POTTS, B., M. MCGOWEN, D. WILLIAMS, S. SUITOR, T. JONES, P. GORE and R. VAILLANCOURT (2007): Advances in reproductive biology and seed production systems of *Eucalyptus*: The case of *Eucalyptus* globulus. IUFRO Working group 2.08.03 – Eucalypts and diversity: balancing productivity and sustainability. Durban, South Africa, 22–26 October 2007.

POUND, L., M. WALLWORK, B. POTTS and M. SEDGLEY (2002): Early ovule development following self- and cross-pollinations in *Eucalyptus globulus* Labill. ssp. *globule*. Annals of Botany **89**: 613–620.

SEDGLEY, M. and R. SMITH (1989): Pistil receptivity and pollen tube growth in relation to the breeding systems of *Eucalyptus woodwardii* (*Symphyomyrtus*: *Myrtaceae*). Annals of Botany **64**: 21–31.

SETTERFIELD, S. and R. WILLIAMS (1996): Patterns of flowering and seed production in *Eucalyptus miniata* and *E. tetrodonta* in a tropical savanna woodland, Northern Australia. Australian Journal of Botany **44**: 107–122.

SOKAL, R. and F. ROHLF (1995): Biometry: the principles and practice of statistics in biological research (3rd Edition). W. H. Freeman and Company, New York pp 702–703.

TRINDADE, H., L. BOAVID, N. BORRALHO and J. FEIJO (2001): Successful fertilization and seed set from pollination on immature non-dehisced flowers of *Eucalyptus globulus*. Annals of Botany **87**: 469–475.

VAN WYK, G. (1977): Pollen handling, controlled pollination and grafting of *Eucalyptus grandis*. South African Forestry Journal **101**: 47–53.

VENN, T. (2005): Financial and economic performance of long-rotational hardwood plantation investment in Queensland, Australia. Forestry Policy and Economics **7**: 437–454.

WILLIAMS, D., B. POTTS and P. BLACK (1999): Testing single visit pollination procedures for *Eucalyptus globulus* and *E. nitens*. Australian Forestry **62**: 346–352.

# GSCA: New Software and Algorithms to Analyse Diallel Mating Designs Based on Restricted Linear Model

By CHUNFA TONG[1)], GUANGXIN LIU, LIWEI YANG and JISEN SHI[2)]

Key Laboratory of Forest Genetics and Biotechnology of Ministry of Education,
Nanjing Forestry University, Nanjing 210037, China

(Received 3rd December 2001)

**Abstract**

The diallel mating designs have been extensively employed to gain genetic information by crop and tree breeders, but analysis of diallel data faces some challenges because the same parent acts both male and female roles. Theoretically, little attention was paid to the statistical inference and hypothesis testing for a fixed diallel linear model. In this paper we provide a uniform solution to any fixed diallel linear model with matrix expression based on the theory of restricted linear models. We derive formulae for estimating diallel parameters and their standard errors, and obtain uniform statistics for hypothesis testing of parameters, factors and differences between general combining abilities (GCA) or specific combining abilities (SCA). To put the result into practice, we have developed a Windows® software program "GSCA" for analyzing a flexible diallel linear model that could contain the GCA, SCA, reciprocal, block and environment effects as well as interaction effects such as GCA by environment. GSCA can perform analyses not only for Griffing's four types of diallel crosses but also for more complicated diallel crosses whether the data structure is balanced or unbalanced. A real example is given to illustrate the convenience, flexibility and power of our software for diallel analysis.

*Key words:* Diallel mating design, restricted linear models, general combining ability, specific combining ability, least squares.

## Introduction

Diallel mating designs have been widely used in crop and tree breeding programs to obtain genetic information of the parents involved for determining breeding strategies (SPARAGUE and TATUM, 1942; JINKS, 1954; BURLEY et al., 1966; SNYDER and NAMKOONG, 1978; YEH and HEAMAN, 1987; JONSSON et al., 1992; WU and MATHESON, 2005; GARDNER et al., 2007). A diallel cross consists of of all possible crosses between a number of varieties. If there are $p^2$ combinations, consisting of $p$ selfings and $p(p-1)$ crosses, since reciprocal crosses, alternating pollen and ovum parents may be differential by maternal or paternal effects (MAYO, 1987). As $p$ increases, $p^2$ becomes impossibly large. For this reason, many methods have been developed for the examination of partial diallel crosses (MAYO, 1987).

[1)] Corresponding author: CHUNFA TONG. Key Laboratory of Forest Genetics & Biotechnology of Ministry of Education, Nanjing Forestry University, Longpan Road No. 159, Zip code 210037, Nanjing, Jiangsu Province, China. Phone and Fax: 86-25-85427760. E-Mail: tongchf@njfu.edu.cn

[2)] Corresponding author: JISEN SHI. E-Mail: jshi@njfu.edu.cn



DOI:10.1515/sg-2012-0016
edited by Thünen Institute of Forest Genetics

SPRAGUE and TATUM (1942) originally defined the most important concepts of general combining ability (GCA) as the average performance of a line in hybrid combination, and specific combing ability (SCA) referring to specific crosses that exhibit superiority or inferiority to the average performance of the lines involved. MAYO (1987) more clearly defines GCA as the average performance of a strain in a series of crosses and SCA as deviation in a particular cross from performance predicted in the basis of GCA.

