\sigma$		0.7061	0.6773	0.6312	0.5863	0.5532	0.5339	
		0.1226	0.1876	0.2553	0.3131	0.3534	0.3765	
$\mu-2.0\sigma$		0.7514	0.7102	0.6559	0.6057	0.5696	0.5487	
		0.1335	0.2006	0.2700	0.3290	0.3701	0.3934	
$\mu-1.5\sigma$		0.7857	0.7323	0.6699	0.6148	0.5758	0.5535	
		0.1484	0.2178	0.2892	0.3495	0.3913	0.4151	
$\mu-1.0\sigma$		0.7965	0.7328	0.6636	0.6042	0.5629	0.5393	
		0.1709	0.2434	0.3173	0.3793	0.4221	0.4464	
$\mu-0.5\sigma$		0.7783	0.7060	0.6309	0.5678	0.5244	0.4999	
		0.2072	0.2840	0.3615	0.4258	0.4698	0.4946	

(continued)

Table 5. (continued)

$\theta(Aa t_2 < X < t_1)$		t_1					
$\theta(AA t_2 < X < t_1)$		$\mu+0.5\sigma$	$\mu+1.0\sigma$	$\mu+1.5\sigma$	$\mu+2.0\sigma$	$\mu+2.5\sigma$	$\mu+3.0\sigma$
1.0σ	μ						
	t_2						
	$\mu-3.0\sigma$	0.6342	0.6080	0.5762	0.5503	0.5339	0.5258
		0.2126	0.2714	0.3207	0.3558	0.3765	0.3864
	$\mu-2.5\sigma$	0.6723	0.6361	0.5987	0.5696	0.5517	0.5429
		0.2269	0.2855	0.3348	0.3701	0.3909	0.4008
	$\mu-2.0\sigma$	0.7025	0.6573	0.6148	0.5829	0.5636	0.5543
		0.2420	0.3003	0.3495	0.3848	0.4057	0.4157
	$\mu-1.5\sigma$	0.7150	0.6636	0.6177	0.5840	0.5637	0.5539
		0.2596	0.3173	0.3664	0.4017	0.4228	0.4329
	$\mu-1.0\sigma$	0.7060	0.6513	0.6034	0.5684	0.5474	0.5373
		0.2840	0.3415	0.3907	0.4263	0.4476	0.4579
	$\mu-0.5\sigma$	0.6763	0.6197	0.5703	0.5342	0.5125	0.5019
		0.3202	0.3779	0.4278	0.4641	0.4860	0.4966
1.5σ	μ						
	t_2						
	$\mu-3.0\sigma$	0.5910	0.5748	0.5595	0.5487	0.5429	0.5405
		0.3210	0.3548	0.3788	0.3934	0.4008	0.4038
	$\mu-2.5\sigma$	0.6162	0.5941	0.5758	0.5636	0.5572	0.5545
		0.3363	0.3682	0.3913	0.4057	0.4130	0.4159
	$\mu-2.0\sigma$	0.6301	0.6042	0.5840	0.5709	0.5640	0.5612
		0.3489	0.3793	0.4017	0.4158	0.4230	0.4260
	$\mu-1.5\sigma$	0.6309	0.6034	0.5823	0.5687	0.5617	0.5588
		0.3615	0.3907	0.4126	0.4265	0.4337	0.4367
	$\mu-1.0\sigma$	0.6197	0.5920	0.5706	0.5567	0.5494	0.5464
		0.3779	0.4062	0.4279	0.4419	0.4492	0.4522
	$\mu-0.5\sigma$	0.5989	0.5712	0.5493	0.5349	0.5273	0.5240
		0.4004	0.4284	0.4503	0.4676	0.4724	0.4756

Table 6. Expected conditional fractions of all two-locus genotypes in the genotypic group combination of doubly or singly complete dominance in linkage equilibrium

Expected Conditional Fractions $f(gt_i gg)$	Genotypic Group Combination (gg)			
	A-B-	AAB-	AaB-	aaB-
Genotypes(gt_i)				
AABB	$\frac{f_A f_B}{(1+f_a)(1+f_b)}$	$\frac{f_B}{1+f_b}$	0	0
AaBB	$\frac{2f_a f_B}{(1+f_a)(1+f_b)}$	0	$\frac{f_B}{1+f_b}$	0
aaBB	0	0	0	$\frac{f_B}{1+f_b}$
AABb	$\frac{2f_A f_b}{(1+f_a)(1+f_b)}$	$\frac{2f_b}{1+f_b}$	0	0
AaBb	$\frac{4f_a f_b}{(1+f_a)(1+f_b)}$	0	$\frac{2f_b}{1+f_b}$	0
aaBb	0	0	0	$\frac{2f_b}{1+f_b}$

* Note: f_A , f_a , f_B and f_b are allelic frequencies of alleles A, a, B and b, respectively.

