Theor Appl Genet (1985) 69: 553–558

Evaluation of biometrical methods for estimating the number of genes

1. Effect of sample size*

D. K. Mulitze** and R. J. Baker

Crop Development Centre, University of Saskatchewan, Saskatoon, Saskatchewan S7N 0WO, Canada

Received July 15, 1984; Accepted August 4, 1984
Communicated by A. L. Kahler

Summary. The effect of sample size on estimating the number of genes by the inbred-backcross and genotype assay procedures was investigated. Modifications were proposed for each procedure. Ninety-five percent confidence intervals for estimated numbers of genes and the minimum sample size required to discriminate between various genetic hypotheses were calculated for both procedures. Sample size had a greater impact on the estimation of gene number by the genotype assay procedure than by the inbred-backcross procedure, especially for small sample sizes. For the inbred-backcross procedure, the optimal number of backcrosses varied with the number of genes. Estimates of the number of genes are theoretically less reliable when estimated by the genotype assay procedure than by the inbred-backcross procedure, and are sensitive to the choice of assay generation. Generally, the inbred-backcross procedure is preferred. Even with the fulfillment of all genetic assumptions for each method and absence of error in measuring genotypic values, substantial upward or downward biases in the estimates of the number of genes are expected from both the inbred-backcross and the genotype assay procedures.

Key words: Inbred-backcross procedure – Genotype assay – Number of genes – Quantitative genetics

Introduction

A number of biometrical procedures have been developed for estimating the number of genes governing quantitative traits in autogamous diploids.

The method of moments (Castle 1921; Burton 1951; Wright 1968) was among the first procedures for analyzing the differences in a quantitative trait between two homozygous parents. Subsequent procedures included Mather's K_1 (Mather 1949), the partitioning method (Powers et al. 1950; Powers 1963), discriminant analysis (Weber 1959), the inbred-backcross procedure (Wehrhahn and Allard 1965), the convolution approach (Tan and Chang 1972), genotype assay (Jinks and Towey 1976; Towey and Jinks 1977), and the doubled haploid method (Choo and Reinbergs 1982). Methods vary in their assumptions, in time and resources required, and in precision and reliability of estimates of the number of genes.

The inbred-backcross procedure (Wehrhahn and Allard 1965) involves the production of inbred lines following several backcrosses to the recurrent parent and their subsequent classification in replicated field trials as different from or not different from the recurrent parent. Unless many genes govern the trait, most inbred-backcross lines are expected either to be genotypically identical to the recurrent parent or single gene deviates. The number of inbred-backcross lines expected to deviate from the recurrent parent by a specific gene is mq with a 95% confidence interval of approximately $mq \pm 2[mq(1-q)]^{1/2}$, where m is the number of inbred-backcross lines, $q = 1/2^{b+1}$, and b is the number of backcrosses.

The genotype assay procedure (Jinks and Towey 1976) requires an estimate of the proportion (P_h) of randomly selected F_n plants, derived from crossing two homozygous parents, that are heterozygous for at least one locus. The proportion P_h is estimated by testing for unequal means of F_{n+2} lines derived from two or more randomly selected F_{n+1} progeny of each assay plant. The minimum and maximum number of segregating loci is then estimated by equating the observed proportion, P_h, to theoretical expectations which are a function of the assay generation, number of lines derived from each F_n assay plant, number of loci, and two limiting genetic models. For their theoretical expectations, Jinks and Towey calculated the probability that a sample of F_{n+1} progeny from an F_n plant would not all have the same genotypic value. Because there was not a one-to-one correspondence between genotype and genotypic value, Jinks and Towey derived upper and lower limits for the proportion of assay plants whose progeny were expected to vary for genotypic values. Assuming unequal gene effects and no domi-

* Part of this study is based on the Ph.D. thesis of the senior author

** Present address: Cereal Improvement Program, International Centre for Agricultural Research in the Dry Areas (ICARDA), P.O. Box 5466, Aleppo, Syria nance, each F_{n+1} genotype would have a unique genotypic value, resulting in a maximum proportion (P_{max}) of assayed plants being classified as heterozygous. With equal effects and complete dominance, several different genoytpes may have the same genotypic value, resulting in a minimum proportion (P_{min}) of assay plants being classified as heterozygous. Towey and Jinks (1977) also derived expectation for two intermediate situations: (i) equal effects and no dominance ($P_{int.A}$), and (ii) unequal effects and complete dominance ($P_{int.B}$). Jinks and Towey (1976) used their P_{max} and P_{min} curves to give the minimum and maximum estimates of the number of genes segregating for various quantitative traits of *Nicotiana rustica.*

The two methods described above are similar in that they both require binary classification. Inbred-backcross lines are classified as either parental or non-parental; genotype assay plants are classified as either heterozygous or homozygous. Assumptions common to both procedures are that there is normal diploid meiosis, no linkage, no epistasis, and no selection. The purpose of this study was to consider a basic question concerning methodology; that is, does sample size have a significant impact on estimation of the number of genes by the inbred-backcross or genotype assay procedures?

