Theor Appl Genet (1984) 67:143–148

Methods for estimating gene numbers for quantitative characters using doubled haploid lines

J. W. Snape, A. J. Wright and E. Simpson
Plant Breeding Institute, Maris Lane, Trumpington, Cambridge CB2 2LQ, England

Received February 24, 1983; Accepted June 16, 1983
Communicated by A. Robertson

Summary. Three methods of estimating the numbers of genes segregating for quantitative characters using doubled haploid lines are presented. The first uses estimates of the range and genetical variance of an F_1 or F_2 derived population. The second adapts the genotype assay method of Jinks and Towey (1976) to F_2 derived lines. The third uses the variances of an F_2 derived population. Statistical problems of obtaining meaningful estimates using these methods are discussed and it is concluded that genotype assay is the best method for distinguishing between few and many genes. These methods are illustrated using data from an experiment containing doubled haploid lines of barley developed using the *H. bulbosum* system.

Key words: Doubled haploids – Gene numbers – Barley – Quantitative characters – *Hordeum bulbosum*

Introduction

Techniques for producing doubled haploid lines are now well developed in many plant species (Nitzsche and Wenzel 1977). The major use of such lines is in advancing segregating generations of breeding material to homozygosity thereby reducing the time necessary to produce a new variety and also increasing the efficiency of selection (Snape 1982). However, doubled haploid lines can also be powerful tools for genetical analysis, particularly for quantitative characters, since they form an immediate F_∞ generation with the advantages that this can provide (Snape and Simpson 1982). This paper extends the scope of such analyses further by presenting methods for estimating the numbers of genes, or more correctly, effective factors, segregating in a cross, using doubled haploid lines derived from F_1 and F_2 generations. Such information will reveal the complexity of the genetical control of a character and this will, in turn, indicate the probability of selecting genotypes of a desired expression.

Methods of estimation

1 Estimation using genetical parameters of an F_1 derived doubled haploid population

The most common conventional method of calculating the number of effective factors controlling quantitative characters uses estimates of the genetical components of variation obtained from the segregating generations of a cross (Mather and Jinks 1982). The ratio of the square of half the parental or cross range to the additive genetical variance has been the most frequently used estimate. Since a population of doubled haploid lines is an F_∞ population equivalent to that derived by inbreeding, the same parameters can be extracted. Thus if a large population is produced, the two opposite extreme doubled haploid lines should be in complete association for all the increasing and decreasing alleles segregating between the parents. Consequently, the cross range, in Mather and Jinks (1982) terminology is $2[d] = 2kd$, where k is the number of segregating loci and d the additive effect of an individual locus, which is assumed equal for all loci. The genetical variance of the doubled haploid population is D, the additive genetical variance, which in terms of individual effects is $k d^2$.

Hence an estimate of k is obtained from the ratio of the square of half the range to the population genetical variance.

$$\frac{[d]^2}{D} = \frac{k^2 d^2}{k\ d^2} = k.$$

As an alternative to half the range, the deviation of the most extreme line from the population mean can be substituted (Choo and Reinbergs 1982).

The assumptions underlying this method are the same as for any F_∞ population, namely,

(i) that an additive genetical model is adequate, and that linkage and epistasis are absent,

(ii) that the effects of individual loci are equal,

(iii) that the two opposite extreme doubled haploid lines contain all the increasing and decreasing alleles, respectively, segregating between the parents.

Failure of any of these assumptions will result in estimates deviating from the true number.

(i) Epistasis will not affect the difference between the two extreme doubled haploid lines but will inflate the estimate of the additive genetical variance. Thus the estimate of k would be obtained from $[d]^2/(D + I)$, where I is the additive espistatic variance, and consequently, k would be underestimated. If the deviation of the most extreme homozygote from the mean is used, then both the numerator and denominator are affected, k being estimated from the ratio $([d] + [i])^2/(D + I)$ which will overestimate the true number. Skewness of the distribution caused by genotype × environment interaction can also cause serious bias with this method.

Linkage will reduce the number of effective factors estimated and the extent of the bias will depend on the degree of linkage disequilibrium in the population. The bias can be reduced by developing the doubled haploid population from the F_2 rather than the F_1. If both F_1 and F_2 derived populations are available the effects of linkage and epistasis can be detected and the direction of possible biases evaluated (Snape and Simpson 1981).

