Theor Appl Genet (1990) 79:775–780

An analytical model for the estimation of chromosome substitution effects in the offspring of individuals heterozygous at a segregating marker locus

M. R. Dentine and C. M. Cowan

Department of Dairy Science, University of Wisconsin, Madison, WI 53706, USA

Received August 1, 1989; Accepted December 20, 1989
Communicated by E. J. Eisen

Summary. Use of marker genes for quantitative traits has been suggested as a supplement to selection for livestock species. Linkage relationships can be estimated by using data from offspring of a heterozygous parent, if offspring can be positively assigned segregation of one or the other of the marker alleles. In field data, some data on offspring can be characterized and used to estimate the difference in chromosome substitution effects, but other matings result in uncertain transfer of the marker alleles. In this study, an alternative estimation procedure is proposed that would allow incorporation of data on all offspring of a heterozygous parent, even those where chromosome segregation is ambiguous. If the frequency of the marker alleles is known in the population of mates of a heterozygous individual, the mean and variance of the heterozygous offspring can be used in a generalized least-squares model to estimate the chromosome substitution effect. When gene frequencies are not known, maximum likelihood estimates can be obtained from the data for use in a conditional estimate. Monte Carlo simulations of data following the assumed genetic model were analyzed as proposed, and parameter estimates were characterized. Estimates of chromosome substitution effects were reasonable approximations of input values. Distributions of *t*-statistics testing the null hypothesis of no difference between marked chromosome segments were unbiased, with only slightly larger variance than expected. Addition of data from heterozygous offspring improved the efficiency of detection of chromosome substitution effects by more than four times when marker gene frequencies were low.

Key words: Major genes – Marker-assisted selection – Quantitative traits

Introduction

Recent biochemical techniques have enhanced the ability to characterize individual loci and have renewed interest in finding individual loci with large quantitative effects. Strategies to screen the genome for quantitatively important loci have been proposed that contrast individuals of similar background genotypes, to remove the effects of linkage disequilibrium and to isolate the effects of a single locus (Thoday 1967). Some of these have been successful in plants (Weller 1986) and have involved crosses of lines homozygous for different marker alleles and the analysis of backcross or F_2 offspring (Nienhuis et al. 1987; Weller 1987; Young and Tanksley 1989).

These schemes have been used in laboratory animal species (Kluge and Geldermann 1982; Spickett and Thoday 1966) but are less useful in livestock species, due to time and cost considerations of making specific crosses that may not produce commercially useful animals. Additionally, livestock lines are rarely homozygous for neutral marker genes, and even lines with distinctly different selection objectives, such as dairy and beef cattle, have intermediate gene frequencies (Cowan et al. 1989; Geldermann et al. 1985; Hallermann et al. 1988). Linkage disequilibrium in field data can lead to apparent correlations of marker genes with quantitative traits that may be sample-dependent and thus not useful for later selections.

Most quantitative genetic information on livestock species has been gained by analysis of field data under actual production management and is somewhat retrospective. Several strategies have been proposed to incorporate this approach using data analysis of field records in the search for genes of large effect (Famula 1986; Hoeschele 1998 a). Smith and Simpson (1986) pointed out that the search for quantitative trait loci of moderate effects must be accomplished within family for popula-

tions segregating for marker loci. One possibility has been the use of individuals heterozygous for unique or rare markers, where all offspring can be characterized for segregation of one or the other marker from the common parent. In cases where the alleles in the parent are not rare, some matings allow definitive segregation assignments while others do not contribute information. This concept for segregating markers was generalized by Soller and Beckmann (1988) who used data from "informative matings," although their strategy required crosses of lines with fixation at the quantitative trait locus.