Theory and results

Inbred-backcross procedure

Wehrhahn and Allard (1965) showed that the probability that an inbred-backcross line would deviate from the recurrent parent at any particular locus is $q = ½^{b+1}$, where b is the number of backcrosses used in developing the line. It follows that the proportion of inbred-backcross lines that deviate from the recurrent parent at one or more of k independent loci is expected to be

$$d = 1-(1-q)^k \tag{1}$$

This proportion of non-parental lines increases with the number of loci by which the donor and recurrent parents differ and decreases with the number of backcrosses (Fig. 1).

In evaluating a set of inbred-backcross lines, the number of loci carrying genes which govern the difference between parents can be estimated from the observed proportion ($\hat{d}$) of non-parental lines by

$$\hat{k} = \ln(1-\hat{d})/\ln(1-q). \tag{2}$$

With a normal approximation, the 95% confidence interval for $\hat{d}$, given d, becomes $d \pm 1.96[d(1-d)/m]^{½}$. Upon substituting $1-(1-q)^k$ for d, taking natural logarithms, and solving for $\hat{k}$, the 95% confidence interval for $\hat{k}$ becomes

$$\ln\{1-d-1.96[d(1-d)/m]^{½}\}/\ln(q) < \hat{k} < \ln\{1-d+1.96[d(1-d)/m]^{½}\}/\ln(q). \tag{3}$$

Ninety-five percent confidence intervals for estimated numbers of loci ($\hat{k}$), given k = 2 to 15 loci, were calculated for sample sizes of 75, 150 and 300 (Table 1). When the sample size was too small for the normal approximation to apply (Steel and Torrie 1980, Table 21.1), confidence limits for $\hat{d}$ were interpolated from the exact binomial distribution (Steel and Torrie 1980, Table A. 14B). Confidence limits for the estimated numbers of loci were rounded off to the nearest integer.

Confidence interval widths increased as k and b increased, and generally as sample size decreased (Table 1). Interval widths are a function of the slopes of the curves for the theoretical expectations (Fig. 1), increasing as the slope decreases.

Table 1. Ninety-five percent confidence intervals for estimated numbers of loci for three sample sizes in the inbred-backcross procedure

Sample size	Actual no. of loci				
	2	4	6	10	15
			(one backcross)		
75	1-3	3-5	4-8	8-13	–
150	2-3	3-5	5-7	8-13	–
300	2-2	3-5	5-7	8-12	12-19
			(two backcrosses)		
75	1-3	3-6	4-8	7-13	11-20
150	1-3	3-5	5-7	8-12	12-18
300	1-3	3-5	5-7	9-12	13-17
			(three backcrosses)		
75	1-4	2-6	4-9	7-14	11-20
150	1-3	2-6	4-8	8-13	12-19
300	1-3	3-5	5-7	8-12	13-17

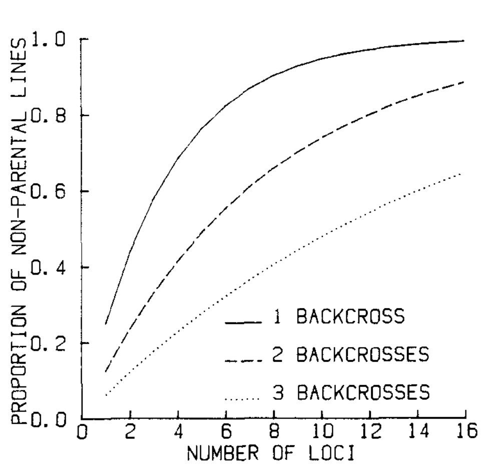

Fig. 1. Theoretical proportions (*d*) of non-parental inbred-backcross lines for one, two and three backcrosses

Sample size is also critical in trying to distinguish between various genetic hypotheses. A properly planned experiment might require, for example, a sample size sufficient to distinguish (with 95% certainty) between di- or tri-genic inheritance. With two backcrosses, the expected proportions of non-parental lines would be 0.2344 and 0.3301, respectively. Using the standard error method of Mather (1951), the minimum sample size required would be approximately

$$m = \left\{1.96[(r_1(1-r_1))^{1/2} + (r_2(1-r_2))^{1/2}]/(r_1-r_2)\right\}^2, \qquad (4)$$

where r_1 and r_2 are the expected proportions of non-parental lines. At least 335 inbred-backcross lines would be required to distinguish between di- and tri-genic inheritance with two backcrosses (Table 2). With one backcross, only 190 lines would be required to test the same hypothesis at 95% certainty.