(ii) The bias introduced by the inequality of gene effects at individual loci has been discussed by Mather and Jinks (1982). Generally, inequality will result in an underestimate of k and the greater the inequality the smaller the estimate of k.

(iii) The most serious limitation to the accuracy of estimation is the bias introduced by the use of a finite sample which may not contain the extreme genotypes. The probability, P_r, that a sample of n lines from a cross segregating for k factors contains a line with r increasing genes is,

$$P_r = 1 - \left(1 - \overset{k}{\underset{r}{C}} \left(\tfrac{1}{2}\right)^k\right)^n$$

where $\overset{k}{\underset{r}{C}} = k!/r!\,(k-r)!$

If the probability that this line is the lowest scoring genotype in the sample is P'_r then

$$P'_r = P_r\left(1 - \sum_{i=0}^{r-1} P'_i\right).$$

These formulae show that the probability of obtaining an extreme genotype with all k decreasing or increasing alleles is small even when few loci are segregating. For example, its probability in a sample of 100 lines segregating for 10 loci is 0.093.

When estimates are based on the difference between the two extremes in the sample then the probability that this exhibits all possible gene differences is the square of the probability of obtaining one extreme (i.e. 0.009 in the above example). More generally, the probability that the two extreme lines differ by r increasing alleles (where $r > \frac{1}{2}k$) is given by the sum of products:

$$\sum_{i=0}^{k-r} P'_i\, P'_{(k-t-i)}. \qquad (1)$$

This indicates that there is a low probability of detecting all gene differences using the cross range even with relatively small numbers of factors segregating. Because of this problem Choo and Reinbergs (1982) proposed the use of the deviation of the most extreme line from the mean of the sample.

The probability of obtaining a sample containing either a line with r increasing or one with r decreasing alleles, P''_r, is

$$P''_r = 1 - \left(1 - 2\overset{k}{\underset{r}{C}} \left(\tfrac{1}{2}\right)^k\right)^n$$

and the probability that one of these lines is the most extreme genotype of a sample is, P^*_r,

$$P^*_r = P_r\left(1 - \sum_{i=0}^{r-1} P^*_i\right). \qquad (2)$$

This modification gives an improvement of P^*_r over P'_r which is approximately equivalent to doubling the sample size. However serious underestimation of the true number of loci segregating remains likely. Furthermore, when this is the case, both methods suffer from another source of downward bias arising from the estimation procedure itself, because while the deviation measured is a function of the r gene differences, the variance estimate depends on all k loci. When the difference between the two extremes is used, the estimated quantity, a, is r^2/k. When the most extreme line is used $a = (2r - k)^2/k$, and the bias is larger.

The total effect of these two sources of bias is illustrated in Fig. 1 for both methods. Here the probability of obtaining a given estimate or greater is

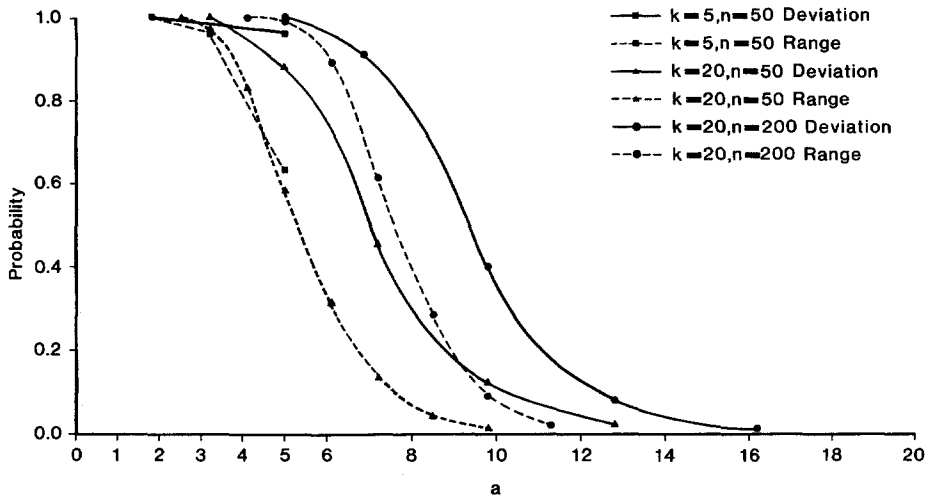

Fig. 1. Probability of obtaining a given estimate, a, or greater for true gene numbers, k, of 5 or 20 in samples of size, n, of 50 or 200 on the basis of the cross range or the deviation of the extreme line. (When k=5, n=200, P=1.0 for both methods)

plotted against expected gene estimates, a. These values were calculated for situations of 5 and 20 genes with sample sizes of 50 or 200, using equations (1) and (2).

