

Diallel designs, analyses, and reference populations

A. J. Wright

Some issues which evidently remain as areas of contention in diallel analysis are reviewed. In the estimation of genetic variance components for an ancestral reference population in equilibrium, analysis should be applied to F1 data only, and proceed as for a random effects model. No meaningful reference population is constituted by the parental sample itself, but an equilibrium reference population which could be derived from these parents can be defined. To give unbiased estimates of parameters for this population, analysis must include S1 data, and the parents must be homozygous and in linkage equilibrium. Estimation of dominance variance must allow for the fact that the diallel population itself is not in linkage equilibrium by correcting the specific combining ability sum of squares for the mean S1 vs. F1 difference. Some different analyses of variance of the diallel table are discussed in the context of the above reference populations.

INTRODUCTION

The diallel cross has probably attracted more attention and been the subject of more theoretical examination and practical application over the past 30 years than has any other mating design. In spite of this, comparison of the many sources, including some recent publications (Baker, 1978; Singh and Paroda, 1984; Pooni *et al.*, 1984), suggests that some controversy remains, particularly with respect to the choice of design (*i.e.*, including or excluding S1 progenies or reciprocal crosses), the nature of the population whose genetic parameters are being estimated, and the type of analysis to be applied and the assumptions necessary. The purpose of this note is to attempt to clarify some of these issues, drawing freely from earlier literature. The emphasis is on underlying principles rather than rigorous exposition of computational methods which are comprehensively dealt with elsewhere.

LEVELS OF ANALYSIS

Diallel analysis can be conducted on at least three levels, each successive level aiming to provide more detailed information but requiring more limiting assumptions. The first level analysis can be referred to as combining ability analysis and the second as genetic variance component analysis, both of

which have fixed and random effects forms. The third is a more detailed analysis of genetic variances which will be discussed later. Unless otherwise stated, it will be assumed throughout that all maternal or other reciprocal effects are absent.

Combining ability analysis

The classification of the full-sib families into half-sib groups allows the description of each F1 in terms of the general combining abilities (gca) of its parents and a specific combining ability (sca). If the parental set is the whole population of interest, then analysis is confined to this description and to tests of significance of the two types of combining ability. If the parental set can be regarded as a sample from some larger population, then estimates of the variances σ_g^2 and σ_s^2 can be derived from a subdivision of the total family variance:

$$\sigma_F^2 = 2\sigma_g^2 + \sigma_s^2.$$

No genetical assumptions need to be made about the reference population to which these estimates relate, and it might typically comprise homozygous lines with potential as parents of commercial hybrids. The distinction between fixed and random effects leads to different expectations of mean squares (Griffing, 1956*b*; Wearden, 1964).

Variance component analysis

It is only with respect to a reference population which is in Hardy Weinberg equilibrium with respect to individual loci and linkage equilibrium with respect to all pairs of loci (*i.e.*, uncorrelated distributions of genotypes at different loci) that analysis can proceed to the second level and σ_g^2 and σ_v^2 can be equated to genetic variance components. These conditions are met only by a large population produced by repeated generations of random outcrossing without selection.

Additive, dominance and various epistatic variances can be defined for such a population (Cockerham, 1963; 1980), but a two factor design such as the diallel can only estimate two of them, and epistasis has to be assumed or shown to be absent for estimation to proceed. The $Wr-Vr$ analysis provides a test for epistasis (Jinks and Hayman, 1953) and a further test is available if the F2 generation is also grown (Jinks, 1956; Hayman, 1957). With disomic inheritance and under certain assumptions about parental equilibrium which will be described later, additive and dominance variances can be estimated from σ_g^2 and σ_v^2 , using the relations

$$\sigma_A^2 = 4\sigma_g^2/(1+F)$$

$$\sigma_D^2 = 4\sigma_v^2/(1+F)^2$$

where F is the inbreeding coefficient of the parents. In the notation of Mather and Jinks (1971) for a two allele model of gene action:

$$\sigma_A^2 = \frac{1}{2}D_R = 2\sum u(1-u)[d+h(1-2u)]^2$$

and

$$\sigma_D^2 = \frac{1}{4}H_R = 4\sum u^2(1-u)^2h^2,$$

where u is the frequency of the increasing allele, d and h are the homozygote and heterozygote effects, and summation is over loci. The two allele notation is used for convenience, and is not a necessary assumption at this stage (Kempthorne, 1956). The additive variance, σ_A^2 , carries no assumptions or implications as to perfect additivity of gene action, but describes the quantity of variation which is accounted for by a linear relationship of genotype to gene frequency in a Hardy Weinberg population, and is of significance in determining its behaviour under selection (Falconer, 1981).

When homozygosity of parents for no more than two different alleles per locus is assumed, the third level of analysis may be applied and σ_A^2 subdivided into further components. These are of

little direct relevance to the equilibrium population but convey information about gene action and frequencies (Jinks and Hayman, 1953; Hayman, 1954b), so

$$2\sigma_A^2 = D + H_1 - F^* - H_2,$$

where

$$D = 4\sum u(1-u)d^2$$

$$H_1 = 4\sum u(1-u)h^2$$

$$F^* = 8\sum u(1-u)(2u-1)dh$$

$$H_2 = 16\sum u^2(1-u)^2h^2$$

where $H_2 = H_R$ as defined earlier, and F^* is so written to distinguish it from F denoting inbreeding coefficient, D is the variance appropriate to a population of homozygous lines which might be derived from the reference population, and can be conveniently referred to as the homozygote variance.

DEFINITION OF THE REFERENCE POPULATION

The reference population for genetic analysis has to be in Hardy-Weinberg and linkage equilibrium and these properties are fulfilled only by large populations reproducing by panmixia without selection. There are two possibilities which will be referred to as the ancestral reference and the descendant reference. As implied by these terms, the ancestral reference is a population from which the diallel parents can be considered a sample or from which they were derived by inbreeding without selection. The descendant reference is the equilibrium population which could be generated by repeated cycles of random mating initiated by the diallel parents, and so is defined by the genes in the frequencies in which they are carried. Since the parents can never be considered themselves to constitute an equilibrium population, this is the only meaningful reference relating exactly to the genetic properties of the parents per se, and is the one used by Kuehl *et al.* (1968). These two definitions have genetic implications over and above those imposed by the statistical assumptions of random and fixed genotype effects to which they correspond (Griffing, 1956b; Wearden, 1964).

The reason for the distinction between the two situations can be appreciated if a parental group is considered which is a sample from an ancestral reference but also gives rise to a descendant population. The parents represent a link between the two populations whose striking characteristic is its limited size, p . In other contexts, this small sample

might be referred to as a population bottleneck. Because of the genetic drift caused by this bottleneck, the gene frequencies in the descendant population differ from those of the ancestors, and it is inbred, with an inbreeding coefficient of $F = 1/2p$ or $1/p$ according to whether the parents are a sample of non-inbred plants or their inbred derivatives.

Ancestral reference

Because diploid inheritance is assumed the diallel FIs are a sample of crosses from the ancestral reference population, irrespective of the inbreeding status of their parents, whereas the SIs are inbred and so belong to a different population. The estimation of σ_A^2 and σ_D^2 must therefore be based on general and specific combining ability variance estimates derived from analysis of the FI data only. This conclusion was first drawn by Griffing (1956a) and further emphasised by Kempthorne (1956), Griffing (1958), and Kuehl *et al.* (1968). Griffing (1956b, Method 3) and Matzinger and Kempthorne (1956) described methods of analysis of this diallel design, although a similar method was used earlier by Yates (1947). Hayman (1960, 1963) modified his earlier genetic analyses of the full diallel cross (Hayman, 1954b) so as to use SI data only as a means of deriving estimates of D , F and H_1 from those of σ_A^2 and σ_D^2 already obtained from the FIs.