GRIFFING (1956a) summarized diallel crosses into 4 categories (i.e. full diallel, full diallel without selfings, half diallel with selfings, and half diallel without selfings) and provided formulae of calculating the fixed effects of GCAs and SCAs as well as the variance components of GCAs and SCAs as random effects for balanced data. Griffing's analysis ignored the possible selfings, since these can introduce bias, but as noted by GILBERT (1958), if the particular parents are of interest in themselves, it may be more important to include the selfings (MAYO, 1987). The uniqueness that each observation has two levels of the same main effect, and the common phenomena of missing plots or missing crosses (especially in forest trees) in a diallel mating design make it difficult to estimate related genetic parameters in a diallel statistical model (WU and MATHESON, 2001; XIANG and LI, 2001).

Two kinds of statistical linear models are commonly employed to analyze diallel crosses with balanced or unbalanced data. One is the fixed-effects linear model in which GCA and SCA are treated as fixed effects to be estimated to rank the parents for selection (HUBER et al., 1992; WU and MATHESON, 2000), and the other is the random-effects linear model where GCA and SCA are considered as random effects for variance component estimation and further for estimating heritabilities and genetic correlations (WU and MATHESON, 2001; XIANG and LI, 2001). NELDER (1977) discussed various views as to the procedures which involve 'fixed', 'mixed' and 'random' models. In this study, we focus on the fixed-effects linear model to estimate GCA and SCA and to provide hypothesis testing for these parameters and various factors. Early solutions for fixed GCA and SCA effects (GRIFFING, 1956a; BECKER, 1975; FALCONER, 1981; HALLAUER and MIRANDA, 1981) were limited to balanced data and based on ordinary least squares (OLS) with the restrictions that the sum of all effect estimates for an factor equals to zero (HUBER et al., 1992). Later on, HUBER et al. (1992) and WU and MATHESON (2000) described the fixed linear models in matrix notations and gave the OLS estimates by reducing parameters using the sum-to-zero restrictions. Although their method can deal with unbalanced data, it requires analyst to reconstruct the linear model almost by hand and some problems remain such as hypothesis testing for the reduced parameters.

A series of analytical tools have been developed for the cumbersome computations of the GCA and SCA in diallel crosses. These tools are primarily divided into two classes: 1) packages written in high-level computer programming languages, and 2) programs based on standard commercial packages such as the Statistical Analysis System (SAS). The Fortran program DIALL was developed by SCHAFFER and USANIS (1969) only to estimate GCA and SCA variance components, whereas another Fortran program written by SNYDER (1975) could calculate the fixed GCA and SCA effects. The drawbacks of these Fortran programs include unfriendly user interface, inability to handle large data size and inflexibility in choosing a fixed- or random-effects model (JOHNSON and KING, 1998; XIANG and LI, 2001). Since SAS is a powerful tool for statistical analysis, a number of SAS programs have been developed for diallel analysis (ZHANG and KANG, 1997; JOHNSON and KING, 1998; WU and MATHESON, 2000; WU and MATHESON, 2001; XIANG and LI, 2001; MURRAY et al., 2003; ZHANG et al., 2005). Some of these SAS programs can estimate the fixed GCA and SCA effects and their standard errors and provide hypothesis testing for genetic parameters and factors. Others can estimate variance components of GCA and SCA. The SAS codes of these programs are relatively complicated because the same parent generally plays both the male and female roles in a diallel mating design so that the SAS procedures cannot be directly applied to analyze diallel data. Users must be familiar with SAS programming so as to modify the SAS codes when they adopt these programs to analyze their data. Now that Microsoft Windows® is overwhelming popular operating system in personal computers, Windows based software is desirable specifically for analyzing diallel crosses.

In this paper we describe how to construct the fixed linear model and its linear restrictions in matrix notations for a diallel mating design. With the matrix expression of the diallel linear model, the estimates and their standard errors of the parameters such as the fixed GCA and SCA effects are given by a single formula based on the theory of linear models with linear restrictions, respectively (WANG and CHOW, 1994; RAO et al., 2008). Uniform statistics are obtained for hypothesis testing of each parameter and various factors such as GCA, SCA and the interaction between GCA and environment. A formula is presented for hypothesis testing of the difference between GCAs or SCAs. Windows® software has been developed for analyzing a flexible diallel linear fixed effect model that could contain the GCA, SCA, reciprocal, block and environment effects as well as interaction effects such as GCA by environment. The software can perform analyses not only for Griffing's 4 diallel mating designs but also for more complicated diallel designs whether the data structure is balanced or unbalanced. The published radiata pine data (WU and MATHESON, 2000) was analyzed to illustrate the convenience, flexibility and power of our software for diallel analysis.