For populations segregating at both marker and quantitative loci, selection is possible within half-sib families where the common parent is heterozygous at a marker locus. Differences in the effects of loci on the marked chromosome can be followed by the segregation of chromosomal segments containing the marker locus (Geldermann 1975; Stam 1986). The sum of all linked effects, corrected for the average amount of recombination at each locus, will be seen as a chromosome substitution effect (Geldemann 1975) detectable in the offspring. This approach does not require a single gene of major effect, but could encompass a fortuitous cluster of several genes of moderate effect close to the marker (Roberts and Smith 1982; Spickett and Thoday 1966; Stam 1986).

The number of individuals that must be characterized with respect to the marker and the quantitative trait has been calculated for each of these detection schemes. Cases where the segregation of the marker cannot be followed with certainty have not been utilized in the estimation procedures and represent extra costs, as additional individuals must be characterized although the data will not be used. One improvement has been suggested by Weller et al. (1988) that is particularly applicable to sex-limited traits such as milk production. The performance records of granddaughters are used for information on the quantitative trait, while the marker characterization is done on their sires (sons of the original heterozygous individual). This procedure reduces the number of individuals that must be characterized for the marker alleles, but requires an original grandsire heterozygous for at least one relatively rare allele, or the use of only homozygous sons in the analysis.

The objective of this paper is to investigate a method of analyzing data from offspring of a parent heterozygous for a set of segregating marker alleles that incorporates data on all progeny, including those with uncertain transfer from the common parent.

Materials and methods

The genetic model proposed includes the usual assumptions of quantitative inheritance but suggests that in some individuals, several clustered genes of moderate effect may create homologous chromosome segments with different genetic value within an individual. Suppose a detectable marker with two alleles is situated at locus L_0 surrounded by other quantitative loci L_1 to L_n on two homologous chromosomes designated as α and β within an individual. For an individual heterozygous at this marker locus, the chromosome substitution effect is defined as the difference between those offspring receiving the alternate chromosome segments (those marked with an A allele at L_0 versus those marked with the B allele) adjusted for recombination. Under random mating, the homologous chromosome segments inherited from the mates can be assumed to be of equal value for both types of offspring. Thus, the difference for a quantitative trait Y in offspring characterized as receiving the marker on the α chromosome from a common parent versus those receiving the β marker is the chromosome segment substitution effect $D_{(\alpha-\beta)}$, where

$$D_{(\alpha-\beta)} = \sum_{i=0}^{n} [L_i(\alpha) - L_i(\beta)]\,[1-2c_i]$$

and c_i is the average recombination fraction between L_0 and L_i. The marker locus may contribute to D or may serve only as a linked indicator of Mendelian segregation of a chromosome segment. Loci that are linked to the marker but homozygous within the individual will not contribute to D. Offspring of an AB-marked individual with an AA genotype of their own will have inherited the α segment from the common parent (Fig. 1) while those with BB inherited the β segment. The genetic variance of the trait within these individuals is due to all other quantitative genes not linked to this marker and is assumed homogenous, although the means differ.

Those offspring with an AB marker genotype (with expected frequency of 0.5) are of two types, with some receiving the α segment and some the β. If the population of mates segregating for the A and B alleles with the frequency of A set to p, the proportion of AB offspring that received the α segment from the sire will be $(1-p)$ under random mating. The mean of all AB individuals is then the weighted mean of the two types. The uncertainty of the segregation within the AB class of offspring

