Groundnut Breeding Training Modules
Groundnut Breeding
Training Modules

Compiled by
Bonny R. Ntare

International Crops Research Institute
for the Semi-Arid Tropics

2013

Abstract
This set of 9 training modules were prepared for a training workshop in groundnut breeding targeting national Agricultural Research Scientists and technicians involved in groundnut breeding. The modules included: setting goals and identifying target environments for breeding, breeding theory; experimental designs for controlling field variability, multi-environment trials-design and analysis; broad sense heritability estimates and selection response; correlations among traits-implications for screening; choosing parents and managing a pedigree breeding program; intellectual property rights and Marker Assisted Breeding for cultivar development.

Acknowledgements
This training guide is adapted from the Rice Breeding Modules prepared by the International Rice Research Institute (IRRI).

Author
Bonny R Ntare
Principal Scientist, Groundnut Breeding
International Crops Research Institute for the semi-arid Tropics (ICRISAT)
BP 320 Bamako, Mali

Cover pictures : Bonny R. Ntare
CONTENTS

Acronyms and abbreviations..............................................................................................................vii

MODULE 1

Setting goals and identifying the target environment.................................................................1
Introduction.......................................................................................................................................1
1.1 The target population of environments...............................................................................2
1.2 Factors used in delineating the TPE.......................................................................................3
1.3 Methods for grouping sites and farms into breeding target environments.....................3
1.4 The relationship between the selection environment and the TPE..................................6
1.5 Setting goals and prioritizing traits......................................................................................7

MODULE 2

Breeding theory ...............................................................................................................................11
2.1 Statistics review – linear models, means, variances, LSDs, and repeatability...................11

MODULE 3

Experimental designs for controlling field variability.............................................................19
3.1 Introduction.............................................................................................................................19
3.2 Objectives .............................................................................................................................19
3.3 Experimental design and analysis tools for separating the effects of genotypic and
environmental effects..................................................................................................................19
3.4 Randomized complete-block designs..................................................................................20
3.5 Incomplete-block designs......................................................................................................21

MODULE 4

Multi-environment trials – design and analysis ......................................................................27
4.1 Introduction.............................................................................................................................27
4.2 Objectives.............................................................................................................................28
4.3 The purpose of variety trials ..................................................................................................28
4.4 Linear models describing measurements obtained from field trials..................................29

MODULE 5

Broad-sense heritability estimates and selection response......................................................43
5.1 Rationale.................................................................................................................................43
5.2 Objectives...............................................................................................................................43
5.3 $H$ for MET models................................................................................................................44
<table>
<thead>
<tr>
<th>ACRONYMS AND ABBREVIATIONS</th>
</tr>
</thead>
</table>
| ADN | Acide désoxyribonucléique  
*Deoxyribonucleic acid (DNA)* |
| ATM | Accord de transfert de matériel  
*Material transfer agreement (MTA)* |
| CDB | Convention sur la diversité biologique  
*Convention on biological diversity* |
| CRD | Dispositif complètement randomisé  
*Completely randomized design* |
| DVD | *Digital Versatile Disc* |
| EMS | Carrés moyens attendus  
*Expected mean squares* |
| ES | *Environnement de sélection*  
*Selection environment* |
| FAO | Organisation des Nations Unies pour l’Alimentation et l’Agriculture  
*Food and Agriculture Organization of the United Nations* |
| FTO | Liberté d’opérer  
*Freedom to operate* |
| GE | Génotype x environnement  
*Genotype x environment* |
| GEI | Interaction génotype x environnement  
*Genotype x environment interaction* |
| GSY | Modèle génotype x site x année  
*Genotype x site x year* |
| ICRISAT | Institut international de recherche sur les cultures des zones tropicales semi-arides  
*International Crops Research Institute for the semi-arid Tropics* |
| IRRI | Institut international de recherche sur le riz  
*International Rice Research Institute* |
| LSD | La plus petite différence significative  
*Least Significant Difference* |
| MET | Multi-environment trial  
*Essai multi-environnemental* |
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Abbreviation in English</th>
<th>Definition in English</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEC</td>
<td>Population d’environnements cibles</td>
<td>Target population of environments</td>
</tr>
<tr>
<td>PI</td>
<td>Propriété intellectuelle</td>
<td>Intellectual property (IP)</td>
</tr>
<tr>
<td>QTL</td>
<td>Loci quantitatifs</td>
<td>Quantitative trait loci</td>
</tr>
<tr>
<td>RCBD</td>
<td>Dispositifs à blocs randomisés complets</td>
<td>Randomized complete-block designs</td>
</tr>
<tr>
<td>REML</td>
<td>Méthode de la vraisemblance restreinte maximale</td>
<td>Restricted maximum likelihood</td>
</tr>
<tr>
<td>SAM</td>
<td>Sélection assistée par marqueurs</td>
<td>Marker assisted selection (MAS)</td>
</tr>
<tr>
<td>SED</td>
<td>Erreur-type de la différence entre 2 moyennes</td>
<td>Standard error of a difference between 2 means</td>
</tr>
<tr>
<td>SEM</td>
<td>Erreur-type de la moyenne d’un cultivar</td>
<td>Standard error of a mean</td>
</tr>
<tr>
<td>SML</td>
<td>Système multilatéral</td>
<td>Multilateral System (MLS)</td>
</tr>
<tr>
<td>SNRA</td>
<td>Systèmes nationaux de recherche agricole</td>
<td>National Agricultural Research Systems (NARS)</td>
</tr>
<tr>
<td>TIRGVAAT</td>
<td>Traité international sur les ressources génétiques végétales pour l’alimentation et l’agriculture</td>
<td>International treaty on plant genetic resources for food and agriculture (ITPGRFA)</td>
</tr>
</tbody>
</table>
MODULE 1

Setting goals and identifying the target environment

Introduction

Successful breeding programs have clear objectives. Usually, the objective is to produce a cultivar that is superior to farmers’ varieties in a particular target population of environments (TPE). But what is meant by “superior”?

- Is a superior variety one that is higher-yielding under high-input management used by commercial farmers or under low-input management typical of small-holders?
- Is a superior variety one with a high yield potential in favorable seasons, or one that will produce an assured yield in years of severe drought or disease pressure?
- Is a superior variety one that is high-yielding but low quality, or low-yielding but very high quality?
- Is a superior variety one that performs well in research-station trials, or one that farmers wish to adopt?

There is no absolute definition of a superior variety. In particular situations any of the criteria listed above may constitute superiority. However, if the goal of the breeding program is to have impact in farmers’ fields, breeders need to produce varieties that are better than the farmers’ current variety, under management that the farmer is likely to use, and in the opinion of the farmer. Successful breeders therefore carefully assess the production, quality, and market requirements of the farmers they serve, and the characteristics of the environment in which their varieties will be grown. In this section, we will examine the factors that need to be considered in defining target environments and setting goals for breeding programs.
1.1 The target population of environments

Because the TPE consists of many farms and future seasons, it is best considered to be not a single environment, but a variable set of future production environments (Cooper and Byth, 1996). The idea of the TPE as a population of fields and future seasons is very useful to the breeder. This is because rainfall and pest populations can vary greatly from season to season, and because soil quality, drainage, and management can vary among farms, even within a very small geographical region. This underlying environmental variability results in genotype x environment interaction (GEI), or change in the relative performance of varieties from field to field and season to season. Breeding programs must base selection decisions on data that are predictive of future performance in the TPE, averaged over several farms and seasons. In most situations, breeders will wish to develop cultivars that are superior to currently-used varieties in most years and on most farms within the TPE.