## Statistical Methods

*Restricted Linear Model and Least Square Estimates*

Consider a simple linear model for a diallel mating design, which is usually specified as

$$y_{ijk} = \mu + G_i + G_j + S_{ij} + \varepsilon_{ijk} \quad (1)$$

after GRIFFING (1956a) which follows SPRAGUE and TATUM (1942) where $y_{ijk}$ is the $k$th observation of the $ij$th cross; $\mu$ is the overall mean; $G_i$ and $G_j$ are the GCA



DOI:10.1515/sg-2012-0016
edited by Thünen Institute of Forest Genetics

effects of the $i$th and $j$th parents, respectively; $S_{ij}$ is the SCA effect of the $i$th and $j$th parents that satisfies $S_{ij}=S_{ji}$ if both exist; $\varepsilon_{ijk}$ is the within plot error term assumed to be normally distributed with mean 0 and variance $\sigma^2_e$. This model is usually used for a half diallel cross in which only one set of $p(p-1)/2$ $F_1$'s are included, but their reciprocals are not included (GRIFFING, 1956b).

To make these effects estimable, some linear constrains must be imposed on the model. The restrictions generally take the form:

$$\sum_i G_i = 0, \sum_i S_{ij} = 0 \text{ and } \sum_j S_{ij} = 0. \tag{2}$$

The linear model (1) and the restrictions (2) can be expressed in the following matrix form,

$$\begin{cases} y = X\beta + \varepsilon \\ L\beta = 0 \end{cases}, E(\varepsilon) = 0, \operatorname{cov}(\varepsilon) = \sigma_e^2 I \tag{3}$$

where $y$ is an $n \times 1$ vector of observation, $\beta$ is a $p \times 1$ vector of parameters, $X$ is an $n \times$ p design matrix with value of 1 or 0 for each element, $L$ is a matrix with $p$ columns and $\varepsilon$ is a vector of random error terms.

In order to illustrate the process of building the restricted linear model for analysis of a diallel cross, we take a concrete example. For a half-diallel mating design of 4 parents, if there were observations of $y_{121}$, $y_{122}$, $y_{123}$, $y_{131}$, $y_{132}$, $y_{141}$, $y_{142}$, $y_{143}$, $y_{231}$, $y_{232}$, $y_{341}$, $y_{342}$, $y_{343}$ with a cross and some data missing, we would have the linear model,

$$\begin{pmatrix} y_{121} \\ y_{122} \\ y_{123} \\ y_{131} \\ y_{132} \\ y_{141} \\ y_{142} \\ y_{143} \\ y_{231} \\ y_{232} \\ y_{341} \\ y_{342} \\ y_{343} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ G_1 \\ G_2 \\ G_3 \\ G_4 \\ S_{12} \\ S_{13} \\ S_{14} \\ S_{23} \\ S_{34} \end{pmatrix} + \begin{pmatrix} \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{123} \\ \varepsilon_{131} \\ \varepsilon_{132} \\ \varepsilon_{141} \\ \varepsilon_{142} \\ \varepsilon_{143} \\ \varepsilon_{231} \\ \varepsilon_{232} \\ \varepsilon_{341} \\ \varepsilon_{342} \\ \varepsilon_{343} \end{pmatrix} \tag{4}$$

with linear restrictions,

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ G_1 \\ G_2 \\ G_3 \\ G_4 \\ S_{12} \\ S_{13} \\ S_{14} \\ S_{23} \\ S_{34} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{5}$$

It can be seen that the problem of computing GCAs and SCAs for any diallel mating design can be expressed as model (3) by this process.

Let $S(A)$ denote the vector space spanned by the column vectors of matrix $A$ and $R(A)$ the rank of matrix $A$. In model (3), if the matrices $X$ and $L$ satisfy: i) $S(X') \cap S(L') = \{0\}$, and ii) $R(X)+R(L)=p$, then the restriction $L\beta=0$ is called satisfying a *side condition*. Under such a side condition, the least square unbiased estimates of $\beta$ and $\sigma^2_e$ are given by

$$\begin{cases} \hat{\beta} = (X'X + L'L)^{-1} X'y \\ \hat{\sigma}_e^2 = SSE/(n-r) \end{cases} \tag{6}$$

where $SSE = y'y - \hat{\beta}'X'y$ is the residual sum of squared errors in model (3) and $r$ is the rank of the design matrix $X$ (WANG and CHOW, 1994). In addition, the covariance matrix of is easily given as

$$Var(\hat{\beta}) = \sigma_e^2 (X'X + L'L)^{-1} X'X (X'X + L'L)^{-1}. \tag{7}$$

*Hypothesis Testing*

In analysis of a diallel cross model, hypothesis tests are involved in various factors and differences between GCAs or SCAs. The linear null hypothesis of interest can be expressed as

$$H_0 : H\beta = 0 \tag{8}$$

where $H$ is an $h \times$ p matrix. Under the null hypothesis $H_0$, model (3) is added extra restrictions and becomes

$$\begin{cases} y = X\beta + \varepsilon \\ C\beta = 0 \end{cases}, E(\varepsilon) = 0, \operatorname{cov}(\varepsilon) = \sigma_e^2 I \tag{9}$$

where $C = \binom{L}{H}$. In general, the restriction $C\beta=0$ does not satisfy the side condition. Therefore, the least square estimate of the parameter vector $\beta$ in model (9) is not of the form as in (6), but is given by