These graphs show that while the modified method is always the more efficient, both methods grossly underestimate the true number of factors. Although increases in sample size enhance the probability of detecting more loci the likelihood of discrimination between 10 and 20, or even between 5 and 10 loci remains low.

2 Genotype assay using F_2 derived doubled haploid lines

An alternative method of estimating the number of effective factors has been proposed by Jinks and Towey (1976). Their method, termed genotype assay, involved assessing the frequency of individuals in a generation that are heterozygous, for one or more loci, by detecting segregation in their progeny. The observed frequency of heterozygotes can then be directly equated to theoretical expectations of the numbers of loci segregating.

This method is applicable to populations of doubled haploid lines developed from a sample of F_2 individuals. In the simplest case suppose two doubled haploid lines are developed from each F_2 individual. If an F_2 individual is heterozygous at r loci then the probability that two doubled haploid lines developed from it are genetically different is,

$$1-(\tfrac{1}{2})^r.$$

Assuming that all genotypic differences can be detected as phenotypic differences, the frequency of heterozygotes detectable by examining pairs of doubled haploid lines, P_{max}, is, therefore,

$$\frac{1}{2^k}\sum_{r=0}^{k}\binom{k}{r}[1-(\tfrac{1}{2})^r]$$

which reduces to

$$[1-(3/4)^k]. \tag{3}$$

Solving this equation for different values of k gives the expected proportions of heterozygous individuals.

In practice, not all genotypic differences will be expressed as phenotypic differences because internal balance can cause different genotypes to be phenotypically identical, for example, AAbb ≡ aaBB with equal gene effects (Mather 1973). In these circumstances the observed frequency of heterozygotes is an underestimate and use of the P_{max} equation will produce an underestimate of k.

For the case of equal gene effects giving a maximum degree of internal balance, the frequency of heterozygotes that can be detected is at a minimum. Therefore the proportion, P_{min}, is given by the equation,

$$P_{min}=\frac{1}{2^k}\sum_{r=0}^{k}\binom{k}{r}\left[1-\sum_{n=0}^{r}\left(\binom{r}{n}(\tfrac{1}{2})^r\right)^2\right]. \tag{4}$$

Figure 2 shows the theoretical proportions of heterozygotes that can be detected for the maximum and minimum conditions for k = 1 to 20. The true number of effective factors for any cross or character will fall between these limits. The shape of the graph shows that the method is relatively sensitive for distinguishing small numbers of factors but becomes very insensitive when the proportion of segregating families exceeds about 75%. However, in most circumstances it is more important to know whether a character is controlled by a few or many loci rather than whether 20 or 40 genes are involved. Greater sensitivity could be obtained by developing more than two doubled haploid lines from each F_2 individual.

The advantage of the method of genotype assay is that no assumptions concerning the genetical control of the character are necessary. The only cause of under-