In describing the TPE for a particular breeding program, it is important to recognize that seasonal variation can result in very different conditions in the same field in different years. This is particularly true in rainfed environments. In a given location or region, for example, flooding and submergence may occur in some years, while in other years periods of drought may occur at critical stages of crop development. Characterizing its frequency of occurrence in the TPE can help breeders decide how much time, labor, and money should be invested in screening for a particular stress. For example, in a certain location, weather records may indicate the following rainfall pattern:

<table>
<thead>
<tr>
<th>Favorable years:</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years with drought at seedling stage:</td>
<td>4</td>
</tr>
<tr>
<td>Years with drought at flowering:</td>
<td>6</td>
</tr>
<tr>
<td>Years with disease epidemics</td>
<td>0</td>
</tr>
</tbody>
</table>
At such a location, varieties need to be screened for yield potential in favorable environments and for ability to withstand seedling and reproductive-stage stress. Screening for disease tolerance does not appear warranted.

### 1.2 Factors used in delineating the TPE

There may be several TPE within a single region. Sometimes there is great environmental variability among locations, or even within single farms. This is particularly common in groundnut, because the fields in which groundnut is grown often vary greatly in hydrological and soil characteristics, depending on their position in the toposequence. Sometimes, there are big differences in the way farmers of different socioeconomic groups manage their fields. For example, wealthy farmers may use more fertilizer or pesticides than poorer farmers. If great environmental or socioeconomic variability exists within a region, then it may be best to divide the TPE into separate breeding targets, and either serve each subdivision with its own breeding program, or focus resources on the most important TPE.

Defining the target environment is thus the first and most important decision in planning a breeding program. The target environment is defined in terms of environmental and socioeconomic factors.

### 1.3 Methods for grouping sites and farms into breeding target environments

The breeding program is usually described in terms of:

- a geographical region (for example, West Africa), and
- an ecosystem (e.g., semi-arid).

Within this type of broad environmental classification, the breeder may still be faced with a range of soil types, management regimes, and variety preferences. S/he must therefore consider
whether or not separate breeding targets exist within the TPE. Both quantitative tools and common sense are needed to determine whether to break a TPE into smaller breeding targets. Although it may be possible to describe environmental or socioeconomic differences among farms or fields in a region, it is not necessarily the case that different breeding regions are needed for these different classes of environments. Separate breeding targets are only needed when cultivars perform differently (i.e., when there is rank change) in the different environments. Methods for defining target environments will be described in more detail in a later section, but will be briefly discussed here. A number of tools can be used to help define target environments, including:

**Environmental parameters**

Sites with similar rainfall patterns, and soil types may be grouped for breeding purposes

**Statistical analyses of multi-environment trials**

If multi-environment trials involving large sets of varieties have been conducted over several seasons, several statistical methods can be used to group sites into TPE. These methods will be discussed in more detail in a later section, but include:

- *Cluster or principal components analysis on genotype means or genotype x environment interaction effects, also known as pattern analysis.*

  These methods group trials in a way that maximizes GEI among groups and minimizes it within groups. Pattern analysis is most useful when the TPE is very large and diverse, and when researchers do not have a good hypothesis about the causes of GEI. This is rarely the case.

- *Tests of fixed environmental effects*

  Usually, breeders have a working hypothesis about the most important cause of GEI within the region they serve. For example, the TPE may include both commercial farms where fertilizers are applied and subsistence farms using
low levels of fertilizer or none. It is possible to test the hypothesis that there is cultivar x fertility level interaction in a multi-location cultivar trial by including fertility level as a fixed factor in a combined analysis of variance.

- Examination of correlations of cultivar means across sites
  An easy and effective way of assessing the extent of GEI across environments within a target region is to examine the correlations of genotype means from variety trials conducted across sites or groups of sites. If these correlations are above 0.3 for a single 3-replicate trial, GEI is unlikely to be large. If they are lower than 0.3, there are two possible interpretations; either GEI is important, or the trials have high error terms and there is little genetic variation among cultivars under evaluation. Examining the F-value for cultivars within a single trial can help differentiate these two alternatives; if the F value for cultivars is not significant in a given trial, means from that trial will usually not be significantly correlated with other trials.

**Exercise**

Correlations of cultivar means across pairs of sites from multi-location cultivar trials are presented below, along with F-values for cultivars from each site. Use these correlations to group trials into two potential TPE.

*Correlations among the mean yields of groundnut breeding lines at 6 locations (Variety F values on diagonal)*

<table>
<thead>
<tr>
<th>Location</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7.06</td>
<td>0.45</td>
<td>Ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>B</td>
<td>6.64</td>
<td>0.39</td>
<td>-0.55</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>C</td>
<td>5.81</td>
<td>-0.42</td>
<td>ns</td>
<td>0.63</td>
<td>0.36</td>
<td>0.47</td>
</tr>
<tr>
<td>D</td>
<td>3.68</td>
<td>0.63</td>
<td>20.7</td>
<td>74.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1.4 The relationship between the selection environment and the TPE

After the TPE has been defined, the breeder must design the SE. The SE should be chosen or constructed to maximize the power of trials and nurseries to predict performance in the TPE. If several “populations” of environments occur within the TPE (for example, a particular field may be subject to drought in some years and disease epidemic in others), then several SEs may be needed to maximize overall progress from selection. For example, all lines may have to be screened both for disease and drought tolerance.

Although the SE is constructed to predict performance in the TPE, the breeder must always remember that the SE is not the TPE, and gains made in the SE will not necessarily be expressed in farmers’ fields. It is important for breeders to monitor the correlation between performance on-farm and performance in the SE.

There are 3 important features of a useful SE:

1. The SE must predict performance in at least some seasons and locations within the TPE. In other words the genetic correlation ($r_G$) between TPE and SE must be high.

2. The SE must clearly and repeatably differentiate among genotypes under evaluation. In other words, the heritability (H) for screening in the SE must be high.

3. The SE must permit relatively large numbers of genotypes to be screened at low cost. In other words, the SE must permit a high selection intensity ($i$) to be achieved.

For some stresses, the features of an effective SE are well known. For example, much research effort has been done to identify effective screening methods for tolerance to abiotic and biotic stresses. Methods have been developed that are known to be highly repeatable, predictive of performance
under farmers’ fields, and appropriate for screening large populations. For drought tolerance, the situation is quite different. Many screening methods for reproductive-stage drought tolerance have been proposed, but only recently has evidence begun to accumulate about the effectiveness of these methods in predicting performance under natural drought stress in the TPE.

1.5 Setting goals and prioritizing traits

It is important for any breeding program to have clear goals. These goals need to take into account the current priorities of farmers, but must also consider the way these priorities may change in the future.

**Determining farmer preferences**

Breeders should not assume that they know the priorities of farmers, or that selection on the research station will necessarily result in varieties that farmers prefer. There are many ways in which farmer trait priorities and varietal preferences can be studied, including:

- *Focus group discussions* in which farmers are asked about the positive and negative features of their current varieties.
- *Preference analysis*, in which groups of farmers are asked to rate a set of varieties in a demonstration plot, and to describe the good and bad features of the varieties in the plot.

These methods are discussed in more detail in participatory plant breeding. But breeders can learn a great deal by talking to as many farmers as possible about their production problems and varietal preferences. However, farmers are also usually very concerned with cooking and eating quality, and will reject high-yield varieties that do not meet their quality requirements. Failing to focus on quality is a basic error that has led to the failure of many breeding programs to have an impact.
Occasionally, farmers may not express a preference for characteristics that breeders feel would be useful. This may be the case for major changes in growth or development permitting new cropping options. For example, a reduction in growth duration from 130 to 90 days may permit the sowing of a dry-season crop where none is currently produced, but farmers may not be aware of the existence of such early-maturing material, and may therefore not express a preference for it. Farmers may need to observe such new materials in on-farm demonstration trials before they recognize its potential benefits.

**Setting goals**

Successful breeding programs often have very specific objectives that are useful in planning and measuring progress. These objectives are often best expressed in terms of *developing a replacement for a currently popular variety*. For example, a program may wish to replace a traditional high-yielding variety with one of similar yield but higher yield under farmer management. Several planning decisions follow logically from these objectives.