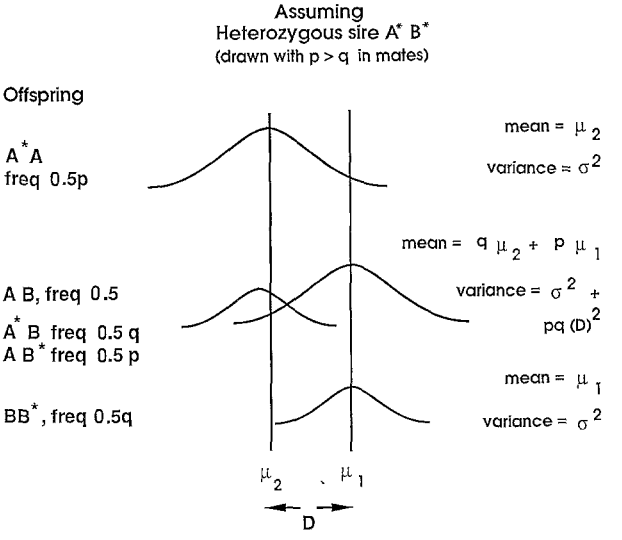

Fig. 1. Schematic frequency distribution of offspring of a sire heterozygous at a marker locus (AB) with a chromosome substitution effect of D. The sire is presumed mated to a population with an average chromosome substitution effect of zero. Figure is drawn with the frequency of A in mates (p) greater than frequency of B (q). Those alleles marked with (*) are of sire origin

contributes to the variance of the trait for these individuals. This increase is a function of the proportion of individuals of each type and the difference between the average breeding values. Following the model of Hoeschele (1988 b), the expected variance of the heterozygotes is increased by an additional $p(1-p)D^2$. Given the frequency p, the expected mean and variance of these AB individuals are known, although the exact chromosome inheritance is not, and these individuals can be incorporated into a statistical model to estimate the size of D. This strategy differs from the model employed by Hoeschele (1988 a), since the heterozygotes are not distinguishable and the probability of segregation from the sire is not modified dependent on the trait values within the sample.

Knowledge of $p(A)$ for a segregating marker may be available for a population for traits such as blood types, or can be estimated from the sample data set for new markers such as restriction fragment length polymorphisms. If the sample data is used, the maximum likelihood estimate under the assumption of a homogeneous population of mates and random mating can be obtained (Stevens 1938). The likelihood of N progeny with n_{AA} of one homozygous type and n_{BB} in the alternative homozygous class is proportional to $\left(\frac{p}{2}\right)_{n_{AA}}\left(\frac{1}{2}\right)_{N-n_{AA}-n_{BB}}\left(\frac{1-p}{2}\right)_{n_{BB}}$ with a maximum at

$$p(A)=\frac{n_{AA}}{(n_{AA}+n_{BB})}$$

For n_{AA} or n_{BB} equal to 0, the estimated frequency of A will be 0 or 1, respectively. An estimate of D can be obtained by solving the usual generalized least-squares equations for a linear model.

$Y_{ij}=\mu+\lambda_i(D)+g_{ij}+e_{ij}$, where

Y_{ij} = trait of offspring ij within genotype i;
μ = overall constant;
λ_i = independent variable representing the contrast of linkage relationships ($i=1$ to 3):
- $\lambda=1$ for daughters characterized as marker genotype $AA(i=1)$,
- $\lambda=0$ for daughters characterized as marker genotype $BB(i=2)$,
- $\lambda=$ gene frequency of the B allele $[1-p(A)]$ in the population of mates used for offspring characterized as marker genotype $AB(i=3)$. If p is unknown, the Maximum Likelihood (ML) estimate is substituted;

D = chromosome substitution effect of A versus B;
g_{ij} = quantitative breeding value of the offspring. For cases with a common sire, this includes the dam contribution and Mendelian sampling from the sire for all loci not linked to the markers in the sire;
e_{ij} = unexplained residual element within the marker genotypic class.

Since the residuals of the observations are not identically nor independently distributed and these animals may have additive relationships beyond a common parent, a generalized least squares is required rather than an ordinary LS. If the desired parameter estimates are μ and D contained in a vector $\hat{\boldsymbol{\beta}}$, and $\mathbf{X}$ is the incidence matrix for μ and D composed of column of 1's and λ_i's and $\mathbf{Y}$ is the vector of Y_{ij}, generalized least-squares estimates of β can be obtained by $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\,\mathbf{V}^{-1}\,\mathbf{X})^{-1}\,(\mathbf{X}'\,\mathbf{V}^{-1}\,\mathbf{Y})$ where $\mathbf{V}$ is the variance-covariance matrix for the observations described below, provided the inverse of $\mathbf{V}$ can be computed. Under the assumption that the Y's are distributed as a multivariate normal with variance within homozygous classes σ_w^2 the variance of $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\,\mathbf{V}^{-1}\,\mathbf{X})^{-1}\,\sigma_w^2$ and the null hypothesis of $D=0$ can be tested with the usual t-test for a generalized least-squares model.