The original proposal of Wehrhahn and Allard (1965) differs somewhat from that outlined above. Rather than estimating the number of genes from an estimate of the total proportion of non-parental lines, they recommended evaluation of distinct groups of inbred-backcross lines. Inbred-backcross lines will not always fall into distinct groups. Even when they do, one will have to decide, for example, if a particular group represents lines which deviate from the recurrent parent at a single locus or at either of two loci each carrying alleles with equal effect. It is important to be able to distinguish between such hypotheses if one uses the approach described by Wehrhahn and Allard. The arguments concerning sample size would be similar to those described above. For example, with two backcrosses, the proportion of lines that deviate from the recurrent parent at a particular locus is expected to be 0.125 while the proportion that deviate at one or the other of two loci having genes with equal effect is expected to be 0.219. From equation 4, one can estimate that 242 lines would be required to distinguish between the two alternatives. With three backcrosses, comparable proportions would be 0.062 and 0.133, and 263 inbred-backcross lines would be required to distinguish between one and two genes for any particular group of lines.

In using the total proportion of non-parental lines to estimate numbers of genes, one avoids the often subjective classification of lines into distinct groups while sacrificing the opportunity to estimate the effects of individual genes if lines do fall into distinct groups (Wehrhahn and Allard 1965). However, it seems clear that consideration of sample size is going to be similar in either approach. Large samples will be required to give reasonably precise estimates of numbers of genes and to distinguish between moderately simple genetic hypotheses concerning all non-parental lines or groups of non-parental lines. With few genes, more precise estimates may be derived with one backcross. As the number of genes increases, it will be necessary to increase the number of backcrosses in order to increase the precision of estimates.

Table 2. Theoretical minimum numbers of inbred-backcross lines required to be 95 percent certain of distinguishing between various numbers of loci using the inbred-backcross procedure

No. of loci	No. of backcrosses		
	1	2	3
2 vs. 3	190	335	637
2 vs. 4	56	100	189
2 vs. 5	31	51	96
3 vs. 4	317	508	927
3 vs. 5	95	147	264
3 vs. 6	50	74	131
4 vs. 5	487	704	1,235
4 vs. 6	144	199	344
4 vs. 7	75	99	168
5 vs. 6	713	930	1,563
5 vs. 7	209	260	430
5 vs. 8	108	127	208
5 vs. 10	50	54	86
10 vs. 12	1,044	714	949
10 vs. 15	249	143	178
12 vs. 15	966	481	567

Genotype assay

In developing theoretical expectations, Jinks and Towey (1976) made the implicit assumption that only F_{n+1} progeny with unequal genotypic values would give rise to F_{n+2} lines with unequal means. This is incorrect, for example, in the case of complete dominance where some heterozygotes might give rise to grandprogeny lines with unequal means even though the F_{n+1} plants have the same genotypic values. Assuming a two-locus model with equal effects and complete dominance (P_{min}, Jinks and Towey 1976), progeny with genotypes AABB, AABb, AaBB, and aaBb have the same genotypic value but give rise to F_{n+2} lines with three different means (Table 3). As the number of distinct F_{n+2} grandprogeny means decreases, the probability of sampling two or more lines with unequal means, as well as the expected proportion of heterozygotes, also decreases. From Table 3, it is apparent that a genetic model with equal additive effects ($P_{int.A}$) actually gives the minimum expected proportion. Varying levels of complete or incomplete dominance would increase the number of different

Table 3. Expected F_{n+2} grandprogeny means for genotype assay under four genetic models