$$\hat{\beta}_H = (T^{-1} - T^{-1}C'Q^{-}CT^{-1})X'y \tag{10}$$

where $T = X'X + L'L + H'H$, $Q = CT^{-1}C'$, and $Q^{-}$ is the pseudoinverse of matrix $Q$. The residual sum of squared errors in model (9) is $SSH = y'y - \hat{\beta}'_H X'y$. Under the null hypothesis and normality assumption, it can be proved that SSH and SSE are mutually independent, and hence, we have the statistic

$$F = \frac{(SSH - SSE)/(df_1 - df_2)}{SSE/(n - df_1)} \sim F(df_1 - df_2, n - df_1) \tag{11}$$

where

$$df_1 = \operatorname{rank}\begin{pmatrix} X \\ L \end{pmatrix} - \operatorname{rank}(L) = p - \operatorname{rank}(L) = \operatorname{rank}(X)$$

and

$$df_2 = \operatorname{rank}\begin{pmatrix} X \\ L \end{pmatrix} - \operatorname{rank}(C) = p - \operatorname{rank}\begin{pmatrix} L \\ H \end{pmatrix}.$$

The theory related to results (10) and (11) can be referred to WANG and CHOW (1994) and RAO et al. (2008).

To save computing time, we express the inverse of matrix $T$ in (10) in another form. Let $M = X'X + L'L$, then $T = M + H'H$, and

$$T^{-1} = M^{-1} - M^{-1}H'(I + HM^{-1}H')^{-1}HM^{-1} \tag{12}$$

(RAO et al., 2008).



DOI:10.1515/sg-2012-0016
edited by Thünen Institute of Forest Genetics

There are two kinds of hypothesis tests in analysis of a diallel cross. One is the test for a factor or an effect, in which the null hypothesis is of the form

$$H_0 : \beta_{k+1} = \beta_{k+2} = ... = \beta_{k+s} = 0, \quad (13)$$

and the coefficient matrix in (8) can be written as $H = (0, I_S, 0)$. For examples, in linear model (4),

$$H = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

if the null hypothesis is $H_0$: $G_2 = 0$, and

$$H = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

if the null hypothesis is $H_0$: $G_1 = G_2 = G_3 = G_4 = 0$. In this case, let $M^* = M^{-1}H'$, then we have by (12)

$$T^{-1} = M^{-1} - M^*(I + HM^*)^{-1}M^{*\prime} \quad (14)$$

The other test is for difference between GSAs or SCAs. The null hypothesis is of the form $H_0$: $\beta_i - \beta_j = 0$, and the matrix $H$ in (8) is only a row matrix with the $i$th element 1 and the $j$th element –1. Hence, we can obtain by (12)

$$T^{-1} = M^{-1} - (M_i - M_j)(M_i - M_j)' / (1 + M_{ii} + M_{jj} - 2M_{ij}) \quad (15)$$

where $M_i$ and $M_j$ are the $i$th and $j$th columns of matrix $M^{-1}$, respectively, and $M_{ij}$ is the $ij$th element.

Another problem is the pseudoinverse computing in (10) because the pseudoinverse of a matrix commonly is not unique and the algorithm is relatively complicated. It can be solved by linearly transforming the row vectors of matrix $C$ so that the calculation of the pseudoinverse of matrix $Q$ can be avoided. Suppose that there exists an invertible matrix $\Gamma$ such that $\Gamma C = \binom{C_1}{0}$, where $C_1$ has full row rank, then it can be deduced that

$$\hat{\beta}_H = \left(T^{-1} - T^{-1}C_1'(C_1T^{-1}C_1')^{-1}C_1T^{-1}\right)X'y \quad (16)$$

where only matrix inverses are involved instead of pseudoinverse.

## Software Development

We have developed Windows® software, GSCA, for computing GCA and SCA in extensive diallel mating designs. The algorithm is based on the above theoretical results of restricted linear models and can be deal with missing data. In addition to providing GCA and SCA estimates, GSCA gives hypothesis testing for model, various factors and differences between GCAs or SCAs. The software can be freely downloaded from the webpage: http://fgbio.njfu.edu.cn/tong/GSCA/gsca.htm.

GSCA focuses on treating the following linear model of a diallel mating design with fixed effects:

$$\begin{aligned} y_{ijlkm} = {} & \mu + E_k + B_{l(k)} + G_i + G_j + S_{ij} + R_{ij} + \\ & GE_{ik} + GE_{jk} + SE_{ijk} + RE_{ijk} + e_{ijlkm} \end{aligned} \quad (17)$$

where $y_{ijlkm}$ is the $m$th observation of the $l$th block within $k$th environment for the $ij$th cross; $\mu$ is the overall mean; $E_k$ is the $k$th environment effect; $B_{l(k)}$ is the $l$th block effect in the $k$th environment; $G_i$ and $G_j$ are the GCA effects of the $i$th female and $j$th male respectively; $S_{ij}$ is the SCA effect of the $i$th and $j$th parents; $R_{ij}$ is the reciprocal effect due to the cross between the $i$th female and the $j$th male; $GE_{ik}$ and $GE_{jk}$ are the $k$th environment with the $i$th and the $j$th GCA interactions, respectively; $SE_{ijk}$ is the $k$th environment with the $ij$th SCA interaction; $RE_{ijk}$ is the interaction of the $k$th environment with the reciprocal effect of $R_{ij}$; and $\varepsilon_{ijlkm}$ is the within plot error term. The linear restrictions for model (17) are as follows, $\sum_k E_k = 0$, $\sum_l B_{l(k)} = 0$ for each $k$, $\sum_i G_i = 0$, $\sum_i S_{ij} = 0$ and $\sum_j S_{ij} = 0$ with $S_{ij} = S_{ji}$, $R_{ij} + R_{ji} = 0$, $\sum_i GE_{ik} = 0$ for each $k$, $\sum_i SE_{ijk} = 0$ and $\sum_j SE_{ijk} = 0$ with $SE_{ijk} = SE_{jik}$ for each $k$, and $RE_{ijk} + RE_{jik} = 0$.

