*Heredity* **55** (1985) 139–143
© The Genetical Society of Great Britain

Received 1 March 1985

# The effect of linkage on additive genetic variance with inbreeding an $F_2$

**M. J. Kearsey**

**Department of Genetics, University of Birmingham,
Birmingham B15 2TT, England.**

**The effects of linkage on the additive genetical variation among inbred lines produced by single seed descent (SSD) from an $F_2$ is investigated numerically with five and ten gene models. Although considerable bias can be generated by linkage, the additive genetical variation present after one round of random mating is shown to approximate closely to that among the SSD lines over a wide range of gene distributions and would thus be a reliable estimator for breeding purposes. Simple experimental methods of estimating this variance are discussed.**

## INTRODUCTION

Plant breeders working with inbreeding crops need to be able to predict the likely performance of recombinant inbred lines derivable from a given $F_1$ hybrid. The information upon which such predictions are based needs to come from early generations of the cross so that attention and effort can be focused on the most promising crosses. The theory and practice behind such predictions is described elsewhere (Jinks and Perkins, 1972; Jinks and Pooni, 1976, 1981).

In the absence of non-allelic interaction and genotype environment interaction for the trait the true means of the inbred lines obtained from any cross ($P_1 \times P_2$) should be normally distributed with mean $\mu$ and variance $\sigma^2$. Reliable estimates of $\mu$ and $\sigma^2$ early in the inbreeding programme would enable the breeder to predict what proportion of lines from that cross would outperform some specified target ($T$) using the one tailed normal deviate ($Z = (T - \mu)/\sigma$).

The first of these parameters $\mu$, given the assumptions above, is simply the mean of the two parents and being a first degree statistic can be estimated with precision. The variance, on the other hand, is more difficult to estimate both because of the lower reliability of second degree statistics and because it is biased by linkage.

Given that the character concerned is controlled by $k$ unlinked genes, $\sigma^2 = D^* = \sum_{i=1}^{k} d_i^2$, a quantity which is estimable from the $F_2$, $d_i$ being half the effect of a homozygous gene substitution at the $i$th locus (Mather and Jinks, 1982). With linkage, however, this is no longer true and $D^*$ is biased by cross product terms, $d_i d_j$. It is thus necessary to be able to estimate $D^*$ from early generations with the appropriate linkage bias.

The bias due to linkage disequilibrium is greatest in the $F_2$ and declines with successive generations of selfing but, unlike the situation with repeated random mating, the approach to linkage equilibrium is prevented by the rapid increase in homozygosity. In the presence of linkage therefore the component of the additive genetical variance within full-sib families ($D$ following Mather and Jinks, 1982) will differ in the $F_2$, $F_3$, $F_4$, etc. due to the change in the linkage bias. The $D$'s from the variance within full-sib families of the $F_2$, $F_3$, $F_4$ ... are termed $D$'s of rank 1, 2, 3, etc., respectively (Mather and Jinks, loc. cit.) reflecting the rounds of recombination that have contributed to them.

Jinks and Pooni (1982) presented not only the general formula for $D^*$ ($= DVF_\infty$ in their notation but here abbreviated for simplicity of presentation) in the presence of linkage but also the formulae for the various $D$'s (of different rank) obtainable from the $F_2$'s, $F_3$'s etc. They were able to show, in principle, that the mean of the rank 1 and rank 2$D$'s should give an acceptable approximation to $D^*$ and illustrated this numerically with a 2 gene model and with Nicotiana data. The purpose of this present paper is threefold. Firstly to extend the illustrations they provide by looking at several linked loci with different patterns of allelic distribution along the chromosome since the consequences of multilocus linkage are not easy to visualise. Secondly to suggest an alternative and possibly simpler method for predicting $D^*$. Thirdly to illustrate the extensive potential for further selection that may still exist among lines derived by selfing in the presence of linkage.