- A high-yielding locally-preferred variety should be used as a parent in most crosses. Because it can be difficult to recover quality characteristics in a single cross, the high-quality parent may be used as the recurrent parent in generating a $BC_1$-derived population.

- Quality parameters should be the focus of early-generation selection, because they are highly heritable, whereas yield is not.

- The program should be structured to generate a large population of breeding lines with acceptable quality, which can then be evaluated for yield under farmer management.

**Exercise**

In order of priority, list the main objectives of your breeding program. Identify the variety you are trying to replace.
**Summary**

1. Breeders produce varieties for use by farmers in a set of fields and future years called the *target population of environments* (TPE).

2. Considerable season-to-season and field-to-field variability may occur within the TPE.

3. The TPE is defined in terms of both environmental (drought and disease pressures) and socioeconomic factors (like the ability of farmers to afford purchased inputs).

4. There are many ways to group trial sites into TPE. One of the simplest is to group together sites at which line means are highly correlated.

5. A *selection environment* (SE) is used by the breeder to predict performance of new lines in the TPE. If an SE is effective, the rank of breeding lines in it will be similar to their rank in the target environment. It will also permit large populations to be efficiently screened.

6. It is important to determine farmer preferences, especially for quality, and incorporate them as selection criteria. Failure to do so may result in non-adoption of a new line, even if its agronomic performance in the selection environment is excellent.

**Reference**

Breeding theory

2.1 Statistics review – linear models, means, variances, LSDs, and repeatability

2.1.1 Introduction

One of the most important tasks of the breeder is to distinguish between the effects of genotype and environment in nurseries and trials. The genotypic differences between lines in a nursery or trial can be thought of as the “signal” the breeder is trying to detect; the effects of experimental sections (plots) are “noise”. Genotype and environmental effects are both part of the *phenotypic value*, or measurement taken on a field plot. Genotypic and environmental effects are incorporated into the *linear models* used to analyze the results of field trials. A familiarity with the linear model for phenotypic value is a prerequisite to further discussions on how to increase the precision of variety trials, maximize selection response and use breeding resources efficiently. In this section, we will examine the effect of field variability and replication on the precision with which cultivar effects are estimated.

2.1.2 Objectives

- Review the linear model for plot measurements in variety trials and nurseries, and the statistics derived from it.
- Understand the purpose of replication in breeding programs
- Review the concept of experimental repeatability
- Model the relationship between replication, the standard error of a cultivar mean (SEM), the least significant difference between the means of 2 cultivars, and repeatability.
2.1.3 The linear model for plot measurements

Genotype means
Plot measurements are analyzed using a linear model that permits the separation of genotypic effects from random error or “noise” resulting from plot differences in soil quality or water availability. In a completely randomized design (CRD), or a design with no blocking, a measurement on a plot \( P \) in a nursery or variety trial can be decomposed into the following components:

\[
P_{ij} = \mu + G_i + e_j
\]  
[Eq. 1.1]

Where:
\( \mu \) = the mean of all plots
\( G_i \) = the effect of the \( ith \) genotype, expressed as a deviation from the mean of all plots
\( e_j \) = the “residual” effect of the \( jth \) plot, expressed as a deviation from the mean of all plots

Thus, in an unreplicated nursery, measurements on single plots completely confound (mix) genotype and plot effects. In replicated trials or nurseries, the mean of a variety \( j \) over plots \( (\sum Y_i/r) \) has the expected value \( \mu + G_i \). In practice, the mean of a variety estimated from a trial is:

\[
Y_i = \mu + G_i + \sum e_j/r
\]  
[Eq. 1.2]

where \( r \) is the number of replicates.

As all agronomists know, some locations in the field are more productive than others. If we randomly assign a genotype to many different plots (i.e., if replication is high) the positive and negative plot effects tend to cancel each other out, and \( \sum e_j \) approaches 0. However, when replication is low (\( r<4 \)), the residuals \( (\sum e_j/r) \) can cause the mean we estimate from plots to differ greatly from the true genotype mean \( \mu + G_i \). The reason we replicate breeding trials and nurseries is to minimize the
effect of the plot residuals on our estimate of genotype mean, not to generate an error term for testing the significance of differences (and not to compensate for lost plots)!

**Variances**

The variance of a genotype mean is:

\[ \sigma_Y^2 = \frac{\sigma_e^2}{r} \]  

[Eq. 1.3]

The *standard error of a mean (SEM)* is the square root of \( \sigma_Y^2 \).

Breeders are often interested in the *variance of the difference* between 2 varieties. The difference \( D \) may be expressed as follows:

\[ D = Y_1 - Y_2 \]  

[Eq. 1.4]

\[ = Y_1 + (-1)Y_2 \]

For two independent and uncorrelated variables, the variance of the sum equals the sum of the variances. Note that if \( Y = cX \), where \( c \) is a constant, then \( \sigma_Y^2 = c^2 \sigma_X^2 \). Therefore,

\[ \sigma_D^2 = \sigma_{Y_1}^2 + (-1)^2\sigma_{Y_2}^2 \]

\[ = \frac{\sigma_e^2}{r} + \frac{\sigma_e^2}{r} \]

\[ = 2 \frac{\sigma_e^2}{r} \]  

[Eq. 1.5]

The *standard error of a difference between 2 means (SED)* is the square root of \( \sigma_D^2 \).

The *Least Significant Difference* (LSD) is the critical value of a *t-test* of the difference between 2 means. It is calculated as:

\[ \text{LSD} = t_{\alpha/2, \text{edf}} \times \text{SED} \]

\[ = t_{\alpha/2, \text{edf}} \times \sqrt{\frac{2 \sigma_e^2}{r}} \]  

[Eq. 1.6]
Where $t_{\alpha/2, edf}$ is the $t$ value for a significance level of $\alpha/2$, and the number of degrees of freedom in the error term (edf) of the analysis of variance of the trial or nursery. For $\alpha/2$, $t_{\alpha/2, edf}$ is always approximately 2. As an approximation,

$$\text{LSD}_{0.05} = 3 \text{ SEM} \quad \text{[Eq. 1.7]}$$

It is important to be clear about the meaning of the LSD value. The LSD$_{0.05}$ is the difference between estimates of the means of 2 lines with the same genotypic value that is expected to be exceeded by chance in 5% of the repetitions of a variety trial.

Put in another way, suppose you took 2 samples of seed of one variety, harvested from the same plot, and put each sample in a different bag, giving each a different name, and then you took these 2 samples and compared them in a replicated yield trial in different sets of plots. By chance, the difference in mean yield between these samples of identical material would exceed the LSD$_{0.05}$ value 5% of the time.

The SEM, SED, and LSD are important measures of the precision of a variety trial, or its ability to detect small differences.

### 2.1.4 Genetic and phenotypic variances

The genetic (or genotypic) variance in a cultivar trial is the variance of the cultivar effects, or the $G_i$'s in equation 1.5. It is denoted $\sigma^2_G$.

The phenotypic variance in a cultivar trial is the variance of cultivar means across reps. It is denoted $\sigma^2_p$. Because means are based on plot measurements, which contain both $G_i$'s and $e_j$'s, $\sigma^2_p$ contains both the genetic variance and a portion of the residual variance:

$$\sigma^2_p = \sigma^2_G + \sigma^2_e/r \quad \text{[Eq. 1.8]}$$
where $\sigma^2_e$ is the plot residual or error variance from the ANOVA, and $r$ is the number of reps. $\sigma^2_G$ and $\sigma^2_e$ are estimated from the ANOVA as follows:

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Df</th>
<th>Mean square</th>
<th>Expected mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replicates</td>
<td>$r-1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotypes</td>
<td>$g-1$</td>
<td>$MS_G$</td>
<td>$r\sigma^2_G + \sigma^2_e$</td>
</tr>
<tr>
<td>Residual</td>
<td>$(r-1)(g-1)$</td>
<td>$MS_E$</td>
<td>$\sigma^2_e$</td>
</tr>
</tbody>
</table>

Thus, to estimate $\sigma^2_G$:

$$\sigma^2_G = \frac{(MS_G - MS_E)}{r} \quad \text{[Eq. 1.9]}$$

$\sigma^2_e$ is estimated directly as the error variance of the experiment.