At this point it may be noted that since the reference population itself is in linkage equilibrium, this assumption does not specifically have to be made with regard to the parental sample. Errors in parameter estimates will be introduced by departures from equilibrium, but these are of the same type as any other sampling variation, such as departures from population gene frequencies, and fall into the statistical rather than the genetical domain. This argument has been made in a more general context by Cockerham (1980). Diploid inheritance and absence of epistasis are therefore the only conditions necessary for the estimation of additive and dominance variances in this case, with parental homozygosity and two alleles per locus also required if analysis is continued to the third level.

Descendant reference

It has been noted that the descendant reference population has an inbreeding coefficient of $F = \frac{1}{2}p$ or $1/p$ for non-inbred and inbred parents respectively, but that the FIs in the diallel are non-inbred.

As the SI progenies have a mean inbreeding coefficient of $\frac{1}{2}$ or 1 in these two cases and constitute a $1/p$ th part of the full diallel table, the complete diallel set correctly represents the descendant population gene frequencies and inbreeding coefficient. This full table is in Hardy Weinberg equilibrium with respect to individual loci, but it can be noted at this point that it is not in linkage equilibrium as the SI progenies have a higher probability of homozygosity at all loci than do the FIs.

Pooni *et al.* (1984) claim to show that a complete diallel set is the appropriate design for the analysis of samples from a reference population (ancestral reference). However, they neglect the issues which have been shown here to determine the difference between the two models, and by explicitly assuming that the gene frequencies in the samples are identical to those in the reference, have in fact examined the descendant reference situation as defined.

The estimation of *gca* or breeding values for defined limited reference populations has been discussed by Wright (1973) in the context of synthetic varieties. The *gca* effect of a parent estimated from its diallel array, including the SI term as one of p values, estimates the value of its random bred progeny in the descendant reference population, and justifies the use of array effects as calculated in Griffing's (1956b) Method 1 and Hayman's (1954a) analyses. The use of the term general combining ability to describe the mean of a set of progenies which includes SIs is novel, but is seen to be consistent with the descendant reference. It contradicts Hayman's (1957) view that the SI progeny should be omitted in this case. As pointed out by Pooni *et al.* (1984), the altered weighting which has to be applied to the SI values to maintain orthogonality when only one reciprocal is grown (Morley-Jones, 1965) means that the *gca* estimates are biased in this case.

Although the *gca* estimates are individually meaningful, the arbitrary origin of the parental genotypes means that the variance of these effects, σ_g^2 , has no simple interpretation in terms of σ_A^2 unless the further assumptions are made that the parents themselves are in linkage equilibrium and have a common, known coefficient of inbreeding. While this latter condition is fulfilled if the parents are all homozygous, Kempthorne (1956) and Baker (1979) and others have pointed out that linkage disequilibrium is inevitable in small diallels when several loci are varying. Jinks and Hayman (1953) and Hayman (1957) have shown that the $Wr - Vr$ analysis can reveal disequilibrium provided that

some segregation takes place in the production of the F1, and suggest that loci which do not segregate can be regarded as allelic for the purpose of analysis. However, such linked loci must segregate in later generations and in common with all loci pairs have been defined as independently distributed in the equilibrium descendant population which is used as the reference.