The variance-covariance matrix ($\mathbf{V}$) is the sum of three elements: (1) The variance of the within-sire family residuals, which would be $\mathbf{I}\,\sigma_w^2$ for individuals with data with the same accuracy and no environmental correlations. The expected variance of these residuals is the variance within half-sibs due to all unlinked genes and unexplained residuals and equal to $3/4\,\sigma^2_{\text{additive}}+\sigma^2_{\text{residual}}$. If the heritability is assumed known, this variance can be specified. (2) The correlations between residuals due to additive relationships can be incorporated using the usual numerator relationships matrix (Henderson 1976). (3) The variance of the heterozygote observations around the weighted heterozygote mean will be increased by an amount $p\,q\,D^2$. Since the estimate of D will depend on the variance-covariance matrix which includes a function of D, an iterative solution will be necessary.

Simulations

A Monte Carlo simulation was programmed in Fortran generating pseudo-normal random numbers and binomial probabilities as data, according to the proposed model, with various marker gene frequencies for the mates of a heterozygous sire and varying D. Mates were assumed to be from a homogenous population with gene frequency $p(A)=0.1, 0.3, 0.5, 0.7$, or 0.9. The probability of the sire marker segregation was 0.5 for each allele. Residual effects were assumed to be independent and identically distributed as $N(0, \sigma_w^2)$.

The model generated a number of data sets from n individuals, which were then analyzed according to the proposed statistical model. Initial estimates of σ_w^2 were pooled estimates within homozygous subclasses. Distributions of t-tests of estimates of D from 2,500 populations of 25 or 50 offspring with an input D of 0 and $p=0.3$ were calculated. Gene frequencies of the marker allele were estimated from the sample data using Maximum Likelihood. The first estimate of D was the difference between trait means of the AA and BB subclasses. Subsequent iterations employed the estimates of σ_w^2 and D from the previous run until convergence was reached. Convergence criteria was a less than 0.1% change in $p\,q\,D^2/\sigma_w^2$ from consecutive rounds of iteration. Resulting parameter estimates were summarized. Empirical power curves for traits at intermediate gene frequencies ($p=0.3$ or 0.7) were constructed using 250 sample populations of 25, 50 or 100 offspring. The efficiency of estimation using the empirical standard errors of the estimates of D were compared to the standard errors from t-tests using the expected numbers in the homozygous classes.

Results

Estimates of the differences in chromosome substitution from Monte Carlo simulation using 500 samples of 50 progeny approximated input parameters (Table 1). Estimates of D showed some dependence on the gene frequency of the markers in the population of mates with closer estimates at unbalanced frequencies and underestimates at $p=0.5$ when $D>0$. The empirical variance of these estimates was higher for $p=0.5$ than for $p=0.1$ or $p=0.9$ ($F=1.58$, $P<0.001$), as expected from the increase in the $p\,q\,D^2$ term for the heterozygotes at intermediate gene frequencies.

The expected normal distribution around zero was seen for t-tests with population sizes of 25 and 50 with