### 2.1.5 The repeatability of variety trials

The *repeatability* of a variety trial is the proportion of the variation among line means that is due to the variation in genotype effects. This statistic, also denoted *broad-sense heritability* ($H$), is calculated as:

$$H = \frac{\sigma^2_G}{\sigma^2_p} = \frac{\sigma^2_G}{[\sigma^2_G + \sigma^2_e/r]} \quad \text{[Eq. 1.10]}$$

$H$ is an important measure of the precision (reliability) of an experiment. It is an estimate of the correlation expected between different repetitions or runs of exactly the same experiment, or the correlation you would expect between means estimated for the same set of cultivars, estimated twice by re-randomizing the cultivars to different plots in the same field. Values of $H$ are between 0 and 1. Inspection of Eq. 1.10 shows that $H$ is not a constant! As the number of replicates increases, or as the error variance decreases, $H$ approaches 1. Thus, you can “buy” higher $H$ for your experiment by investing in more replication.
Note that $H$ estimates pertain only to the trial in which they were estimated. H estimates from trials with fewer than 30 varieties are not very reliable.

Note also that H does not tell us anything about the Mendelian transmissibility of a trait from parent to offspring. It is a purely statistical measure of the repeatability of a trait in a particular experiment.

### 2.1.6 Modelling the effect of replication on the precision and repeatability of variety trials

You can use equations 1.6 and 1.10 to model the effect of changing the replicate number in a trial. These equations are therefore useful to breeders in planning their programs. Because replication is expensive; it is important for breeders to examine whether trials are over or under-replicated.

**Modelling Exercise**

The analysis of variance for pod yield for a groundnut variety trial conducted at Samanko, in the rainy season of 2003 is presented below:

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Df</th>
<th>Mean square</th>
<th>Expected mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replicates</td>
<td>3</td>
<td>688426</td>
<td></td>
</tr>
<tr>
<td>Genotypes</td>
<td>21</td>
<td>759729</td>
<td>$r\sigma^2_G + \sigma^2_e$</td>
</tr>
<tr>
<td>Residual</td>
<td>66</td>
<td>159692</td>
<td>$\sigma^2_e$</td>
</tr>
</tbody>
</table>

a) Estimate $\sigma^2_G$ and $\sigma^2_e$

b) Calculate the predicted SEM, SED, LSD.05, and H for trials with 1-8 reps. The appropriate t-value is 1.997.
<table>
<thead>
<tr>
<th>Number of replicates</th>
<th>SEM</th>
<th>SED</th>
<th>LSD&lt;sub&gt;.05&lt;/sub&gt;</th>
<th>Reduction in LSD due to adding last replication</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

c) If two 3-replicate variety trials like the one described above were conducted adjacent to each other in the same field and in the same season, what would be the expected correlation in variety means between the 2 trials?

d) What are the 2 possible ways to increase the precision and repeatability of a variety trial?
Summary

- In field trials and nurseries, genotype and plot effects are always confounded. Plot effects are the “noise” in breeding experiments that obscures estimates of genotypic value.
- In breeding programs, the purpose of replication is to separate plot and genotype effects, reducing their confounding, improving the precision of trials and increasing our ability to identify superior genotypes.
- Genotypic and error variances estimated from the analysis of variance can be used to estimate both LSD and repeatability or broad-sense H.
- Increasing replication will reduce the LSD, increase H, and increase the precision and repeatability of the trial.
- The gains in precision from increasing replication diminish quickly with each additional replicate. It is rarely cost-effective to use more than 4 replicates.
- Equations for LSD and H can be used to model the effect of increasing or decreasing replication on the precision and repeatability of trials.
3.1 Introduction

Field variability is often very great in rainfed groundnut trials, causing plot and genotypic effects to be difficult to separate. Adjacent plots tend to be more alike than plots separated by some distance. Fields may exhibit relatively smooth fertility or drainage gradients, or variability may be patchy, with abrupt changes in yield over a short distance. The main tools for overcoming the confounding effects of field variability on the ability to separate genotypic and environmental effects are replication and blocking. The application of these tools in the most useful complete- and incomplete-block designs is discussed in this section.

3.2 Objectives

- Review the features of modern experimental designs that allow separation of the effects of genotype and environment in field trials and nurseries
- Introduce the concept of repeatability
- Understand the application of randomized complete-block designs (RCBDs), alpha-lattices, and augmented designs

3.3 Experimental design and analysis tools for separating the effects of genotypic and environmental effects

- **Replication** is the most effective means of overcoming the effects of field variability. Replication directly reduces the contribution of plot residuals to estimates of means. As we saw in the modeling exercise, increasing replication
reduces the SEM, and therefore the LSD, in proportion to the square root of the number of replicates (see Eq. 1.6). Thus, doubling the number of replicates reduces the LSD by a factor of $1/\sqrt{2}$, and to halve the LSD, replication must be quadrupled.

- **Blocking**, or the arrangement of experimental sections (plots) into homogeneous subgroups, may help reduce the effects of field variability, particularly when soil quality or drainage changes gradually within the field. Blocking is likely to be less useful when variability is extreme over small intervals and changes in plot value are abrupt.

### 3.4 Randomized complete-block designs

Randomized complete-block designs (RCBDs) group one complete replicate in each block. They are useful when the number of lines in the trial is not large, so that there is less soil and drainage variability within than among blocks. The variability among blocks is thus removed from the plot residuals. For complete-block designs, the plot residual term $e_{ij}$ in the model in Eq. 4.1 is divided into a complete-block effect $R$ (also called a replicate effect, since the replicates are synonymous with blocks in the RCBD) and a within-replicate plot residual $e$. Thus RCBDs remove effects of blocks from the plot residual, reducing the confounding of genotype with plot residual effects. The model for measurements on a plot is:

$$P_{ij} = \mu + G_i + R_j + e_{k(j)} \quad [\text{Eq. 3.1}]$$

Because all varieties occur in all blocks, the replicate or complete-block effects do not contribute to the SEM (and consequently to the SED and LSD).
3.5 Incomplete-block designs

When the number of genotypes per replicate is large, there is likely to be great soil heterogeneity even within the block. This variability may be partly controlled by grouping plots within large replicates or complete blocks into much smaller incomplete blocks. The most commonly-used incomplete-block experimental designs are alpha-lattices for replicated trials and augmented designs or gridding for unreplicated trials. Models for these designs are more complex than for RCBDs and CRDs, and will not be presented here. In brief, the partitioning of complete blocks into smaller, more homogeneous incomplete blocks permits more of the residual variation among plots to be removed from estimates of genotype means.

Gridding

Gridding is a simple form of control of residual variation that can be used in unreplicated nurseries, and does not require repeated checks. The nursery is divided into a set of incomplete blocks. Selection is done within blocks. For example, a nursery or unreplicated trial of 500 lines may be planted in the field diagrammed below, divided into 5 ranges of 100 lines each. The field has a toposequence gradient, with the deepest soil being in the bottom left and top right corners. If we do not impose any sort of blocking or gridding on the field and we attempt to select the best 20% of the lines on the basis of phenotypic acceptability, we will select mainly from the ends of ranges 1 and 5, which have the deepest soil. We will be selecting mainly on the basis of $e$, not $G$. However, if we divide each range into 5 smaller, more homogeneous blocks of 20 plots each (represented by the dotted lines on Fig. 3.1), and select the best 4 lines in each block, we will be selecting the best 20% mainly on the basis of differences in genotypic value rather than spoil depth.
Augmented designs

Augmented designs also use grids or incomplete blocks to remove some field variation from the plot residuals. In an augmented design, a large set of experimental lines is divided into small incomplete blocks. In each incomplete block, a set of checks is included; every check occurs in each incomplete block, but the experimental lines are included in only one block. Because the design is unreplicated, the repeated checks are used to estimate the error mean square and the block effect. The block effect is estimated from the repeated check means and then removed from the means of the test varieties. This reduces error and increases precision somewhat. However, the repeated checks used to estimate block effects add a substantial number of plots to the trial. Block effects could
also be estimated as effectively from the means of the test varieties in each block. This would save considerable space and labor. In general, augmented designs have few advantages over unreplicated nurseries in which block effects are estimated without repeated checks.