Genotype of F_{n+1} plant	Genetic model			
	Additive unequal (P_{max})	Dominant unequal ($P_{int.B}$)	Additive equal ($P_{int.A}$)	Dominant equal (P_{min})
AABB	$2a_1+2a_2$[a]	$2.0a_1+2.0a_2$	4a	4.0a
AABb	$2a_1+a_2$	$2.0a_1+1.5a_2$	3a	3.5a
AAbb	$2a_1$	$2.0a_1$	2a	2.0a
AaBB	a_1+2a_2	$1.5a_1+2.0a_2$	3a	3.5a
AaBb	a_1+a_2	$1.5a_1+1.5a_2$	2a	3.0a
Aabb	a_1	$1.5a_1$	a	1.5a
aaBB	$2a_2$	$2.0a_2$	2a	2.0a
aaBb	a_2	$1.5a_2$	a	1.5a
aabb	0	0	0	0
No. of distinct F_{n+2} genotypic means	9	9	5	6

[a] Allelic substitution effects at locus A-a and B-b designated by a_1 and a_2, respectively, and by a when effects at both loci are equal

grandprogeny means and the expected proportion would approach P_{max}.

To show that P_{min} and $P_{int.B}$ are always intermediate between P_{max} and $P_{int.A}$, it is necessary to revise the formulae for P_{min} (Jinks and Towey 1976) and $P_{int.B}$ (Towey and Jinks 1977). For an F_n plant heterozygous at r = 1, 2, ... k independent loci, the probability of choosing p grandprogeny lines with equal means is

$$P_r^* = \sum_{t=0}^{r} \sum_{s=0}^{t} \left\{ \frac{r!}{s!\,(t-s)!\,(r-t)!}\, 2^{t-s-2r} \right\}^p , \qquad (5)$$

where t is the number of loci in the heterozygous or homozygous desirable allelic phases and s is the number in the homozygous desirable phase in the F_{n+1} plant. The proportion of detectable F_n heterozygotes then becomes

$$P_{min} = 2^{k-kn} \sum_{r=0}^{k} {}^kC_r\, (2^{n-1}-1)^{k-r} (1-P_r^*). \qquad (6)$$

For $P_{int.B}$, a FORTRAN program was written to calculate the revised expected proportions by computing the number and probabilities of all possible grandprogeny genotypic means from assay plants heterozygous at r = 1, 2, ... k loci, assuming unequal effects and complete dominance. Revised $P_{int.B}$ proportions were equal to, or only slightly less than, P_{max} with greater divergence as the number of loci and the assay generation increased.

Corrected P_{min} theoretical expectations were approximately midway between those of $P_{int.A}$ and P_{max} (Fig. 2). As the assay generation was delayed, all theoretical expectations decreased and became increasingly convergent. With assay in F_5 or later generations, the curves become so convergent as to result in essentially point estimates of the number of genes (Fig. 2). Increasing the number of F_{n+1} progeny sampled would result in detection of even more heterozygotes and greater convergence of the curves.

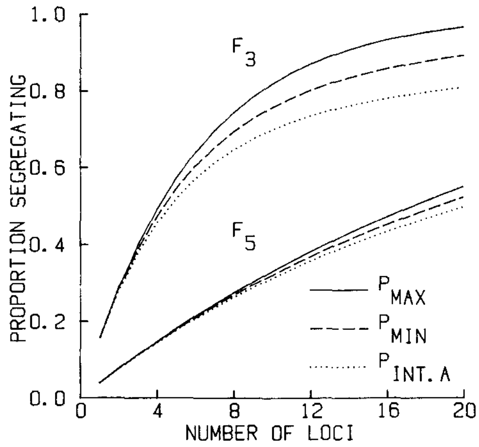

Fig. 2. Genotype assay probability curves for P_{max}, P_{min} (corrected), and $P_{int.A}$ with p = 2 F_{n+1}-derived F_{n+2} lines per assay plant

Ninety-five percent confidence intervals for the estimated number of loci were constructed to assess the impact of sample size on estimation by genotype assay. Since, in practice, one cannot be certain of the true genetic model, confidence limits were constructed from both the P_{max} and $P_{int.A}$ probability curves. The lower limit was taken as the lower limit for P_{max} while the upper limit was taken as the upper limit for $P_{int.A}$. For the lower limit, let $q = P_{max} = 1 - z^k$ where $z = 1 + [(2^p + 2 - 4^p)/(2^{n+2p-1})]$ (Towey and Jinks 1977). The