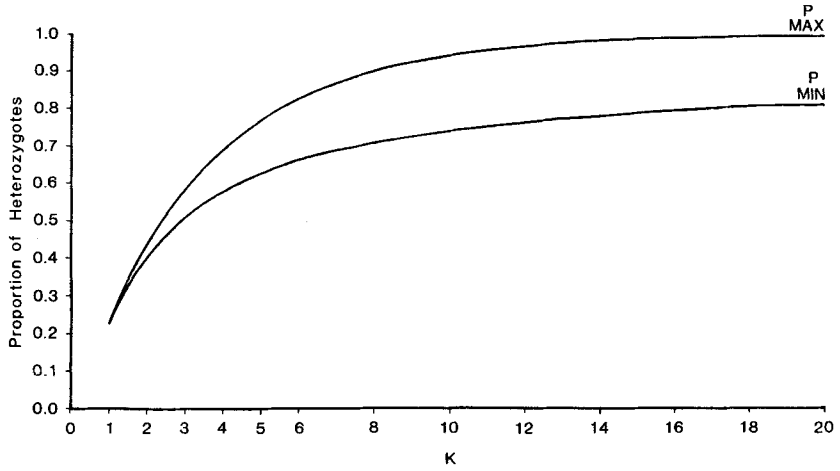

Fig. 2. Expected proportions of F_2 heterozygotes detected by differences between pairs of derived doubled haploid lines for different gene numbers k, when all gene differences are expressed, (P_{max}), and when there is a maximum degree of internal balance (P_{min})

estimation of the number of factors is the sensitivity of the experiment in detecting genetical differences between pair means.

If, however, the assumption of equal gene effects can be made, then further information can be obtained from the distribution of the size of the differences between pair means. The generalised probability equation for the detectable proportion of pairs differing by m genes when k are segregating, assuming internal balance, is,

$$\frac{1}{2^k}\sum_{r=m}^{k} {}^{k}C_{r}\left[2\left(\sum_{n=0}^{(r-m)} {}^{r}C_{n}\,{}^{r}C_{(n+m)}\right)(1/2)^{2r}\right]. \qquad (5)$$

Thus the distribution of pairs differing by 0 to k genes can be calculated. Table 1 shows the proportions of families differing by m genes when k are segregating for values of m = 1–6 and k = 1–10. For example, if k = 7, then 44.8% of pairs will differ by the effect of one gene, 18.7% by two and 4.7% by three gene effects. By comparing the expected distribution of differences to the observed distribution, confirmatory information on the number of factors segregating may be obtained.

Table 1. Expected proportions of families differing by m loci when k are segregating

k	m						
	0	1	2	3	4	5	6
1	0.750	0.250					
2	0.594	0.375	0.031				
3	0.492	0.434	0.070	0.004			
4	0.423	0.457	0.107	0.012	0.001		
5	0.375	0.462	0.139	0.022	0.002	0.0001	
6	0.339	0.458	0.165	0.034	0.004	0.0003	
7	0.311	0.448	0.185	0.047	0.007	0.0007	
8	0.289	0.437	0.201	0.059	0.012	0.0015	0.0001
9	0.272	0.425	0.213	0.071	0.016	0.0026	0.0003
10	0.257	0.414	0.222	0.082	0.021	0.0040	0.0005

3 Estimation using F_2 doubled haploid population variances

The estimate of the number of factors obtained from using the genetical parameters of an F_1 or F_2 derived population is equivalent to the K_1 estimate of Mather and Jinks (1982). Similarly a population of F_2 derived doubled haploid lines can yield an estimate equivalent to the K_2 estimate of Mather and Jinks, which is usually obtained from the variances of an F_3 population. Thus the true variance between the means of families derived from F_2 individuals is $\frac{1}{2}D = \frac{1}{2}k\,d^2$ and the variance of the variances between doubled haploids within F_2 derived families is $\frac{1}{4}D^2 = \frac{1}{4}k\,d^4$. Hence an estimate of k is obtained from:

$$\frac{(\text{true Var. } F_2\text{DH fam. means})^2}{(\text{Var. of within } F_2\text{DH fam. vars})} = \frac{\frac{1}{4}k^2d^4}{\frac{1}{4}k\,d^4} = k.$$

This estimate is not subject to the sampling bias inherent in estimating the genetical range. However it will be influenced by unequal gene effects which can cause underestimation. Also the precision of the estimate may be less than that obtained using the parameters method because of the larger errors in estimating variances and variances of variances rather than means (Mather and Jinks 1982).

Example

Materials and methods

The three methods of analysis were applied to data on days to ear emergence measured on doubled haploid lines of barley developed using the *H. bulbosum* system. These were produced from the cross between the spring varieties 'Vada' and 'Sultan' and consisted of a population of 40 F_1 derived lines and a population of 86 F_2 derived lines made up of pairs from 43 separate F_2 plants. These were grown with the parents in a

randomised field experiment with five replications. Each doubled haploid line was represented in each replication by a single plot of 11 plants and each parent by three such plots. Days to ear emergence were measured on a plot basis from the date of the first plot to head.