The model can be reduced to some simpler forms that could analyze Griffing's four kinds of diallel crosses. For example, for a half-diallel mating design in a single site, the model will be altered to

$$y_{ijlm} = \mu + B_l + G_i + G_j + S_{ij} + e_{ijlm} \quad (18)$$

with linear restrictions, $\sum_l B_l = 0$, $\sum_i G_i = 0$, $\sum_i S_{ij} = 0$ and $\sum_j S_{ij} = 0$.

The raw data to be analyzed by GSCA should be formatted in a text file as shown in *Table 1*. The first line lists factor and trait names of "Pi", "Pj", "Blk" ,"Env" ,"Trt1", "Trt2", etc., and must be in such order, where "Pi" stands for female parent, "Pj" for male parent, "Blk" for block, "Env" for environment, "Trt1" for trait 1, "Trt2" for trait 2, and so on. From the second line on, each line is the data for an individual that corresponds to factors and traits in the first line. Since GSCA can deal with the flexible model, either block or environment factor or both can be missed in the raw data file.

An appropriate linear model could be chosen by GSCA itself or by hand when you use GSCA for analysis of a diallel cross. When the data is successfully opened by GSCA, if you click the menu "Analysis" and then the option "Run", a dialog window (*Fig. 1*) will pop out for parameter choosing. The default linear model given by GSCA contains the main effects of GCA, SCA, reciprocation, environment and block if they exist. The interaction effects such as GCA by environment and SCA by environment could be chosen by clicking your mouse on the parameter selection window (*Fig. 1*).

*Table 1.* – Data format of GSCA. Either block or environment factor or both can be missed.

| Pi | Pj | Blk | Env | Trt1 | Trt2 |
|---|---|---|---|---|---|
| 1 | 2 | 1 | 1 | 213 | 175 |
| 1 | 2 | 1 | 1 | 211 | 158 |
| 1 | 2 | 2 | 1 | 186 | 169 |
| 1 | 2 | 2 | 2 | 220 | 227 |
| 1 | 2 | 2 | 1 | 172 | 192 |
| 2 | 3 | 1 | 2 | 203 | 179 |
| 2 | 3 | 1 | 1 | 151 | 260 |
| 2 | 3 | 1 | 1 | 184 | 187 |
| 2 | 3 | 1 | 2 | 220 | 203 |
| 2 | 3 | 2 | 1 | 178 | 183 |
| ..... | | | | | |



DOI:10.1515/sg-2012-0016
edited by Thünen Institute of Forest Genetics

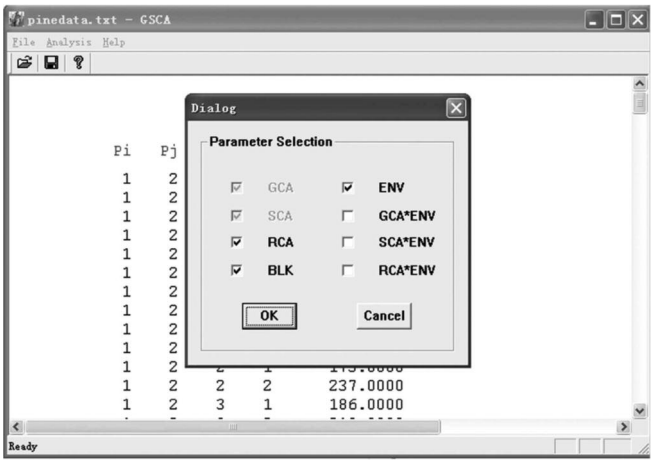


*Figure 1.* – The main window of GSCA with the pop-out window for model selection.

We used the radiata pine data published in WU and MATHESON (2000) as an example to illustrate GSCA's function and its usage convenience. The data is the diameter at breast height (DHB) of radiata pine measured at age 11 in two environments for a $6\times6$ half-diallel mating design with 3 blocks and 4-tree plots at each site. WU and MATHESON (2000) performed the analysis using the following linear model,

$$y_{ijlkm} = \mu + E_k + B_{l(k)} + G_i + G_j + S_{ij} + GE_{ik} + GE_{jk} + SE_{ijk} + e_{ijlkm} \quad (19)$$

which is the reduced form of model (17). We used this model again to implement the calculations of the radiata pine data with GSCA. The analysis here was carried out for scenario 1 of missing crosses (1,3), (2,5) and (5,6) in WU and MATHESON (2000). Thus, the data is unbalanced in both cross and plot levels. The output by GSCA is listed in Appendix A. It is observed that GSCA not only gives the results that the SAS program DIAFIXED.SAS does but also the results of hypothesis testing for differences between GCAs or between SCAs.