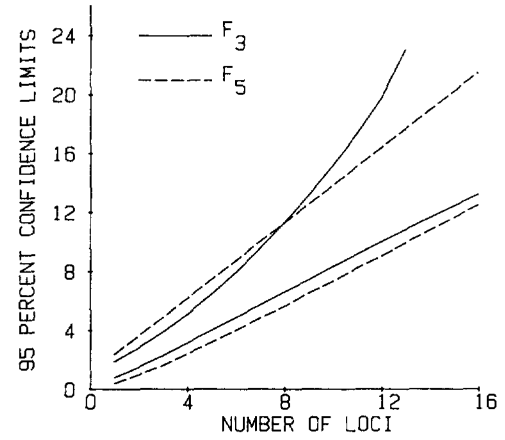

Fig. 3. Ninety-five percent confidence limits for the estimated number of loci for samples of 150 F_3 or F_5 genotype assay plants ($p=2$)

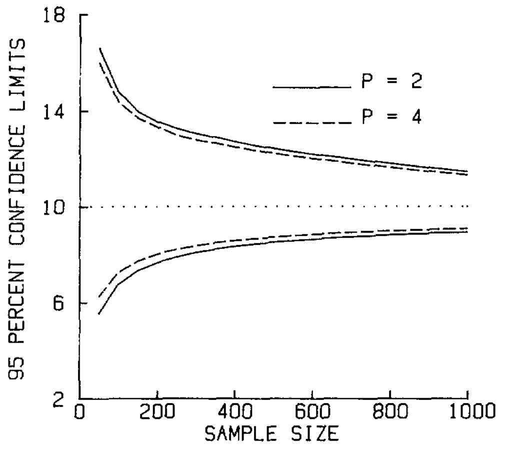

Fig. 4. Ninety-five percent confidence limits for various sample sizes of F_5 assay plants, assuming ten segregating loci and $p=2$ or 4 F_6-derived F_7 lines per assay plant

expected variance of q is q (1–q)/m where m is the number of F_n assay plants. By assuming a P_{max} model, using a normal approximation to the binomial distribution, and solving for k, the lower 95% confidence limit becomes $\ln\{1-q-1.96[q(1-q)/m]^{1/2}\}/\ln(z)$. The upper limit was calculated using the $P_{int.A}$ proportions, interpolation of the binomial confidence limit (Steel and Torrie 1980, Table A. 14 B), and reference back to the $P_{int.A}$ probability graph.

Confidence interval widths increased as the number of loci increased and as the slope of the original P_{max} and $P_{int.A}$ curves decreased (Figs. 2 and 3). The upper confidence limit showed increasing slope with increasing numbers of loci, particularly in early generations. Interval widths were smaller in F_3 than in F_5 assay generations up to about ten loci (Fig. 3). Interval widths decreased with increased sample size, especially when sample sizes were increased from 50 assay plants (Fig. 4). Doubling the number of grandprogeny lines per assay plant (p) did not decrease interval widths as much as doubling sample size (Fig. 4). Interval widths with $p=30$ were only slightly smaller than with $p=4$. Interval widths were most notably affected by generation of assay, sample size, and number of loci.

Minimum sample sizes required to discriminate between two genotype assay proportions expected for different numbers of loci also were calculated. Two expected P_{max} or $P_{int.A}$ proportions were substituted for r_1 and r_2 in equation (4). At least 277 F_3-derived families (with $p=2$) would be required to discriminate with 95% certainty between di- and tri-genic inheritance under a P_{max} genetic model. For a $P_{int.A}$ model, 349 plants would have to be assayed and, in the F_5, over 1,000 plants would have to be assayed for either model (Table 4). Minimum required sample sizes under $P_{int.A}$ exceeded those required under P_{max} and were greater in the F_5 than in the F_3 unless the hypotheses involved greater numbers of loci (Table 4).

Table 4. Theoretical minimum required sample sizes for 95% certainty of discriminating between genotype assay proportions expected for various numbers of loci under two genetic models in F_3 and F_5

No. of loci	F_3		F_5	
	P_{max}	$P_{int.A}$	P_{max}	$P_{int.A}$
2 vs. 3	277	349	1,008	1,051
2 vs. 4	83	111	302	315
2 vs. 5	43	61	151	160
3 vs. 4	426	610	1,493	1,561
3 vs. 5	124	187	409	439
3 vs. 6	63	100	202	220
4 vs. 5	609	963	1,815	2,010
4 vs. 6	173	292	514	570
4 vs. 7	86	153	256	284
5 vs. 6	814	1,461	2,368	2,632
5 vs. 7	228	433	662	739
5 vs. 8	114	228	313	358
5 vs. 10	50	110	128	150
10 vs. 12	715	2,406	1,347	1,682
10 vs. 15	149	542	246	324
12 vs. 15	519	1,974	757	1,042

Discussion

Sampling variance can cause substantial upward or downward biases in the number of genes estimated by the inbred-backcross or genotype assay procedures. For the inbred-backcross procedure, increased sample size would reduce the confidence interval widths. However, the number of backcrosses would also have a significant effect, especially on the minimum sample size required to test genetic hypotheses about the number of genes. Researchers using the inbred-backcross procedure commonly make two backcrosses. This may not be the optimum strategy if, in fact, relatively few genes are involved and if conclusions are based upon the estimated proportion of non-parental lines. In this case, one backcross may be sufficient and would certainly require less time and effort.