Results

a) Estimation using genetical parameters. The mean, range and additive genetical variance for the F_1 and F_2 derived doubled haploid populations are shown in Table 2, together with the parental means and standard errors. Significant differences between the mid-parental value and the F_1 and F_2 means were found, indicating the presence of additive × additive epistasis in the control of this character. Also a significant difference between the F_1 and F_2 means indicates that some of this epistasis is due to interactions between linked loci. Further evidence of linkage is shown by the significantly larger additive genetical variance in the F_2 derived population. This increase in variance as well as the greater range in the F_2 derived population suggests that the linked loci are predominantly in repulsion phase in the parents.

The estimates of the numbers of effective factors segregating for this character, obtained from the statistics in Table 2, are shown in Table 3. Values vary between 4 and 8 depending on the generation and method. However with the small population sizes used here these estimates are likely to be inaccurate because of the sampling bias discussed above, where estimates > 4 factors have only a low probability of detection.

Table 2. Means, range and variances of the 'Vada' × 'Sultan' F_1 and F_2 derived doubled haploid populations

Generation	*n*	Mean (days)	S.E.	Range	Genetical variance
'Vada'	1	3.73	0.280	–	–
'Sultan'	1	6.60	0.280	–	–
F_1DH population	40	2.97	0.178	1.20 – 5.80	1.0295
F_2DH population	86	3.63	0.177	1.40 – 7.00	1.7336

Table 3. Estimates of numbers of effective factors for ear emergence time using genetical components

Generation	Method	k
F_1DH	(½ Range)²/F_1DH variance	5 – 6
	(extreme F_1DH line – F_1DH mean)²/ F_1DH variance	7 – 8
F_2DH	(½ Range)²/F_2DH variance	4 – 5
	(extreme F_2DH line – F_2DH mean)²/ F_2DH variance	6 – 7

Table 4. F_2 pair means (days) and estimates of numbers of factors

Family				Family		
F_2DH	1	5.4	4.4*	F_2DH 23	4.4	5.0
	2	6.4	6.6	24	3.6	3.4
	3	3.2	3.8	25	3.0	2.2
	4	2.8	4.2***	26	4.0	4.6
	5	5.4	5.2	27	3.8	3.6
	6	3.8	3.6	28	6.2	7.0
	7	6.2	4.8***	29	3.4	3.0
	8	3.2	3.6	30	5.8	3.6**
	9	1.4	2.6**	31	2.6	4.8****
	10	3.6	1.8***	32	1.6	2.8**
	11	2.2	3.4**	33	4.4	6.8****
	12	3.0	4.0*	34	2.0	2.2
	13	2.8	3.4	35	2.4	2.4
	14	3.4	3.6	36	5.2	3.2****
	15	3.4	2.4*	37	2.8	2.6
	16	2.0	3.8***	38	2.6	4.2***
	17	4.2	1.4*****	39	3.2	5.2****
	18	5.4	5.2	40	3.6	1.8***
	19	3.0	1.6***	41	2.0	1.6
	20	2.8	4.4***	42	3.4	3.2
	21	3.8	4.0	43	3.0	3.2
	22	3.6	3.6			

SED = 0.686: Significance of differences
* = 0.20 – 0.10; ** = 0.10 – 0.05; *** = 0.05 – 0.01;
**** = 0.01 – 0.001; ***** < 0.001

Estimates of k
Proportion of lines different at 0.05 = 0.33 : k = 1 – 2
Proportion of lines different at 0.10 = 0.40 : k = 2
Proportion of lines different at 0.20 = 0.47 : k = 2 – 3

b) Genotype assay. The mean ear emergence times for the F_2 derived pairs are shown in Table 4 together with the estimates of the numbers of factors, obtained using Figure 2.

Values of 1 – 3 factors are obtained, although there must obviously be more than one since the cross range exceeds the parental range. The maximum estimate, however, is still lower than the minimum estimate obtained from using genetical parameters. With the levels of sensitivity in the present experiment mean differences between pairs greater than 0.9 days are necessary for 0.20 level significance. Thus only effective factors causing differences of this magnitude or greater are included in these genotype assay estimates. This suggests that there may be two or three factors of major effect combined with an unknown number of factors of minor effect segregating in the cross.