There are further problems associated with the estimation of σ_D^2 for this population. Griffing's (1956*b*) Method 1 analysis of the diallel including S1s provides an sca sum of squares based on deviations of both F1 and S1 progenies from mean array effects, and this estimate or its equivalent *b* item in Hayman's (1954*a*) analysis has been used to estimate σ_D^2 (Hayman, 1954*b*). However, it has already been noted that the diallel population is not in linkage equilibrium even when the parents are. Whenever more than one locus is segregating and their dominance effects are not so opposed in direction and magnitude so as to cancel exactly, inbreeding depression results which causes lower average values on the leading diagonal than off it, with a consequent inflation of the sca deviations and the estimate of σ_D^2 derived from them. In the reasoning which led them to conclude that σ_A^2 can be used to estimate σ_D^2 , Pooni *et al.* (1984) summed the variance contributions of individual loci but neglected the genotypic correlations caused by the mixing of F1 and S1 progenies. Kuehl *et al.* (1968) have shown that an unbiased estimate is obtained only when the mean F1 vs. S1 contrast is removed from the sca sum of squares, thus using Hayman's (1954*a*) *b - b1* items. This can be contrasted with the use of *b3* alone, equivalent to Griffing's (1956*b*) Method 3 specific combining ability item, for the ancestral reference. Clearly this problem is likely to be more serious for traits and species which show strong inbreeding depression, and when small diallels are used.

Kuehl *et al.* (1968) found that, provided the parents are in equilibrium, the parameters of the descendant reference can be estimated more precisely than those of the ancestors because no sampling of alleles is involved.

The following table summarises the requirements and properties of the various types of analysis discussed.

Estimation of σ_A^2 and σ_D^2		
Reference population	Conditions required	Analytical procedures
Ancestral	(i)	Griffing (1956 <i>b</i>)
Descendant	(i), (ii), (iii)	Kuehl <i>et al.</i> (1968)

Estimation of D, H1, F*, H2		
Reference population	Conditions required	Analytical procedures
Ancestral	All	Hayman (1960)
Descendant	All	Jinks & Hayman (1953) Hayman (1954 <i>b</i>)

Conditions:

- (i) Diploid inheritance, no epistasis or reciprocal effects.
- (ii) Parents included in design.
- (iii) Parents homozygous and in linkage equilibrium.
- (iv) Two alleles per locus.

ANALYSES OF VARIANCE

The analysis of variance of diallels excluding the leading diagonal has already been mentioned and is comparatively uncomplicated. Since this design should only be used for an ancestral reference population, a random effects model is appropriate with the usual consequences for the expectation of mean squares (Griffing, 1956*b*; Wearden, 1964).

Several analyses have been proposed for diallel tables which include S1s, and have been discussed by Baker (1979) and Singh and Paroda (1984). Some of these are based on Griffing's (1956*b*) analysis for the full table, but in addition to fitting array or gca effects, subdivide the sca variation. Hayman's (1954*a*) analysis has already been mentioned, and its use as a means of estimating σ_A^2 and σ_D^2 for the descendant reference. Morley-Jones (1965) modified this analysis for the half diallel. Gardner and Eberhart (1966) proposed an analysis for a half diallel among the parents which are themselves random mating populations, and although both Baker (1979) and Singh and Paroda (1984) have treated it as distinct from the Morley-Jones (1965) analysis, it is in fact identical if the same parentage is used.

The *b* partitioning of Hayman and Morley-Jones is equivalent to an analysis of the deviations of the progeny from their mid-parent values, and Walters and Gale (1976) and Walters and Morton (1978) have suggested that Hayman's *a* (array) item be replaced by the sum of squares of S1 values (*s*) so as to give an alternative analysis with a more obvious genetic structure. However, since *s* is not orthogonal to the *b2* item, this analysis is not statistically acceptable. The only apparent alternative to Hayman's and Morley-Jones' analyses is provided by replacing *l2* (= *b2*) in Walters and Gale's analysis by Griffing's Method 3 general combining ability sum of squares computed from F1 data. This would represent an orthogonal extension of Griffing's Method 3 analysis to

include S1 data, and would offer an ancestral reference counterpart to the Hayman analysis which is better fitted to the descendant reference.