## Discussion

We have applied the theory of linear models with linear restrictions to describe the inherited nature of general diallel mating designs. Formulae of estimating fixed genetic parameters and the F-ratio statistics for hypothesis testing of parameters were derived to get the genetic information for determining breeding strategies. The results are adapted to any diallel matings including Griffing's four types of diallel crosses and even more complex mating designs that may have environmental factors and their interactions with genetic effects. The methods can handle balanced and unbalanced data. Unbalanced designs are a common phenomenon, especially in tree breeding programs.

Compared with previous statistical methods for diallel analysis, our statistical methods based on restricted linear models are superior to Griffing's diallel methods (GRIFFING, 1956a) and the OLS analysis (HUBER et al., 1992; WU and MATHESON, 2000). GRIFFING (1956a) provided different formulae to calculate the fixed GCA and SCA effects using genotype means for the four types of diallel crosses, but this method is limited to balanced data structures and lacks the flexibility to be extended to more complicated mating designs. The OLS methods make the parameters estimable by reducing redundant parameters utilizing the sum-to-zero restrictions. However, the procedure for reducing parameters is so trivial that it is almost operated by hand for specific mating design and is difficult to implement with computer programming. Furthermore, the standard errors of the reduced parameters and statistics for hypothesis testing cannot be obtained directly from the reduced linear model. On the contrary, we propose a universal approach to estimating the fixed parameters and giving statistics for hypothesis testing of single or multiple parameters or the difference between them.

Our statistical formulae (eqs. 6~7,10~11) for parameter estimation and hypothesis testing are readily calculated because they are expressed in matrix forms and a lower dimensional matrix inverse is involved. First, since the number of parameters $p$ is generally far less than the sample size $n$, calculating the inverse of $p\times p$ matrix $M=X'X+L'L$ in these formulae does not need much time. Second, although the vector of parameter estimates (eq. 10) under the null hypothesis contains the pseudoinverse of matrix $Q$, which is more complex than a normal inverse in algorithm, linear transformations are applied so that eq. (10) is replaced by eq. (16) where the pseudoinverse is avoided instead of common matrix inverse. Third, the key to calculating the F statistics (eq. 11) is to obtain the estimate of the parameter vector, $\hat{\beta}_H$, which depends on the inverse of matrix $T$. This inverse can be simplified and becomes eqs. (14) and (15) by using eq. (12) and considering the coefficient matrix $C$ of the two types of hypothesis testing, respectively. With these technical treatments, it is feasible to calculate the parameter estimates and the statistics of hypothesis testing in a short while.

GSCA is a typical Windows® based software developed for analyzing model (17) based on the statistical results we have obtained in this paper. It has a user-friendly interface and can give a comprehensive output with one click. Compared with the SAS programs prepared for analysis of the fixed diallel linear model (WU and MATHESON, 2000; ZHANG et al., 2005), GSCA has several major advantages:

i) The SAS packages cannot be directly applied to analyze diallel data because of the uniqueness that the same parent plays both the male and female roles in a diallel cross. Hence, the codes of these SAS programs are usually complicated and most breeders feel difficult to understand them. Users must spend much time to understand and modify these SAS codes so that they can use the modified program to analyze their own data. However, users of GSCA have no annoyance to modify any codes;

ii) GSCA provided $t$ value and its p-value for hypothesis testing of difference between GCAs or SCAs which is equivalent to Griffing's LSD methods, whereas the DIAFIXED.SAS program (WU and MATHESON, 2000) did



DOI:10.1515/sg-2012-0016
edited by Thünen Institute of Forest Genetics

not have this function. The DIALLEL-SAS05 program (ZHANG et al., 2005) can provide LSDs and the corresponding thresholds for significance levels of 0.01 and 0.05 based on Griffing's method (GRIFFING, 1956a), but it may be more convenient to use the p-value than the threshold in determining weather a LSD significantly exists; and

iii) GSCA can handle extensive diallel data structure through the pop out window for choosing a proper linear model, while the SAS programs treat a specific linear model.

**Acknowledgements**

We thank the anonymous referee for his/her constructive modifications, suggestions and comments on this manuscript. This work was supported by the National Natural Science Foundation of China (No. 31070601) and a project funded by the Priority Academic Program Development (PAPD) of the Jiangsu Higher Education Institutions, China.

**References**

BECKER, W. A. (1975): Manual of quantitative genetics. Washington State Univ. Press, Pullman, WA.

BURLEY, J., P. M. BURROWS, F. B. ARMITAGE and R. D. BARNES (1966): Progeny test designs for *Pinus patula* in Rhodesia. Silvae Genet. **15**: 166–173.

FALCONER, D. S. (1981): Introduction to quantitative genetics. Longman and Co., New York, NY.

GARDNER, H. D., W. P. WILLIANS and G. L. WINDHAM (2007): Diallel analysis of aflatoxin accumulation in maize. Field Crops Research **102**: 60–63.

GILBERT, N. E. (1958): Diallel cross in plant breeding. Heredity **12**: 477–492.