## THEORY

The general formulae for $D^*$ and the $D$'s of various rank were given by Jinks and Pooni (1982) and are repeated below. In these formulae there are $k$ loci ($i=1$ to $k$) with gene effects $d_i$ and recombination fractions between the $i$th and $j$th loci of $p_{ij}$. The rank of the statistic is given by $r$.

$$D_r = \sum_{i=1}^{k} d_i^2 + 2\left\{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \Delta_{ij}(1-2p_{ij})^r d_i d_j\right\} \qquad (1)$$

and

$$D^* = \tfrac{1}{2}D_1 + \tfrac{1}{4}D_2 + \tfrac{1}{8}D_3 + \cdots + (\tfrac{1}{2})^\infty D_\infty$$

$$= \sum_{r=1}^{\infty} (\tfrac{1}{2})^r D_r \qquad (2)$$

$$= \sum_{i=1}^{k} d_i^2 + 2\left\{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \Delta_{ij}\frac{(1-2p_{ij})}{(1+2p_{ij})} d_i d_j\right\} \qquad (3)$$

where $\Delta_{ij} = +1$ or $-1$ for coupling and repulsion linkages respectively.

As $r$ increases $(\frac{1}{2})^r$ in equation (2) and $(1-2p_{ij})^r$ in equation (1) become small and their product therefore becomes trivial if $r > 2$.

Hence the sum

$$\tfrac{1}{8}D_3 + \tfrac{1}{16}D_4 + \cdots + (\tfrac{1}{2})^\infty D_\infty \simeq \tfrac{1}{4}D_2$$

and thus

$$D^* \simeq \tfrac{1}{2}D_1 + \tfrac{1}{2}D_2.$$

Estimates of $D_1$ and $D_2$ require a pedigree breeding programme at least as far as $F_3$ and then can only be obtained on the assumptions that there is no non-additive variation and that the environmental variation can be obtained from another source. For a further discussion see Jinks and Pooni loc. cit.

An alternative approach to obtaining $\frac{1}{2}D_1 + \frac{1}{2}D_2$ is to mate the $F_2$ at random and to estimate the additive genetical variance of this new population.

The $F_2$ yields a rank $1D$, *i.e.*,

$$D_1 = \sum_{i=1}^{k} d_i^2 + 2\left\{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \Delta_{ij}(1-2p_{ij}) d_i d_j\right\}. \qquad (4)$$

On random mating the linkage disequilibrium term $(1-2p_{ij})$ declines by $(1-p_{ij})$, *i.e.*,

$$D' = \sum_{i=1}^{k} d_i^2 + 2\left\{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \Delta_{ij}(1-2p_{ij})(1-p_{ij}) d_i d_j\right\}. \qquad (5)$$

This is identical to $\frac{1}{2}D_1 + \frac{1}{2}D_2$ and thus provides a direct measure of the required predictor from any half sib mating design.

It is clear from these formulae that in the presence of linkage, the estimates of $D$ will differ from $\sum_{i=1}^{k} d_i^2$ and with certain gene distributions and linkage relationships these departures can be very large. However, when predicting the performance of recombinant inbred lines it is $D^*$ and not $\sum d_i^2$ that is required: indeed in so far as one simply wishes to predict the proportion of lines exceeding some specified value the standard deviation ($\sqrt{D^*}$) is the parameter to be estimated.

Numerical solutions to equations (4) (for $D_1$), (5) (for $D' = \frac{1}{2}D_1 + \frac{1}{2}D_2$) and (3) (for $D^*$) have been obtained for a variety of different combinations of linked genes, with various recombination frequencies. For simplicity no inference is assumed, all recombination frequencies between adjacent loci are equal, as are the gene effects ($d_i$) which are set equal to unity.

## RESULTS

Table 1 demonstrates what can happen with five linked loci either in complete association (a) or in one of four dispersed patterns (b)–(e). Using three different linkage values the appropriate values of (i) $D_1$, (ii) $D' = \frac{1}{2}D_1 + \frac{1}{2}D_2$ and (iii) $D^*$ are presented; in the absence of linkage $D_1 = D' = D = \sum d_i^2 = 5$.

Clearly with the genes completely associated (a) all values of $D$ are grossly inflated. Nonetheless the estimate from (ii) is always close to (iii) and hence, despite the considerable bias relative to $\sum d^2$, $(\frac{1}{2}D_1 + \frac{1}{2}D_2)$ is the more appropriate predictor of inbred line performance.