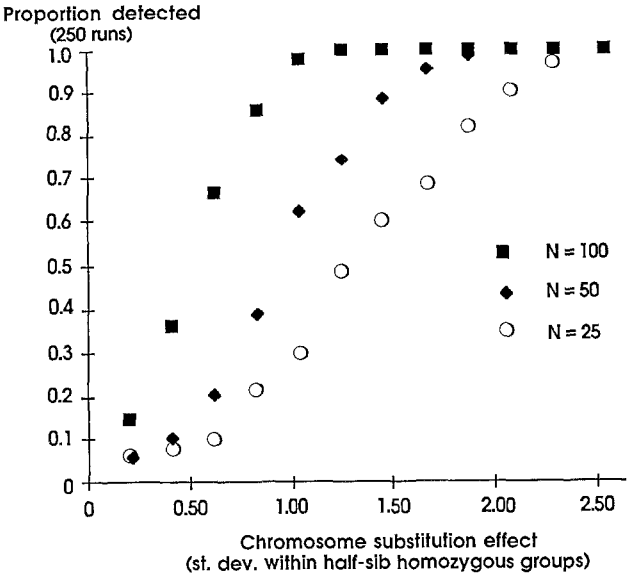

Fig. 2. Power curves for the detection of chromosome substitution effects from data on N offspring of a heterozygous sire for intermediate marker allele frequencies ($p = 0.3$ or 0.7) in mates

Table 1. Results of 500 simulated populations of 50 offspring with single records for a trait at varying gene frequencies (p) in the mates of a sire heterozygous for a marked chromosome pair with difference D for the quantitative trait[a]

D	$p = 0.1$	$p = 0.3$	$p = 0.5$	$p = 0.7$	$p = 0.9$
	Mean of estimated D				
0	0.0053	−0.0150	−0.0185	−0.0045	−0.0270
0.5	0.4884	0.5087	0.4771	0.4909	0.4788
1.0	1.0202	0.9545	0.9407	0.9731	1.0003
1.5	1.4966	1.4869	1.4362	1.4997	1.4877
	SD of estimated D				
0	0.2946	0.3643	0.4140	0.3540	0.3081
0.5	0.3143	0.3731	0.3884	0.3874	0.3136
1.0	0.3135	0.3620	0.3853	0.3820	0.3024
1.5	0.3320	0.3990	0.3889	0.3920	0.3235

[a] Measured in standard deviation units within homozygous individuals (σ_w)

means of the t-statistics of 0.026 and −0.014. The expected standard deviation for t with a population of 25 is 1.03 and for 50 it is 1.005. The results from the simulation were slightly larger at 1.08 and 1.04. The result of this slight increase in variance was that 6.3% of the populations of 25 and 5.1% of the populations of 50 had t-statistics above the usual $P < 0.05$ critical value for the appropriate degrees of freedom. The average number of iterations necessary per data set was 3.14, with a standard deviation of 0.73 for 250 data sets of 25 individuals.

Empirical power curves (Fig. 2) for intermediate gene frequencies showed that reasonable power of detection is available for smaller sample sizes than in other studies (Beckmann and Soller 1983). For D of one standard deviation within the homozygous offspring of a heterozygous parent, approximately 30% of the populations of 25, over 50% of the populations of 50, and nearly all of the populations of 100 had detectable chromosome substitution differences using this method of estimation.

Discussion

In general, the analytical model was successful at recovering the parameters of the data with the amount of variation expected in production traits. Estimates of chromosome substitution effects may be slightly underestimated at intermediate gene frequencies. This situation is preferable to overestimation, which might lead to false conclusions of major gene substitution effects. Results from hypothesis testing procedures were encouraging and resulted in only slightly increased variance of the usual t-statistic. This increase in variance of t could be corrected by using a slightly more conservative α level (such as $t_{n-1,0.01}$) for testing the null hypothesis of $D = 0$ with this estimation procedure.

Following the suggestion by Weller et al. 1988, the Y data on sons could be an estimated breeding value based on progeny resulting in an effective heritability in the range of 0.7 to 0.8 for traits such as milk yields. Sons would be genotyped for the marker information and yield information would be recorded on granddaughters. For these data, much smaller sample sizes could be used than previously proposed (Beckmann and Soller 1983, 1988; Smith and Simpson 1986).