The corrected P_{min} results in a smaller interval between the minimum and maximum estimate of the number of genes using the genotype assay procedure. In later generations, this results in essentially a point estimate of the number of genes. For example, the estimate of 7 to 12 genes controlling flowering time in *Nicotiana rustica* (Towey and Jinks 1976) becomes an estimate of 7 with the new formula. However, even with this improvement, estimates of gene numbers are still quite imprecise when sample size is taken into consideration. Because of the large sampling variances associated with estimates, there is some question about the reliability of estimates given by Jinks and Towey (1976) and Towey and Jinks (1977). They assayed from 18 to 80 plants in the F_2 to the F_6 generations.

Selection of an optimal generation for genotype assay is difficult because of the significant effects of generation and number of loci on expected widths of confidence intervals. Different assay generations would be recommended if one knew a priori whether few or many loci were segregating. With four or fewer loci, one might assay the F_2 generation. As the actual number of loci increases, one would opt for later generations but not beyond the F_6 or F_7. Without a priori knowledge, one might best opt for an intermediate generation such as the F_4 as the generation to be assayed.

In comparing procedures, sample size appears to be a more critical factor for genotype assay than for the inbred-backcross procedure. For an equivalent increase in sample size, confidence intervals are shortened more for the inbred-backcross procedure than for genotype assay. For any given number of loci, the confidence interval widths and the minimum required sample sizes (Tables 2 and 4) for the inbred-backcross procedure with optimal number of backcrosses are less than those required for genotype assay with the optimal generation of assay. From the standpoint of sampling variance alone, one would prefer the inbred-backcross procedure over genotype assay.

This discussion assumes that all genetic assumptions pertaining to each method are satisfied. Linkage, epistasis, and any selection can add to the uncertainty associated with sampling variability. Furthermore, construction of confidence intervals assumed a perfect binary classification for each procedure. Either heritability approaching 100 percent or infallible statistical procedures are required if inbred-backcross lines and assay plants are to be correctly classified. Failing this, confidence intervals will be wider still. In practice, therefore, the uncertainty in estimating the number of genes by these two procedures probably exceeds the already substantial uncertainty due to sampling variance alone.

References

Burton GW (1951) Quantitative inheritance in pearl millet. Agron J 43:407–417

Castle WE (1921) An improved method of estimating the number of genetic factors concerned in case of blending inheritance. Science 54:223

Choo TM, Reinbergs E (1982) Estimation of the number of genes in doubled haploid populations of barley (*Hordeum vulgare*). Can J Genet Cytol 24:337–341

Jinks JL, Towey P (1976) Estimating the number of genes in a polygenic system by genotype assay. Heredity 37:69–81

Mather K (1949) Biometrical genetics. Methuen, London

Mather K (1951) The measurement of linkage in heredity. John Wiley and Sons, New York

Powers LR (1963) The partitioning method of genetic analysis and some aspects of its application to plant breeding. In: Hanson WD, Robinson HF (eds) Statistical genetics and plant breeding, No 982. NAS-NRC, Washington DC, pp 280–318

Powers LR, Locke LF, Garrett JC (1950) Partitioning method of genetic analysis applied to quantitative characters of tomato crosses. USDA Techn Bull 998

Tan WY, Chang WC (1972) Convolution approach to the genetic analysis of quantitative characters of self-fertilized populations. Biometrics 28:1073–1090

Towey P, Jinks JL (1977) Alternative ways of estimating the number of genes in a polygenic system by genotype assay. Heredity 39:399–410

Weber E (1959) The genetical analysis of characters with continuous variability on a Mendelian basis. 1. Monohybrid segregation. Genetics 44:1131–1139

Wehrhahn C, Alard RW (1965) The detection and measurement of the effects of individual genes involved in the inheritance of a quantitative character in wheat. Genetics 51:109–119

Wright S (1968) Evolution and the genetics of populations. Vol I. Genetic and biometric foundations. University of Chicago Press, Chicago Ill, pp 381–403