GRIFFING, B. (1956a): Concept of general and specific combining ability in relation to diallel crossing systems. Aust. J. Biol. Sci. **9**: 463–493.

GRIFFING, B. (1956b): A generalized treatment of the use of diallel crosses in quantittive inheritance. Heredity **10**: 31–50.

HALLAUER, A. R. and J. B. MIRANDA (1981): Quantitative genetics in maize breeding. Iowa State Univ. Press, Ames, IO.

HUBER, D. A., T. L. WHITE, R. C. LITTELL and G. R. HODGE (1992): Ordinary least squares estimation of general and specific combining abilities from half-diallel mating designs. Silvae Genet. **41**(4–5): 263–273.

JINKS, J. L. (1954): The analysis of continuous variation in a daillel cross of Nicotiana rustica varieties. Genetics **39**: 767–788.

JOHNSON, G. R. and J. N. KING (1998): Analysis of half-diallel mating designs: I – A practical analysis procedure for ANOVA approximation. Silvae Genet. **47**: 74–79.

JONSSON, A., I. DORMLING, G. ERIKSSON and L. NORELL (1992) GCA variance components in 36 *Pinus sylvestris* L. full-sib families cultivated at five nutrient levels in a growth chamber. Forest Science **38**: 575–593.

MAYO, O. (1987): The theory of plant breeding. 2nd Edition. Oxford University Press. 334pp.

MURRAY, L. W., I. M. RAY, H. DONG and A. SEGOVIA-LERMA (2003): Clarification and reevaluation of population-based diallel analyses: Gardner and Eberhart analyses II and III revisited. Crop Sci. **43**: 1930–1937.

NELDER, J. A. (1977): A reformulation of linear models. J. R. Statist. Soc. A. **140**: 48–76.

RAO, C. R., H. TOUTENBURG, SHALABH and C. HEUMANN (2008): Linear models and generalizations: Least squares and alternatives. Springer, New York.

SCHAFFER, H. E. and R. A. USANIS (1969): General least squares analysis of diallel experiments. North Carolina State University, Genetics Department Research Report, Number 1, 61pp.

SNYDER, E. B. (1975): Combining-ability determinations for incomplete mating designs. USDA Forest Service, Southern Forest Experiment Station, General Technical Report SO-9.

SNYDER, E. B. and G. NAMKOONG (1978): Inheritance in a diallel crossing experiment with longleaf pine. *In:* USDA For. Serv. Res. Pap. So-140. South. For. Exp. Stn., New Orleans, LA. 31pp.

SPARAGUE, G. F. and L. A. TATUM (1942): General vs. specific combining ability in single crosses of corn. J. Amer. Soc. Agron. **34**: 923–932.

WANG, S. G. and S. C. CHOW (1994): Advanced linear models: Theory and applications. Marcel Dekker, Inc., New York.

WU, H. X. and A. C. MATHESON (2000): Analysis of half-diallel mating design with missing crosses: Theory and SAS program for testing and estimating GCA and SCA effects. Silvae Genet **49**: 130–137.

WU, H. X. and A. C. MATHESON (2001): Analysis of half-diallel mating design with missing crosses: Theory and SAS program for testing and estimating GCA and SCA variance components. Silvae Genet. **50**: 265–271.

WU, H. X. and A. C. MATHESON (2005): Genotype by environment interactions in an Australian-wide radiata pine diallel mating experiment: Implications for regionalized breeding. Forest Science **51**(1): 29–40.

XIANG, B. and B. LI (2001): A new mixed analytical method for genetic analysis of diallel data. Can. J. For. Res **31**: 2252–2259.

YEH, F. C. and J. C. HEAMAN (1987): Estimating genetic parameters of height growth in seven-year old coastal Douglas-fir from disconnected diallels. Forest Science **33**: 946–957.

ZHANG, Y. D. and M. S. KANG (1997): DIALLEL-SAS: A SAS program for Griffing's diallel analysis. Agron. J. **89**: 176–182.

ZHANG, Y. D., M. S. KANG and K. R. LAMKEY (2005): DIALLEL-SAS05: A comprehensive program for Griffing's and Gardner-Eberhart analysis. Agron. J. **97**: 1097–1106.



DOI:10.1515/sg-2012-0016
edited by Thünen Institute of Forest Genetics

## Appendix

GSCA output from the radiata pine data with missing crosses of (1,3), (2,5) and (5,6).

Analysis of Diallel Mating Design for the Fixed Model
Yijklm = u + Ek + Bl(k) + Gi + Gj + Sij + (GE)ik + (GE)jk + (SE)ijk + eijklm
Wed Jul 08 16:33:16 2009

Result for Trait 1:

Test for Model

| Source | DF | Sum of Squares | Mean Squares | F Value | Pr>F |
|---|---|---|---|---|---|
| Model | 27 | 83200.7934 | 3081.5109 | 3.9106 | 0.0000 |
| Error | 215 | 169418.4247 | 787.9927 | | |
| Total | 242 | 252619.2181 | | | |