The probability of finding polygenes with a favorable accumulation of genes of moderate effect has been discussed by Spicket and Thoday (1966) with respect to selection results in *Drosophila melanogaster*. Their selection experiments apparently capitalized on chance recombination events that arranged several favorable genes on a single chromosome. In their analyses, five alleles each of no more than 2 map units explained a large percentage of the difference accumulated between a selection line and a control. In lines of mice selected for growth rates (Salmon et al. 1988), several polymorphisms were fixed with different alleles in the selected population and in a control. These alleles were located near or within the growth-hormone gene and were detected using restriction fragment length polymorphisms and a growth hormone probe. In populations under selection, linkage disequilibrium introduced by the selection process may increase the probability of finding chromosome differences for quantitative traits. Although these genes would eventually become fixed with selection, use of marker genes to detect these chromosome segments while these genes are still segregating could accelerate the process.

Table 2. Comparison of variance of estimates ($\sigma_g^2 = 1$) from a fixed sample size ($n = 25$ or 50), using the proposed generalized model with Monte-Carlo simulation (500 populations, $h^2 = 0.3$) versus the expectation under expected distribution of offspring with a simple t-test of homozygous offspring of a heterozygous parent

	$p = 0.1$		$p = 0.3$		$p = 0.5$	
	50	25	50	25	50	25
Empirical variance of estimate of D from full model including AB	0.301	0.602	0.430	0.849	0.476	0.938
Variance of means [a] (difference in $AA-BB$)	1.46	2.93	0.631	1.26	0.530	1.06
Relative efficiency of model including AB offspring [b]	485%	487%	147%	148%	111%	113%

[a] Variance of (mean AA − mean of BB) $= \sigma^2_{\text{within}}/\xi(n_{AA}) + \sigma^2_{\text{within}}/\xi(n_{BB})$

[b] Relatively efficieny of estimates of D

$$= \frac{\text{variance of difference}_{(AA-BB)}}{\text{variance of estimate}_{(\text{full model})}} \times 100$$

Use of the heterozygote information improved the efficiency of the parameter estimates over methods that do not utilize data from these offspring (Table 2). The table gives the variance of the estimates from the conventional t-test by assuming that the offspring are in the expected frequencies even when these are not integers. In actual data, there would be variation in the number of homozygotes available for the contrast of the homozygous means and some cases where no individuals would be available for the contrast within one of the classes.

In field data, offspring or grandoffspring may be distributed across fixed effects such as herd or year and may share other relatives. Fixed effects besides μ may be added to the model, resulting in the usual Mixed Model equations (Henderson 1988) supplemented by the additional variance in the heterozygote. When offspring are nested within herds or years, previous comparisons would have required at least one of each homozygous group within a level of fixed effect. Since the probability of an offspring being heterozygous would be 50%, many more comparisons would be possible than the simple t-test, especially in cases of unbalanced gene frequencies. The proposed model would increase the number of comparisons by using data whenever at least two genotypes were represented.

The model allows incorporation of additional relationships due to other common relatives and allows the use of all genotyped relatives in the analysis. The $p\,q\,D^2$ term added to the variance of the heterozygotes is due to the uncertainty of potential Mendelian sampling. If some dams were genotyped for the marker, more heterozygous individuals could be assigned a sire chromosome transmission, and this $p\,q\,D^2$ term would not be added to those observations and λ would equal 0 or 1, indicating the known segregation.

If the same number of individuals were genotyped for the marker locus, the gain in efficiency from using information from the uncertain heterozygotes is considerable, the largest gains coming from cases with low frequencies of one of the markers in the mates. In addition, this model allows estimation when one of the homozygous classes is empty, which can occur quite often in small sample sizes or with low gene frequencies.

Acknowledgements. This work was partially funded by the Graduate School at the University of Wisconsin (Project No. 891584) and by a Consortium Grant from Atlantic Breeders Cooperative Inc., Noba Inc., Sire Power Inc., and 21st Century Genetics. The authors would like to recognize the programming assistance of J. Tai in writing the simulation program used.