The presence of two or three factors of large effect can be tested by looking at the distribution of the differences between the pair means. These can be compared to expected distributions calculated from the proportions in Table 1. Expected and observed distributions are shown in Table 5 for models of 2, 3, and 4

Table 5. Expected and observed distributions of gene differences between F_2DH pairs

		No. pairs with gene differences of:		
		0	1	2
Expected	2 gene model	25	17	1
	3 gene model	21	19	3
	4 gene model	18	20	5
Observed	5% sig. level	29	13	1
	10% sig. level	26	15	2

Table 6. Estimation using F_2DH population variances

True variance of F_2DH family means	= 1.2256
Variance of within F_2DH family variances (adjusted for sampling variation)	= 0.8664
k	= 1.73

genes. The observed distributions were obtained by classifying the differences into groups according to significance at the 0.05 and 0.01 levels. Thus, for example, a difference of 1.1 days is significantly different from zero at the 5% level. Consequently pairs differing by less than 1.1 days were considered to be genetically identical. Similarly pairs with differences between 1.1 and 2.2 days were assumed to differ by 1 factor and by greater than 2.2 days by 2 or more factors. From Table 5, clearly the best fit of observed to expected is on a model of two factors of large effect. However, it is also probable, in view of the significant linkage, that each of these is made up of more than one gene of smaller effect.

c) Estimation using F_2 derived doubled haploid variances. The variance of the F_2 doubled haploid family means and the variance of the between pair variances of the data in Table 4, are given in Table 6, together with the estimate of k. The value of k suggests two effective factors which is in agreement with the value obtained from the genotype assay method.

Discussion

Doubled haploids have both theoretical and practical advantages over conventional generations for the estimation of numbers of effective factors. The theoretical advantage arises from the fewer components of genetical variation that can be involved in the genetical control of a character. Dominance and dominance related epistatic effects are absent, which results in an uncorrelated, and therefore more precise, estimate of the additive variance. For the genotype assay method, the effect of relational balance caused by ambidirectional dominance (Mather 1973) is eliminated in the calculation of P_{min} proportions. Experimentally, the complete homozygosity of double haploid lines enables data to be collected on a plot rather than the individual plant basis necessary with conventional early generations. This reduces environmental errors of estimation and, in a species such as barley, eliminates effects of intraplot competition caused by genetical segregation.

It has been shown that gene numbers calculated using the deviation or range of a sample of lines can be seriously underestimated because of the low probability that more than a small proportion of the possible gene differences are represented in any line or pair of lines in the sample. The genotype assay method does not suffer from this bias with respect to genes with large effect because the proportion of families segregating is independent of the sample size, although the accuracy of the estimate is not. Only the numbers of genes whose segregational effects are too small to be consistently detectable at the level of experimental accuracy attained will be underestimated. For this reason the genotype assay method is to be preferred, and in practical terms, it requires no extra effort to develop doubled haploid lines from the F_2 generation instead of the F_1. The three methods of estimation described can therefore be applied simultaneously to data from an F_2 derived population, and are complementary in the sense that they are based on different rationale and use independent parameters. However, some degree of underestimation is a problem common to all methods, and so remains.

References

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 (1967) Complementary and duplicate gene interactions in biometrical genetics. Heredity 22:97–103

Mather K (1973) Genetical structure of populations. Chapman and Hall, London

Mather K, Jinks JL (1982) Biometrical genetics, 3rd edn. Chapman and Hall, London

Nitzsche W, Wenzel G (1977) Haploids in plant breeding. Paul Parey, Berlin Hamburg

Snape JW (1982) The use of doubled haploids in plant breeding. In: Induced variability in plant breeding. Centre for Agricultural Publishing and Documentation, Wageningen, pp 52–58

Snape JW, Simpson E (1981) The genetical expectations of doubled haploid lines derived from different filial generations. Theor Appl Genet 60:123–128

Snape JW, Simpson E (1982) The use of doubled haploid lines for genetical analysis in barley. In: Barley genetics, vol IV. Proc 4th Int Barley Genet Symp. Chapman and Hall, London, pp 704–709