Test for Factors

| Source | DF | Sum of Squares | Mean Squares | F Value | Pr>F |
|---|---|---|---|---|---|
| ENV | 1 | 46970.4679 | 46970.4679 | 59.6077 | 0.0000 |
| BLK(ENV) | 4 | 4671.0714 | 1167.7678 | 1.4820 | 0.2087 |
| GCA | 5 | 4100.9094 | 820.1819 | 1.0408 | 0.3945 |
| SCA | 6 | 14450.0999 | 2408.3500 | 3.0563 | 0.0069 |
| GCA*ENV | 5 | 3652.6213 | 730.5243 | 0.9271 | 0.4643 |
| SCA*ENV | 6 | 10345.0622 | 1724.1770 | 2.1881 | 0.0453 |

Parameter Estimates

| Parameter | Estimate | Std Error | T for H0: Parameter=0 | Pr>\|T\| |
|---|---|---|---|---|
| u | 192.7386 | 1.9379 | 99.4562 | 0.0000 |
| E1 | -14.9620 | 1.9379 | -7.7206 | 0.0000 |
| E2 | 14.9620 | 1.9379 | 7.7206 | 0.0000 |
| B(E)11 | 6.3391 | 3.4272 | 1.8496 | 0.0657 |
| B(E)21 | 1.2844 | 3.4272 | 0.3748 | 0.7082 |
| B(E)31 | -7.6235 | 3.4482 | -2.2108 | 0.0281 |
| B(E)12 | -0.7092 | 3.9467 | -0.1797 | 0.8576 |
| B(E)22 | -1.4048 | 3.7609 | -0.3735 | 0.7091 |
| B(E)32 | 2.1140 | 3.9004 | 0.5420 | 0.5884 |
| G1 | 0.7709 | 3.5666 | 0.2161 | 0.8291 |
| G2 | 2.4463 | 3.3747 | 0.7249 | 0.4693 |
| G3 | -3.8971 | 3.5026 | -1.1127 | 0.2671 |
| G4 | 3.6621 | 3.1748 | 1.1535 | 0.2500 |
| G5 | 1.6653 | 3.9262 | 0.4241 | 0.6719 |
| G6 | -4.6474 | 3.3219 | -1.3990 | 0.1632 |
| S12 | 8.4935 | 4.8291 | 1.7588 | 0.0800 |
| S14 | -12.2601 | 4.3759 | -2.8017 | 0.0055 |
| S15 | -5.3767 | 4.5846 | -1.1728 | 0.2422 |
| S16 | 9.1432 | 4.5590 | 2.0055 | 0.0462 |
| ..... | | | | |
| GE32 | -3.2653 | 3.5026 | -0.9323 | 0.3523 |
| GE42 | 4.5548 | 3.1748 | 1.4347 | 0.1528 |
| GE52 | -1.0453 | 3.9262 | -0.2662 | 0.7903 |
| ..... | | | | |
| SE362 | -8.3822 | 4.5088 | -1.8591 | 0.0644 |
| SE452 | -11.5429 | 4.2423 | -2.7209 | 0.0070 |
| SE462 | 11.4298 | 4.6413 | 2.4626 | 0.0146 |

T Values of Test for Difference Between GCAs

| | G6 | G3 | G1 | G5 | G2 |
|---|---|---|---|---|---|
| G3 | 0.1353 | | | | |
| G1 | 0.9693 | 1.0227 | | | |
| G5 | 1.2502 | 0.9013 | 0.1397 | | |
| G2 | 1.3501 | 1.1216 | 0.3036 | 0.1514 | |
| G4 | 1.6949 | 1.5158 | 0.5704 | 0.3414 | 0.2452 |

*Pr<0.05, **Pr<0.01

T Values of Test for Difference Between SCAs

| | S14 | S36 | ... | S46 | S12 | S16 |
|---|---|---|---|---|---|---|
| S36 | 0.5602 | | | | | |
| S23 | 0.8050 | 0.1925 | | | | |
| S26 | 1.3111 | 0.3394 | ... | | | |
| S15 | 0.9770 | 0.5963 | ... | | | |
| S45 | 1.0990 | 0.6763 | ... | | | |
| S24 | 2.5420* | 2.6210** | ... | | | |
| S34 | 2.3454* | 2.0096* | ... | | | |
| S46 | 2.6659** | 2.0519* | ... | | | |
| S12 | 2.7256** | 2.9000** | ... | 0.4130 | | |
| S16 | 2.9330** | 2.3528* | ... | 0.3791 | 0.0857 | |
| S35 | 4.2186** | 2.5276* | ... | 0.7433 | 0.3496 | 0.2522 |

*Pr<0.05, **Pr<0.01

Herausgeber: Johann Heinrich von Thünen-Institut. Bundesforschungsinstitut für Ländliche Räume, Wald und Fischerei.
Schriftleitung: Institut für Forstgenetik, Sieker Landstrasse 2, D-22927 Grosshansdorf
Verlag: J. D. Sauerländer's Verlag, Berliner Strasse 46, D-63619 Bad Orb
Anzeigenverwaltung: J. D. Sauerländer's Verlag, Bad Orb
Gesamtherstellung: PPPP Norbert Wege e.K., Gladenbach — Printed in Germany.


ISSN 0037-5349
DOI:10.1515/sg-2012-0016
edited by Thünen Institute of Forest Genetics