References

Beckmann JS, Soller M (1983) Restriction fragment length polymorphisms in genetic improvement: methodologies, mapping, and costs. Theor Appl Genet 67:35–43

Beckmann JS, Soller M (1988) Detection of linkage between marker loci and affecting quantitative traits in crosses between segregation populations. Theor Appl Genet 76:228–236

Cowan CM, Dentine MR, Ax RL, Schuler LA (1989) Restriction fragment length polymorphisms associated with growth hormone and prolactin genes in holstein bulls: evidence for a novel growth hormone allele. Anim Genet 20:157–165

Famula TR (1986) Identifying single genes of large effect in quantitative traits using best linear unbiased prediction. J Anim Sci 63:68–76

Geldermann H (1975) Investigations on inheritance of quantitative characters in animals by gene markers. 1. Methods. Theor Appl Genet 46:319–330

Geldermann H, Pieper U, Roth B (1985) Effects of marked chromosome sections on milk performance in cattle. Theor Appl Genet 70:138–146

Hallerman EM, Nave A, Soller M, Beckmann JS (1988) Screening of Israeli holstein-friesian cattle for restriction fragment length polymorphisms using homologous and heterologous deoxyribonucleic acid probes. J Dairy Sci 71:3378–3389

Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–84

Henderson CR (1988) Theoretical basis and computational methods for a number of different animal models. J Dairy Sci 71 [Suppl 2]:1–16

Hoeschele I (1988 a) Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. Theor Appl Genet 76:81–92

Hoeschele I (1988 b) Statistical techniques for detection of major genes in animal breeding data. Theor Appl Genet 76:311–319

Kluge R, Geldermann H (1982) Effects of marked chromosome sections on quantitative traits in the mouse. Theor Appl Genet 62:1–4

Nienhuis J, Helentjaris T, Slocum M, Ruggero B, Schaefer A (1987) Restriction fragment length polymorphism analysis of loci associated with insect resistance in tomato. Crop Sci 27:797–803

Roberts RC, Smith C (1982) Genes with large effects: theoretical aspects in livestock breeding. 2nd World Congr Gene Appl Livestock Prod, Madrid, VI:420–438

Salmon RK, Berg RT, Yeh FC, Hodgetts RB (1988) Identification of a variant growth hormone haplotype in mice selected for high body weight. Genet Res 52:7–15

Smith C, Simpson SP (1986) The use of genetic polymorphisms in livestock improvement. J Anim Breed Genet 103:205–217

Soller M, Beckmann JS (1988) Genomic genetics and the utilization for breeding purpose of genetic variation between populations. Proc 2nd Int Conf Quant Genet, Raleigh/NC, pp 161–188

Spickett SG, Thoday JM (1966) Regular responses to selection. 3. Interaction between located polygenes. Genet Res Camb 7:96–121

Stam P (1986) The use of marker loci in selection for quantitative characters. In: Smith C, King JWB, McKay JC (eds) Exploiting new-technologies in animal breeding genetic developments. Oxford University Press, Oxford, pp 170–182

Stevens WL (1938) Estimation of blood-group frequencies. Ann Eugen 8:362–375

Thoday JM (1967) New insights into continuous variation. Proc 3rd Int Congr Hum Genet, Chicago/IL, pp 339–350

Weller JI (1986) Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic makers. Biometrics 42:627–640

Weller JI (1987) Mapping and analysis of quantitative trait loci in *Lycopersicon* (tomato) with the aid of genetic markers using approximate maximum likelihood methods. Heredity 59:413–421

Weller JI, Kashi Y, Soller M (1988) Estimation of the sample size necessary for genetic mapping of quantitative traits in dairy cattle using genetic markers. J Dairy Sci 71 [Suppl 1]:142

Young ND, Tanksley SD (1989) RFLP analysis of the size of chromosomal segments retained around the *Tm*-2 locus of tomato during backcross breeding. Theor Appl Genet 77:353–359