**Alpha-lattice designs**

Alpha-lattice designs are replicated designs that divide the replicate into incomplete blocks that contain a fraction of the total number of entries. Genotypes are distributed among the blocks so that all pairs occur in the same incomplete-block in nearly equal frequency. The design permits removal of incomplete-block effects from the plot residuals and maximizes the use of comparisons between genotypes in the same incomplete-block.

How effective are alpha-lattice designs in increasing the precision of genotype means estimated from rainfed rice variety trials? There are several ways to address this question. One way is to compare the SEM or a related statistic like the LSD for trials laid out as alpha-lattices, and analyzed both as alpha-lattices and RCBDs. Surprisingly, there are no published reports of such comparisons in groundnut. However in rice, IRRI compared the results of alpha-lattice and RCBD analyses for a large set of rainfed lowland breeding lines evaluated at 7 locations in northeast Thailand in the 2001 wet season, as well as for a smaller set of released upland cultivars and advanced lines tested under severe drought stress in 3 different agronomic management regimes in the dry season of 2002. These trials were done in large plots (5 or 6 rows). They also compared alpha-lattice analysis with RCBD analysis in a 2-replicate trial of 453 lines in an upland drought-stress trial in which single-row plots were used. The results of these comparisons are presented in Table 3.1.
Table 3.1. Comparison of RCBD and alpha-lattice analyses for 7 NE Thailand RL trials (WS 2001) and 3 IRRI upland trials (DS 2002)

<table>
<thead>
<tr>
<th>Trial set</th>
<th>SEM for RCBD (kg/ha)</th>
<th>SEM for alpha-lattice (kg/ha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NE Thailand RL (WS 2001)</td>
<td>406</td>
<td>398</td>
</tr>
<tr>
<td>IRRI upland (DS 2002)</td>
<td>197</td>
<td>187</td>
</tr>
<tr>
<td>IRRI upland – single rows (DS 2004)</td>
<td>20.3</td>
<td>11.7</td>
</tr>
</tbody>
</table>

There was little benefit to the alpha-lattice analysis in the trials with fewer entries and larger plots. However, in the upland trial with single-row plots, the benefit of the alpha-lattice block effect removal was great, reducing the standard error of a line mean by nearly half. In general, alpha-lattice designs are a risk-free way to slightly increase the precision of field trials. It costs no more to conduct an incomplete-block than a complete-block trial.
Summary

- Replication reduces the influence of plot residuals on cultivar means.
- Blocking reduces the size of plot residuals by removing the effect of the block.
- Increasing replication increases repeatability.
- Alpha-lattices are used to reduce the effect of within-complete-block variation. They can provide risk- and cost-free increases in precision and repeatability, particularly in trials involving very large numbers of entries, and when plot size is small.
- Augmented designs are suitable for unreplicated nurseries. They distribute test varieties among incomplete blocks that each has a set of common checks. They use many extra check plots, however, and are unlikely to be better than designs where block effects are estimated from the lines in the trial rather than from the repeated checks.
- Gridding can be used to reduce the effect of soil quality and drainage gradients on selection in unreplicated nurseries.
MODULE 4

Multi-environment trials – design and analysis

4.1 Introduction

As observed in previous sections, the precision and predictive power of individual field trials is very low. The most important tool for predicting cultivar performance within the TPE is the multi-environment trial (MET). New cultivars are tested at several locations and over several years at locations that sample the TPE. The true values of cultivars are unknown. Cultivar means estimated from METs are the best predictors of future cultivar performance, but they are always estimated with error. This error can be minimized by increasing the number of replicates, sites, and years of field testing, but conducting METs is very expensive, requiring much of the time and money available to plant breeding programs. It is important that testing programs be designed to use these resources efficiently, and to maximize the precision of estimation of cultivar means. In this section, the design and analysis of METs will be presented using software that is available to NARES researchers. Tools for assessing the precision of METs will be presented. You will learn how to decide how to allocate testing effort (sites, years, and replications) to maximize the precision and predictive power a variety testing program, given the personnel, funding, and time available to you.
4.2 Objectives

1. To clarify the purpose of variety trials
2. To introduce linear models for multi-environment trials
3. To describe the analysis of variance for METs
4. To model the variance of a cultivar mean estimated from a MET
5. To examine the effect of replication within and across sites and years on measures of precision.
6. To use standard error and LSD modeling to compare different allocations of testing resources for predictive power, cost, and efficiency.

4.3 The purpose of variety trials

It is very important to understand that the real purpose of a variety trial in a breeding program is prediction. The field and season in which the trial is conducted is considered to be a random sample of farmers’ fields and future seasons in the TPE. The recognition that the individual trial environment is a random factor has important implications for our understanding of the precision of a trial, or its power to detect differences in genotypic value among lines in the trial can be observed. The precision of a variety trial is analogous to the magnifying power of a microscope; a high level of precision in a variety trial is needed to detect a small difference in the genotypic value of breeding lines. A lower level of precision is needed to detect a large difference between varieties. The precision of a variety trial is mainly determined by its level of replication within and across environments. The relative precision of different variety trials can be compared by their SEM or LSD.

The purpose of a variety trial is to predict the performance of new varieties, relative to a check, in farmers’ fields and in future seasons within the TPE.
4.4 Linear models describing measurements obtained from field trials

The genotype x environment (GE) model

We conduct the combined analysis of variety trials over locations and years within the TPE to estimate the mean performance of varieties. Estimates of the precision of the means (SEM and LSD) are also required. To estimate the variance of a cultivar mean, we must use a statistical model that describes the factors or sources contributing to that variance. The simplest GEI model for the analysis of a MET is:

\[
Y_{ijkl} = M + E_i + R(E)_{j(i)} + G_k + GE_{ik} + e_{ijkl}
\]  

where:

- \(Y_{ijkl}\) = the measurement on plot \(l\) in environment \(i\), block \(j\), containing genotype \(k\);
- \(M\) = the overall mean of all plots in all environments;
- \(E_i\) = the effect of environment (trial) \(i\);
- \(R(E)_{j(i)}\) = the effect of replicate \(j\) within environment \(i\);
- \(G_k\) = the effect of genotype \(k\);
- \(GE_{ik}\) = the interaction of genotype \(i\) with environment \(k\);
- \(e_{ijkl}\) = the plot residual.

In this model, genotype effects are usually considered fixed; that is, we wish to estimate the performance of the specific genotypes in the trial. Environments and replicates are random factors; we are not interested in the means of the individual trials per se, but rather are interested in them only as sampling the
TPE. (Occasionally, genotypes may also be considered random, if the purpose of the trial is to estimate genetic variances rather than to predict cultivar performance.) GE interactions are random in this model, because the interaction between fixed and random factors is always random. The random GE term contributes to the true error in the test of differences among cultivars, and to the variance of cultivar means.

According to this model, the variance of a cultivar mean is:

\[
\sigma^2_Y = \frac{\sigma^2_{GE}}{e} + \frac{\sigma^2_e}{er} \tag{4.2}
\]

where \(\sigma^2_Y\) is the variance of a cultivar mean, \(e\) is the number of trials, and \(r\) is the number of replicates per trial.