Let us now turn to the dispersed patterns (b)–(e) since complete association is of rather academic interest to the breeder. Despite the fact that in all cases there are three increasing and two decreasing alleles in the better parent, the effect on the $D$'s varies very considerably as Mather and Jinks (1982) suggest. The greatest reduction is always associated with alternating +'s and −'s (b) and becomes less as neighbouring genes have the same sign. On the other hand, with loose linkage (*e.g.*,

**Table 1** Effects of linkage on (i) $D_1$, (ii) $\frac{1}{2}D_1+\frac{1}{2}D_2$ and (iii) $D^*$ with five genes with equal gene effects ($d_i=1$) and various gene combinations. $p$ is the recombination frequency between adjacent loci; $f$ is the dominance ratio N.B. In all cases $D=5$ with no linkage

| | | | | | % error | | |
|---|---|---|---|---|---|---|---|
| $P$ | Gene distribution along chromosome | (i) $D_1$ | (ii) $D'=\frac{1}{2}(D_1+D_2)$ | (iii) $D^*$ | $\frac{\sqrt{D_1}-\sqrt{D^*}}{\sqrt{D^*}}\times 100$ | $\frac{\sqrt{D'}-\sqrt{D^*}}{\sqrt{D^*}}\times 100$ | $H'/f^2$ |
| 0·10 | | | | | | | |
| a | + + + + + | 18·11 | 16·03 | 15·05 | 10 | 3 | 11·57 |
| b | + − + − + | 1·21 | 1·42 | 1·63 | −14 | −7 | |
| c | + + − − + | 1·98 | 2·43 | 2·69 | −15 | −5 | |
| d | − + + + − | 2·49 | 2·98 | 3·20 | −12 | −3 | |
| e | + + + − − | 4·05 | 4·70 | 4·83 | −9 | −1 | |
| 0·25 | | | | | | | |
| a | + + + + + | 11·13 | 9·29 | 8·86 | 12 | 2 | 7·45 |
| b | + − + − + | 2·13 | 2·73 | 2·99 | −16 | −5 | |
| c | + + − − + | 3·63 | 4·13 | 4·21 | −7 | −1 | |
| d | − + + + − | 4·13 | 4·47 | 4·51 | −4 | −1 | |
| e | + + + − − | 5·88 | 5·84 | 5·72 | 1 | 1 | |
| 0·30 | | | | | | | |
| a | + + + + + | 9·47 | 7·96 | 7·68 | 11 | 2 | 6·45 |
| b | + − + − + | 2·56 | 3·21 | 3·41 | −14 | −3 | |
| c | + + − − + | 4·09 | 4·47 | 4·50 | −5 | −1 | |
| d | − + + + − | 4·48 | 4·70 | 4·72 | −2 | 0 | |
| e | + + + − − | 5·97 | 5·77 | 5·67 | 3 | 1 | |

**Table 2** Effects of linkage on (i) $D_1$, (ii) $\frac{1}{2}D_1+\frac{1}{2}D_2$ and (iii) $D^*$ with 10 genes with equal gene effects ($d_i=1$) and various gene combinations. $p$ is the recombination frequency between adjacent loci; $f$ is the dominance ratio. N.B. In all cases $D=10$ with no linkage