REFERENCES

- BAKER, R. J. 1978. Issues in diallel analysis. *Crop Sci.*, 18, 533-536.
- COCKERHAM, C. C. 1963. Estimation of genetic variances. In Hanson, W. D. and Robinson, H. F. (eds) *Statistical Genetics and Plant Breeding*. Pub. 982. N.A.S., N.R.C., Washington, D.C., pp. 53-94.
- COCKERHAM, C. C. 1980. Random and fixed effects in plant genetics. *Theor. Appl. Genet.*, 56, 119-131.
- FALCOLNER, D. S. 1981. *Introduction to Quantitative Genetics*, 2nd Ed. Longman, London and New York.
- GARDNER, C. O. AND EBERHART, S. A. 1966. Analysis and interpretation of the variety cross diallel and related populations. *Biometrics*, 22, 439-452.
- GRIFFING, B. 1956a. A generalised treatment of diallel crosses in quantitative inheritance. *Heredity*, 10, 31-50.
- GRIFFING, B. 1956b. Concept of general and specific combining ability in relation to diallel crossing systems. *Austr. J. Biol. Sci.*, 9, 463-493.
- GRIFFING, B. 1958. Application of sampling variables in the identification of methods which yield unbiased estimates of genotypic variance components. *Austr. J. Biol. Sci.*, 11, 219-245.
- HAYMAN, B. I. 1954a. The analysis of variance of diallel tables. *Biometrics*, 10, 235-244.
- HAYMAN, B. I. 1954b. The theory and analysis of diallel crosses. I. *Genetics*, 39, 789-809.
- HAYMAN, B. I. 1957. Interaction, heterosis and diallel crosses. II. *Genetics*, 42, 336-355.
- HAYMAN, B. I. 1960. The theory and analysis of diallel crosses. III. *Genetics*, 45, 155-172.
- HAYMAN, B. I. 1963. Notes on diallel cross theory. In Hanson, W. D. and Robinson, H. F. (eds.) *Statistical Genetics and Plant Breeding*. Pub. 982. N.A.S., N.R.C., Washington, D.C., pp. 571-578.
- JINKS, J. L. 1956. The F_2 and backcross generations from a set of diallel crosses. *Heredity*, 10, 1-30.
- JINKS, J. L. AND HAYMAN, B. I. 1953. The analysis of diallel crosses. *Maize Genetics Newsletter*, 27, 48-54.
- KEMPTHORNE, O. 1956. The theory of the diallel cross. *Genetics*, 41, 451-459.
- KUEHL, R. O., RAWLINGS, J. O. AND COCKERHAM, C. C. 1968. Reference populations for diallel experiments. *Biometrics*, 24, 881-901.
- MATHER, K. AND JINKS, J. L. 1971. *Biometrical Genetics*. Chapman and Hall, London.
- MATZINGER, D. F. AND KEMPTHORNE, O. 1956. The modified diallel table with partial inbreeding and interactions with the environment. *Genetics*, 41, 822-833.
- MORLEY-JONES, R. 1965. Analysis of variance of the half-diallel table. *Heredity*, 20, 117-121.
- POONI, H. S., JINKS, J. L. AND SINGH, R. H. 1984. Methods of analysis and the estimation of the genetic parameters from a diallel set of crosses. *Heredity*, 52, 243-253.
- SINGH, O. AND PARODA, R. S. 1984. A comparison of different diallel analyses. *Theor. Appl. Genet.*, 67, 541-545.
- WALTERS, D. E. AND GALE, J. S. 1977. A note on the Hayman analysis of variance for a full diallel table. *Heredity*, 38, 401-407.
- WALTERS, D. E. AND MORTON, J. R. 1978. On the analysis of variance of a half diallel table. *Biometrics*, 34, 91-94.
- WEARDEN, S. 1964. Alternative analyses of the diallel cross. *Heredity*, 19, 669-680.
- WRIGHT, A. J. 1973. The selection of parents for synthetic varieties of outbreeding diploid crops. *Theor. Appl. Genet.*, 43, 79-82.
- YATES, F. 1947. Analysis of data from all possible reciprocal crosses between a set of parental lines. *Heredity*, 1, 287-301.