It is important to minimize \(\sigma^2_Y\) in cultivar testing programs. Minimizing \(\sigma^2_Y\) leads to improved prediction of cultivar performance in the future, and increased gains from breeding. Eq. 4.2 can be used to determine the minimum value for \(\sigma^2_Y\) that can be achieved with the resources available to the breeder. The variance components can be estimated from the ANOVA table for a completely balanced set of trials (all varieties tested in all trials). Methods for estimating variance components from unbalanced data sets are available but require more sophisticated software. The expected mean squares from the ANOVA of a MET are linear functions of the variances of the factors in the model 4.1. These are given below in Table 4.1.
Table 4.1. Expected mean squares (EMS) for the ANOVA of the genotype x environment model assuming all factors random

<table>
<thead>
<tr>
<th>Source</th>
<th>Mean square</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environments (E)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replicates within E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotypes (G)</td>
<td>MS&lt;sub&gt;G&lt;/sub&gt;</td>
<td>σ&lt;sup&gt;2&lt;/sup&gt;&lt;sub&gt;e&lt;/sub&gt; + rσ&lt;sup&gt;2&lt;/sup&gt;&lt;sub&gt;GE&lt;/sub&gt; + rgσ&lt;sup&gt;2&lt;/sup&gt;&lt;sub&gt;G&lt;/sub&gt;</td>
</tr>
<tr>
<td>G x E</td>
<td>MS&lt;sub&gt;GE&lt;/sub&gt;</td>
<td>σ&lt;sup&gt;2&lt;/sup&gt;&lt;sub&gt;e&lt;/sub&gt; + rσ&lt;sup&gt;2&lt;/sup&gt;&lt;sub&gt;GE&lt;/sub&gt;</td>
</tr>
<tr>
<td>Plot residuals</td>
<td>MS&lt;sub&gt;e&lt;/sub&gt;</td>
<td>σ&lt;sup&gt;2&lt;/sup&gt;&lt;sub&gt;e&lt;/sub&gt;</td>
</tr>
</tbody>
</table>

The variance components can be estimated as functions of the mean squares estimated from the ANOVA:

\[ \sigma^2_e = \text{MS}_E \]
\[ \sigma^2_{GE} = \frac{(\text{MS}_{GE} - \text{MS}_e)}{r} \]

It should be noted that these variance components have very large standard errors, and should be used only as a rough guide to planning a breeding or testing program. They should only be estimated if the number of degrees of freedom for the mean square is about 50 or greater.

**Example 1: Resource allocation for a breeding program using the GE model**

Consider a groundnut testing program in which the following variances have been estimated in (t/ha)<sup>2</sup>

\[ \sigma^2_{GE} = 0.30 \]
\[ \sigma^2_e = 0.45 \]
Table 4.2 shows the predicted effect of the number of trials and replicates of testing on the standard deviation of a cultivar mean.

**Table 4.2.** The effect of trial and replicate number on the standard deviation of a cultivar mean: genotype x environment model

<table>
<thead>
<tr>
<th>Number of sites</th>
<th>Number of replicates/site</th>
<th>S.E. of cultivar mean (t ha⁻¹)</th>
<th>LSD₀.₀₅ for the difference between cultivar means (t ha⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.87</td>
<td>2.61</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.72</td>
<td>2.16</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.67</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.64</td>
<td>1.92</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.61</td>
<td>1.83</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.51</td>
<td>1.53</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.47</td>
<td>1.41</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.45</td>
<td>1.35</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.50</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.42</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.39</td>
<td>1.17</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.37</td>
<td>1.11</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.43</td>
<td>1.29</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.36</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.34</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.32</td>
<td>0.96</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.39</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.32</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.30</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.29</td>
<td>0.87</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.27</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.23</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.21</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.20</td>
<td>0.60</td>
</tr>
</tbody>
</table>
Several conclusions for planning further testing can be drawn from this table:

1. The number of trials is more important than the number of replicates per trial in determining the precision with which cultivar means are estimated.

2. For this testing system (and for many others), there is rarely much benefit from including more than 3 replicates per trial.

3. When the number of trials exceeds 3, there is unlikely to be much benefit to including more than 2 replicates per trial.

**Limitations of the genotype x environment model in resource allocation planning**

The model set out in Equation (4.1) has a serious limitation in its usefulness for the planning of testing programs. In this model, each trial is considered to be a separate environment. It can be used to make decisions about how many replicates are warranted, but it sheds no light on the number of years or locations at which trials should be conducted. However, trials in a MET series are easily and naturally categorized with respect to the year and location in which they were conducted. If the environment term in Eq. (4.1) is partitioned into year and location effects, a more realistic and informative model can be used for resource allocation exercises, permitting informed decisions to be made about the number of locations, years, and replicates required to achieve an adequate level of precision in cultivar evaluation.

**The genotype x site x year (GSY) model**

A more realistic and complete model for the analysis of cultivar trials than Eq. (4.1) recognizes years and sites as random factors used in sampling the TPE. The resulting model is:

\[ Y_{ijklm} = M + Y_i + S_j + YS_{ij} + R(YS)_{kl(i)} + G_i + GY_{li} + GS_{ij} + GSY_{ij} + e_{ijklm} \]  

[4.3]
This model differs from Eq. (4.1) in that the $E$ term has been partitioned into site (S) and year (Y) effects and their interaction (YS). Similarly, the GE term has been partitioned into GY, GS, and GY components.

The variance of a cultivar mean is now expressed as:

$$
\sigma^2_{\bar{Y}} = \frac{\sigma^2_{GY}}{y} + \frac{\sigma^2_{GS}}{s} + \frac{\sigma^2_{GYS}}{ys} + \frac{\sigma^2_e}{syr}
$$ [4.4]

Inspection of Equation (4.4) indicates that, if $\sigma^2_{GS}$ is large, the variance of a cultivar mean can only be minimized if testing is conducted at several sites. Similarly, if $\sigma^2_{GY}$ is large, testing over several years will be required to achieve adequate precision in the estimation of cultivar means. However, if $\sigma^2_{GYS}$ is the largest component of GEI, then it may be possible to minimize $\sigma^2_{\bar{Y}}$ by either increasing the number of sites or the number of years of testing. Increasing the number of sites is expensive, but will produce the desired information quickly. Increasing the number of years of testing is less expensive but will delay cultivar release.

Variance components for the GLY model can also be estimated from the fully balanced ANOVA of a set of trials repeated over locations and years:
Table 4.3. Expected mean squares (EMS) for the ANOVA of the genotype x location x year (GLY) model assuming all factors random

<table>
<thead>
<tr>
<th>Source</th>
<th>Mean square</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years (Y)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sites (S)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y x S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replicates within Y x S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotypes (G)</td>
<td>$\text{MS}_G$</td>
<td>$\sigma_e^2 + r\sigma_{GYS}^2 + rs\sigma_{GY}^2 + ry\sigma_{GS}^2 + rys\sigma_G^2$</td>
</tr>
<tr>
<td>G x Y</td>
<td>$\text{MS}_{GS}$</td>
<td>$\sigma_e^2 + r\sigma_{GYS}^2 + ry\sigma_{GS}^2$</td>
</tr>
<tr>
<td>G x S</td>
<td>$\text{MS}_{GY}$</td>
<td>$\sigma_e^2 + r\sigma_{GYS}^2 + rs\sigma_{GS}^2$</td>
</tr>
<tr>
<td>G x Y x S</td>
<td>$\text{MS}_{GYS}$</td>
<td>$\sigma_e^2 + r\sigma_{GYS}^2$</td>
</tr>
<tr>
<td>Plot residuals</td>
<td>$\text{MS}_e$</td>
<td>$\sigma_e^2$</td>
</tr>
</tbody>
</table>