| | | | | | % error | | |
|---|---|---|---|---|---|---|---|
| $P$ | Gene distribition along chromosome | (i) $D_1$ | (ii) $D'=\frac{1}{2}(D_1+D_2)$ | (iii) $D^*$ | $\frac{\sqrt{D_1}-\sqrt{D^*}}{\sqrt{D^*}}\times 100$ | $\frac{\sqrt{D'}-\sqrt{D^*}}{\sqrt{D^*}}\times 100$ | $H'/f^2$ |
| 0·1 | | | | | | | |
| a | + + + + + + + + + + | 54·30 | 45·04 | 41·82 | 14 | 4 | 27·67 |
| b | + − + − + − + − + − | 1·55 | 2·11 | 2·57 | −22 | −9 | |
| c | + + − − + + − − + − | 3·47 | 4·35 | 4·86 | −16 | −5 | |
| d | + + + − − − + + − − | 6·22 | 7·42 | 7·83 | −11 | −3 | |
| e | + + + − − − − − + + | 8·68 | 10·53 | 10·72 | −10 | −1 | |
| f | + + + + + − − − − − | 18·13 | 19·09 | 18·38 | −1 | 2 | |
| 0·25 | | | | | | | |
| a | + + + + + + + + + + | 26·00 | 20·89 | 19·86 | 14 | 2 | 13·00 |
| b | + − + − + − + − + − | 3·78 | 5·05 | 5·63 | −18 | −5 | |
| c | + + − − + + − − + − | 6·64 | 7·72 | 7·95 | −9 | −1 | |
| d | + + + − − − + + − − | 9·96 | 10·41 | 10·30 | −2 | 1 | |
| e | + + + − − − − − + + | 13·41 | 12·89 | 12·50 | 3 | 1 | |
| f | + + + + + − − − − − | 18·50 | 16·25 | 15·56 | 9 | 2 | |
| 0·3 | | | | | | | |
| a | + + + + + + + + + + | 21·11 | 17·23 | 16·58 | 12 | 2 | 11·57 |
| b | + − + − + − + − + − | 4·69 | 6·09 | 6·55 | −15 | −4 | |
| c | + + − − + + − − + − | 7·58 | 8·48 | 8·61 | −6 | −1 | |
| d | + + + − − − + + − − | 10·55 | 10·63 | 10·51 | 0 | 0 | |
| e | + + + − − − − − + + | 13·30 | 12·43 | 12·14 | 5 | 1 | |
| f | + + + + + − − − − − | 16·76 | 14·60 | 14·14 | 9 | 2 | |

$P \geqq 0{\cdot}25$) dispersed arrangement (e) actually results in all $D$'s being greater than $\sum d^2$, since on a weighted average the coupling linkages are tighter.

Nevertheless, despite these variable linkage biases, $D'$ always approximates most closely with $D^*$ and is a considerable improvement on $D_1$ as a predictor. Interestingly, however, pattern *b* always leads to the worst prediction, being less accurate even than with complete association (a).

As was stated earlier, it is $\sqrt{D^*}$ (*i.e.*, $\sigma$) that is the important quantity for predictive purposes. Columns 5 and 6 of table 1 show the extent (as a percentage) that $\sqrt{D^*}$ is over or underestimated by $\sqrt{D_1}$ and $\sqrt{D'}$ for every gene arrangement. Clearly $\sqrt{D'}$ is always very close, never out by more than 7 per cent, while $\sqrt{D_1}$ underestimates $\sqrt{D^*}$ by as much as 16 per cent in some cases. In so far as one is simply interested in predicting the proportion of lines likely to exceed some fixed target value $T$, an error even as great as 10 per cent in estimating $\sqrt{D^*}$ would not seriously mislead one. For example, if $(T-\mu)/\sqrt{D^*}=1{\cdot}5$, there would be 7 per cent of lines exceeding $T$. Had $\sqrt{D^*}$ been under or overestimated by 10 per cent, the corresponding prediction would be 5 per cent and 9 per cent respectively; hardly a serious error particularly given the errors of measurement of the predictor.

With 10 loci (table 2) the picture that emerges is very similar, although the bias to $\sum d_i^2$ is proportionately greater. As one would expect when the 10 genes are associated (a) all estimates of $D$ are considerably greater than the value of 10 ($\sum d_i^2$) expected in the absence of linkage. However, with several dispersed patterns (d), (e), (f) the estimates may also be greater than $D$, the more so with looser linkage. Again the alternating sequence (b) causes greatest reduction in $D$, the values of $D$ increasing as adjacent genes become associated ((b) through (f)).

Even with 10 loci $D'$ continues generally to be the best predictor of $D^*$, although one exception is shown in table 2 ($p=0{\cdot}10$, $f$).