Table 7. Minimum sample size n needed to obtain at least m desired progeny with confidence level, $\alpha = 0.90$, and probability of success, $P(\xi)$

$P(\xi)$	m						
	1	5	10	20	30	40	50
0.40	5	18	33	61	89	116	143
0.45	4	16	29	54	79	103	126
0.50	4	14	26	48	70	92	113
0.55	3	13	23	44	63	83	102
0.60	3	11	21	40	58	75	93
0.65	3	10	19	36	53	69	85
0.70	2	9	18	33	49	64	79
0.75	2	9	16	31	45	59	73
0.80	2	8	15	28	42	55	68
0.85	2	7	14	26	39	51	63
0.90	1	7	13	24	36	47	59
0.95	1	6	12	22	33	44	55
0.99	1	5	10	21	31	41	51

Table 8. Minimum sample size n needed to obtain at least m desired progeny with confidence level, $\alpha = 0.95$, and probability of success, $P(\xi)$

P(ξ)	m						
	1	5	10	20	30	40	50
0.40	6	21	36	65	94	121	149
0.45	6	18	32	57	82	107	131
0.50	5	16	28	51	74	96	117
0.55	4	14	25	46	66	86	106
0.60	4	13	23	42	60	78	96
0.65	3	12	21	38	55	72	88
0.70	3	10	19	35	50	66	81
0.75	3	9	17	32	46	61	75
0.80	2	9	16	29	43	56	69
0.85	2	8	14	27	40	52	64
0.90	2	7	13	25	37	48	60
0.95	1	6	12	23	34	45	56
0.99	1	5	11	21	31	42	52

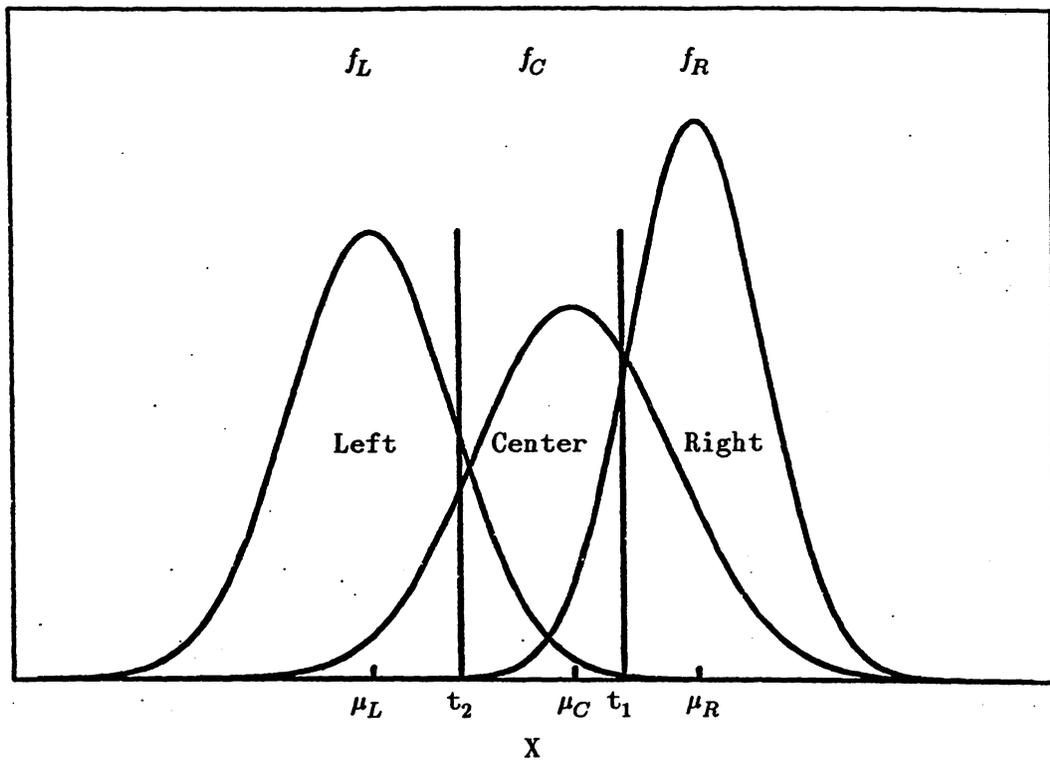


Figure 1. Quantitative trait expressions of three genotypic groups at a locus