As for the GE model the variance components for the GLY model can be estimated as functions of the mean squares estimated from the ANOVA:

$$\sigma_e^2 = \text{MS}_E$$

$$\sigma_{GYS}^2 = (\text{MS}_{GYS} - \text{MS}_e)/r$$

$$\sigma_{GY}^2 = (\text{MS}_{GY} - \text{MS}_{GYS})/rs$$

$$\sigma_{GS}^2 = (\text{MS}_{GS} - \text{MS}_{GYS})/ry$$
Example 2: Resource allocation for a rice breeding program using the GSY model

Cooper et al. (1999) conducted a resource allocation study for the Thai rainfed lowland rice breeding program in northern Thailand. In this study over 1000 unselected breeding lines from 7 crosses were evaluated for 3 years at 8 sites. Variance components estimates were:

\[
\begin{align*}
\sigma^2_{GS} &= 0.003 \pm 0.006 \\
\sigma^2_{GY} &= 0.049 \pm 0.006 \\
\sigma^2_{GYS} &= 0.259 \pm 0.009 \\
\sigma^2_e &= 0.440 \pm 0.006
\end{align*}
\]

These estimates are of interest in themselves. Their relative magnitudes yield important information about the nature of GE interaction in northern Thailand. Inspection of these estimates leads to the following conclusions:

1. \( \sigma^2_{GS} \) is very small relative to the other components, indicating that cultivars do not, on average, perform differently at different locations. There is little evidence, therefore, that specific lines are adapted to specific locations within the TPE.

2. \( \sigma^2_{GYS} \) is the largest component of GEI (a result found in many studies), indicating that cultivar ranks vary randomly from site to site and from year to year.

3. \( \sigma^2_e \) is very large relative to other components of variance, indicating that within-trial field heterogeneity is great. High levels of replication will be needed to achieve acceptable precision. Improvements in field technique and the use of experimental designs that control within-block error should also be considered.

The predicted effect of site, year, and replicate number on the standard error of a line mean evaluated in northern Thailand is presented in Table 4.4.
**Table 4.4.** The effect of location, year and replicate number on the standard deviation of a cultivar mean: GLY model (variance component estimates from Cooper *et al.*, 1999)

<table>
<thead>
<tr>
<th>Number of sites</th>
<th>Number of years</th>
<th>Number of replicates/site</th>
<th>S.E. of cultivar mean (t ha⁻¹)</th>
<th>LSD₀.₀₅ of cultivar means (t ha⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Some conclusions:

1. Testing in approximately 10 trials (5 sites x 2 years or 10 sites x 1 year) is needed to bring LSD values for the difference between 2 cultivars below 1 ton per ha (the LSD is approximately 3 times the standard error of a cultivar mean).

2. Two replicate per trial are adequate if at least 5 trials are used to estimate means.

Estimates of the variance of cultivar means from single trials are biased downwards

It is important for breeders to recognize that the variance of cultivar means estimated from a single trial is severely biased downwards. This is because in an analysis of a single trial, there is no way to separate $\sigma^2_G$ from $\sigma^2_{GY}$, $\sigma^2_{GS}$, and $\sigma^2_{GYS}$. This is because G, GS, GY, and GYS effects are completely confounded (or inseparably mixed) in single trials. In a single trial, only $\sigma^2_e$ can be estimated separately from $\sigma^2_G$ and used in the calculation of the variance of a cultivar mean. Thus, if the variance of a cultivar mean is estimated from a in a single trial,

$$\sigma^2_Y = \frac{\sigma^2_e}{r} \quad [4.5]$$

whereas in a MET,

$$\sigma^2_{\bar{y}} = \frac{\sigma^2_{GY}}{y} + \frac{\sigma^2_{GS}}{s} + \frac{\sigma^2_{GYS}}{ys} + \frac{\sigma^2_e}{syr} \quad [4.6]$$

In a single trial, the values of y and s are 1, and the true variance of a cultivar mean is therefore:

$$\sigma^2_Y = \sigma^2_{GY} + \sigma^2_{GS} + \sigma^2_{GYS} + \frac{\sigma^2_e}{r}$$
rather than $\sigma^2 / r$. If the purpose of the trial is to predict future performance of the varieties under test, it is clear that $\sigma^2_Y$ severely underestimates $\sigma^2_Y$. The literature indicates that the true variance of a cultivar mean for purposes of prediction is at least twice as large as the variance estimated from a single trial. This is why breeders evaluate advanced lines at several sites.

**Deciding whether to divide a target population of environments into two regions for breeding purposes**

The analyses above are useful tools for deciding how many locations, years and replicates of testing are needed to predict performance of a variety with adequate precision. They do not, however, give guidance about whether the TPE should be broken into 2 breeding targets. There are many genotype x environment interaction analyses that can be used to group sites into relatively similar groups based on cultivar performance. These methods, including cluster, pattern, and AMMI analyses, are useful when there is no pre-existing hypothesis to test about the pattern of adaptation of varieties to sites. However, when there is a good hypothesis to test (based on breeder knowledge, farmer practice, or agroclimatic data), there is a straightforward model that can be used for testing whether locations within the TPE can really be grouped into more homogeneous subsets. In this model, trial sites are grouped into subgroups based on location or some other fixed factor. For example, if a breeder wished to evaluate whether two subregions, say the northern and southern parts of the TPE, should be considered separate breeding targets, he or she would classify all trials in the north into one group, and all fields in the south into a second group. A combined analysis would be performed over trials, with the location factor broken down into two subcomponents: subregion (north or south) and trials within subregion. This is illustrated for the 2-way GxE model.
In this model, subregions are considered fixed effects. Locations within subregions are a random sampling factor, like replicates or years. The hypothesis that varieties perform very differently in different regions can be tested by testing the variety x subregion mean square against the variety x locations within subregion mean square. (This model can be easily extended to the genotype x location x year model). Because this is a mixed model the analysis needs to be done with software that can support mixed model analyses, including the latest release of IRRISTAT.

The ANOVA table with expected mean squares for a balanced case is given below:

**Table 4.5. Expected mean squares (EMS) for the ANOVA of the genotype x subregion for testing fixed groupings of trial sites**

<table>
<thead>
<tr>
<th>Source</th>
<th>Mean square</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subregions (S)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Locations within subregions (L(S))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replicates within L(S)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotypes (G)</td>
<td>MS_G</td>
<td>$\sigma^2_e + r\sigma^2_{GL(S)} + r\sigma^2_{GS} + r\sigma^2_{GS}$</td>
</tr>
<tr>
<td>G x S</td>
<td>MS_{GS}</td>
<td>$\sigma^2_e + r\sigma^2_{GL(S)} + r\sigma^2_{GS}$</td>
</tr>
<tr>
<td>G x L(S)</td>
<td>MS_{GL(S)}</td>
<td>$\sigma^2_e + r\sigma^2_{GL(S)}$</td>
</tr>
<tr>
<td>Plot residuals</td>
<td>MS_e</td>
<td>$\sigma^2_e$</td>
</tr>
</tbody>
</table>
Example of a test of significance of genotype x subregion interaction

Breeders in Laos would like to know if the southern and central parts of the country need to be considered separate TPE. A trial involving 22 tradition varieties was conducted at 6 locations in the wet season of 2004, three in the center and three in the south.