## DISCUSSION

The results presented in tables 1 and 2 illustrate vividly how different patterns of linked genes can affect the additive genetic variance. Even with fairly loose linkage (0·3) between adjacent loci, the additive variance produced by a fixed number of genes can be increased or decreased two-fold depending on their distribution along the chromosome. Tighter linkage exerts even stronger effects.

Thus the variance of inbred line means ($D^*$) can be considerably different from that value obtained with linkage equilibrium, $\sum d_i^2$. The latter then is irrelevant for predicting the performance of inbred lines derived from an F$_2$, although were it possible to estimate $\sum d_i^2$ one could better predict the ultimate limit of response to selection. Over a wide range of linkage values and patterns of gene arrangement the best predictor of $D^*$ is in fact $D'$ $(=\frac{1}{2}D_1+\frac{1}{2}D_2)$. In so far as it is $\sqrt{D^*}$ that is needed to predict the performance of the top inbred lines, $\sqrt{D'}$ results in a discrepancy of only a few per cent and is always a very adequate predictor.

In order for $D'$ to be a useful tool for making predictions in the early generations of a selfing series, it is important to provide the breeder with a simple method of obtaining a reliable estimator of $D'$. Jinks and Pooni (loc. cit.) suggest estimating $\frac{1}{2}D_1$ and $\frac{1}{2}D_2$ separately by means of material derived from simultaneously selfing and sibmating an F$_2$. Providing some measure of environmental variance is also available from a non-segregating generation, it is possible to estimate $D_1$ and $D_2$ and hence $D'$.

A different approach would be to mate the F$_2$ at random (see equation (5)) and to estimate $D'$ directly from this derived population by some appropriate design. Thus, if it is possible to cross the material easily, then any half-sib design such as the North Carolina experiments I, II and III (Comstock and Robinson, 1954) or the Triple Test Cross (Kearsey and Jinks, 1968) can provide an estimate of $D'$ since the covariance of half-sibs $(\sigma^2_{HS})=\frac{1}{8}D'$. Alternatively, the population derived by randomly mating the F$_2$ could be selfed and the variance of the true means of such families (analogous to F$_3$'s) estimated from the analysis of variance of the progeny, yielding—

$$\sigma_s^2=\tfrac{1}{2}D'+\tfrac{1}{16}H'$$

where $H'$ represents the dominance variance.

In the presence of linkage

$$H'=\sum_{i=1}^{k}h_i^2+2\left\{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}h_ih_j[(1-2p_{ij})(1-p_{ij})]^2\right\}\qquad(6)$$

where $h_i$ = the deviation of the heterozygote from the mid homozygote at the $i$th locus.

Selfed families are generally easier to produce than half-sib families and, in the absence of dominance, yield an estimate of $D'$ with four-fold greater power. On the other hand, dominance variation will bias our estimate based on $\sigma_s^2$ and we shall now examine the extent of this bias.

### *(a) With no linkage*

$$D'=\sum_{i=1}^{k}d_i^2,$$

$$H'=\sum_{i=1}^{k}h_i^2=f^2\sum_{i=1}^{k}d_i^2$$

where $f = h_i / d_i$ the dominance ratio

$$\therefore \quad 2\sigma_s^2 = D' + \tfrac{1}{8}H'$$

$$= \sum d_i^2\left(1 + \frac{f^2}{8}\right).$$

But $D^* = \sum d_i^2$

$$\therefore \quad \text{bias} = f^2/8.$$

Thus with complete dominance ($f = 1$) the bias would be only 12·5 per cent, while with lower dominance levels it would decline exponentially.

### *(b) With linkage*

Providing there is directional dominance the bias to the dominance variance due to linkage does not depend on whether the genes are in coupling or repulsion (equation (6)) for it is always positive. Thus with 5 and 10 loci of equal effect ($d_i = 1$), $H' = 5f^2$ and $10f^2$, respectively when the genes are unlinked. Linkage considerably inflates these values as is shown in the last columns of tables 1 and 2. With 10 per cent recombination between adjacent loci, for example, $H'$ becomes $11{\cdot}57f^2$ and $27{\cdot}67f^2$ for 5 and 10 loci irrespective of the original gene arrangements.

The data in tables 1 and 2 enable us to explore the accuracy by which we could predict $\sqrt{D^*}$ by using root $2\sigma_s^2$ ($= \sqrt{D' + \frac{1}{8}H'}$) as the predictor and hence examine the price of ignoring dominance. Table 3 illustrates the extent to which $\sqrt{D^*}$ is overestimated by $\sqrt{2\sigma_s^2}$ for various dominance ratios and gene distributions for 10 per cent recombination. For dominance ratios less than 0·6 the bias is small ($\leqq 10$ per cent) and of little consequence. With greater levels of dominance, however, the degree of overestimation is more serious particularly for those gene arrangements for which the correlation between adjacent loci is low (*e.g.*, (b), (c), (d)). Similar calculations with $p = 0{\cdot}25$ result in very low biases indeed, the reduction in additive variance due to linkage being closely balanced by the increase due to dominance.

It would thus appear that for a wide range of linkage and dominance values a breeder could use $2\sigma_s^2$ as an adequate predictor of the likely variance of the inbreds to be derived by SSD.

**Table 3** The effect of estimating $\sqrt{D^*}$ from $\sqrt{2\sigma_s^2}$ ($= \sqrt{D' + \frac{1}{8}H'}$). Data are percentages by which $\sqrt{D^*}$ is overestimated for different dominance ratios. Recombination set at 0·10 between adjacent loci. (See text for details). The true values of $\sqrt{D^*}$ are given for comparison

| | Gene distribution along chromosome | $\sqrt{D^*}$ | Dominance ratio | | | |
|---|---|---|---|---|---|---|
| | | | 1·0 | 0·8 | 0·7 | 0·6 |
| 5 loci | | | | | | |
| a | + + + + + | 3·88 | 8 | 6 | 5 | 5 |
| b | + − + − + | 1·28 | 32 | 20 | 14 | 9 |
| c | + + − − + | 1·64 | 20 | 12 | 8 | 5 |
| d | − + + + − | 1·79 | 17 | 11 | 7 | 5 |
| e | + + + − − | 2·20 | 13 | 8 | 6 | 4 |
| 10 loci | | | | | | |
| a | + + + + + + + + + + | 6·47 | 8 | 6 | 6 | 5 |
| b | + − + − + − + − + − | 1·60 | 48 | 30 | 22 | 14 |
| c | + + − − + + − − + − | 2·20 | 27 | 16 | 12 | 8 |
| d | + + + − − − + + − − | 2·80 | 18 | 11 | 8 | 5 |
| e | + + + − − − − − + + | 3·27 | 14 | 9 | 7 | 5 |
| f | + + + + + − − − − − | 4·29 | 10 | 8 | 6 | 5 |

Finally inspection of tables 1 and 2 makes it clear that $D^*$ is less than $\sum d^2$ over a wide range of gene distributions. This means that a second or third round of crossing and inbreeding may well yield considerable advance over that possible in the first round of SSD.

## REFERENCES

COMSTOCK, R. E. AND ROBINSON, H. F. 1952. Estimation of average dominance of genes. In Gowen, J. W. (ed.) *Heterosis* Iowa State College Press, Ames, Iowa. pp. 494-516.

JINKS, J. L. AND PERKINS, J. M. 1972. Predicting the range of inbred lines. *Heredity*, *28*, 399-403.

JINKS, J. L. AND POONI, H. S. 1976. Predicting the properties of recombinant inbred lines derived by single seed descent. *Heredity*, *36*, 253-266.

JINKS, J. L. AND POONI, H. S. 1981. Properties of pure breeding lines produced by dihaploidy, single seed descent and pedigree breeding. *Heredity*, *46*, 391-395.

JINKS, J. L. AND POONI, H. S. 1982. Predicting the properties of pure breeding lines extractable from a cross in the presence of linkage. *Heredity*, *49*, 265-270.

KEARSEY, M. J. AND JINKS, J. L. 1968. A general method of detecting additive dominance and epistatic variation for a metrical trait. I. Theory. *Heredity*, *23*, 403-409.

MATHER, K. AND JINKS, J. L. 1982. *Biometrical Genetics* (3rd Edn). Chapman and Hall, London.