The ANOVA table is presented below:

**Table 4.6.** ANOVA for 22 varieties tested in two Lao PDR subregions (north and south), with 3 locations per subregion, in WS 2004

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subregions (S)</td>
<td>1</td>
<td>5459785</td>
<td></td>
</tr>
<tr>
<td>Locations within subregions (L(S))</td>
<td>4</td>
<td>17284169</td>
<td></td>
</tr>
<tr>
<td>Replicates within L(S)</td>
<td>18</td>
<td>292059</td>
<td></td>
</tr>
<tr>
<td>Genotypes (G)</td>
<td>21</td>
<td>3644949</td>
<td>4.77**</td>
</tr>
<tr>
<td>G x S</td>
<td>21</td>
<td>764412</td>
<td>0.76</td>
</tr>
<tr>
<td>G x L(S)</td>
<td>84</td>
<td>1006974</td>
<td>6.58**</td>
</tr>
<tr>
<td>Plot residuals</td>
<td>378</td>
<td>153101</td>
<td></td>
</tr>
</tbody>
</table>

In this case, the genotype x subregion (G x S) term is not significant when tested against the pooled genotype x location within subregion (G x L(S)) term. Division of this TPE into 2 subregions base on central versus southern location therefore is unwarranted, at least on the basis of this 2004 trial. A combined analysis of an experiment conducted over several years would give a more reliable result.
Summary

1. The purpose of a variety trial is to predict future performance in the TPE.

2. Genotype x environment interaction is large, and reduces the precision with which cultivar means can be estimated.

3. Variance component estimates for the GLY model can be used to develop testing programs that maximize precision for a given level of resources.

4. Within relatively homogeneous TPE, the genotype x site x year variance component is usually the largest. When this is the case, strategies that emphasize either testing over several sites or testing over several years are likely to be successful.

5. Little benefit is obtained from including more than 3 replicates (and often more than 2) in a MET.

6. Standard errors and LSDs estimated from single sites are unrealistically low because they do not take into account genotype x environment interaction.

7. A decision about whether a TPE should be split into 2 separate breeding targets can be made by grouping trial sites or locations into subgroupings based on some fixed environmental factor, and then testing the significance of the genotype x subgroup interaction.

References

MODULE 5

Broad-sense heritability estimates and selection response

5.1 Rationale

In order to plan breeding programs and allocate resources efficiently, breeders must have a clear idea of the *repeatability* or *broad-sense heritability* \((H)\) of estimates of genotypic value. \(H\) integrates information on genetic variation and environmental “noise” into one statistic that is very useful in planning breeding programs. Just as the SEM can be used to model the effect of changes in replication within and across environments on the precision of trials and nurseries, \(H\) is used to in models of selection response to predict the effect of different allocations of screening resources and population size on gains from selection.

5.2 Objectives

1. To define \(H\) for the 2-way and 3-way MET models, and to model the effect of replication within and across sites and years.

2. To describe the relationship between \(H\) and selection response

3. To describe the relationship between \(H\) and the correlation of cultivar means across trials.

4. To model the effects of changes in \(H\) and selection intensity on selection response.
5.3 \( H \) for MET models

*The two-way genotype x environment (GE) model*

We conduct variety trials to predict the performance of the lines under test in farmers’ fields and in future seasons. Therefore, \( H \) estimates need to take into account GEI if they are to be realistic measures of the repeatability of trials. Recall the model (Eq. 4.1) for the combined analysis of variety trials in which each trial is considered an “environment”:

\[
Y_{ijkl} = M + E_i + R(E)_{j(i)} + G_k + GE_{ik} + e_{ijkl} \quad \text{[5.1]}
\]

For this model:

\[
\sigma_p^2 = \sigma_G^2 + (\sigma_{GE}^2 /e) + (\sigma_e^2 /re) \quad \text{[5.2]}
\]

and therefore:

\[
H = \frac{\sigma_G^2}{\sigma_G^2 + (\sigma_{GE}^2 /e) + (\sigma_e^2 /re)} \quad \text{[5.3]}
\]

Where \( e \) and \( r \) are the numbers of environments and replications per environment, respectively. The expected mean squares from the ANOVA of a MET are linear functions of the variances of the factors in the model 5.1. These are given in Table 5.1.

**Table 5.1.** Expected mean squares (EMS) for the balanced ANOVA of the genotype x environment model assuming all factors random.

<table>
<thead>
<tr>
<th>Source</th>
<th>Mean square (t/ha)²</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environments (E)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replicates within E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotypes (G)</td>
<td></td>
<td>( \sigma_e^2 + r\sigma_{GE}^2 + re\sigma_G^2 )</td>
</tr>
<tr>
<td>G x E</td>
<td></td>
<td>( \sigma_e^2 + r\sigma_{GE}^2 )</td>
</tr>
<tr>
<td>Plot residuals</td>
<td></td>
<td>( \sigma_e^2 )</td>
</tr>
</tbody>
</table>
The variance components can be estimated as functions of the mean squares estimated from the ANOVA:

\[ \sigma^2_e = \text{MS}_E \]

\[ \sigma^2_{GE} = (\text{MS}_{GE} - \text{MS}_e)/r \]

\[ \sigma^2_G = (\text{MSG} - \text{MS}_{GE})/re \]

**Example of estimation of \( H \) from the combined analysis over trials for the 2-way model**

Twenty-two (22) groundnut rice varieties were evaluated over 6 locations in 4-replicate trials. The ANOVA is presented below:

**Table 5.2. ANOVA for 22 groundnut varieties tested over 6 sites**

<table>
<thead>
<tr>
<th>Source</th>
<th>Mean square</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environments (E)</td>
<td></td>
<td>( \sigma^2_e + r\sigma^2_{GE} + re\sigma^2_G )</td>
</tr>
<tr>
<td>Replicates within E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotypes (G)</td>
<td>3644950</td>
<td>( \sigma^2_e + r\sigma^2_{GE} )</td>
</tr>
<tr>
<td>G x E</td>
<td>958462</td>
<td>( \sigma^2_e )</td>
</tr>
<tr>
<td>Plot residuals</td>
<td>153102</td>
<td></td>
</tr>
</tbody>
</table>

\[ \sigma^2_e = \text{MS}_E \quad = 153102 \]

\[ \sigma^2_{GE} = (\text{MS}_{GE} - \text{MS}_e)/r = (9958462-15302)/4 \quad = 201340 \]

\[ \sigma^2_G = (\text{MS}_G - \text{MS}_{GE})/re = (3644950-958462)/24 \quad = 111520 \]

For a single 2-replicate trial,

\[ H = \frac{\sigma^2_G}{\sigma^2_G + (\sigma^2_{GE}/e) + (\sigma^2_e/re)} \]

\[ = \frac{111520}{[111520 + (201340/1) + (153102/4)]} \]

\[ = 0.32 \]
**What does H really mean?**

This estimate is the average repetition of 4-replicate trials in this region, managed as these trials were managed and in similar seasons. This value of $H$ is also an estimate of the correlation expected between line means estimated from 4-rep variety trials conducted at different sites in southern and central Laos.

**The effect of replication within and across trials on H**

Of course, inspection of the equation shows that $H$ increases with increasing replication within and across trials. As is the case for reducing the LSD, increasing replication across trials has a greater impact on $H$ than increasing within-trial replication. This effect is made clear by modeling the value of $H$ for the Lao rainfed rice variety trials:

**Table 5.3. Effect of number of trials and replicates on H predicted for groundnut rice variety trials at 6 sites**

<table>
<thead>
<tr>
<th>Trials</th>
<th>Replicates per trial</th>
<th>H</th>
<th>$\sqrt{H}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.24</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.29</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.32</td>
<td>0.57</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.49</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.55</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.58</td>
<td>0.76</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.61</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.67</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.70</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Because selection response is more closely related to $\sqrt{H}$ than to $H$, it is the right-hand column above that shows the true relationship between replication and selection response. Some important features of the response should be noted:

- When GE interaction is large relative to the plot residual variance, there will be much greater benefit to increasing the number of trials than to increasing replicates per trial.
- When more than 3 sites are included in a trial, there is little benefit to including more than 2 replicates per site.
- For variety means estimated from testing programs including more than 5 sites, repeatability may be very high, even with single-replicate trials.

**$H$ estimated via the genotype x site x year (GSY) model**

Recall that the most realistic and complete model for the analysis of cultivar trials recognizes years and sites as random factors used in sampling the TPE. $H$ for the 3-way (GSY) model is:

$$H = \frac{\sigma_G^2}{\sigma_G^2 + (\sigma_{GYy}^2) + (\sigma_{GSs}^2) + (\sigma_{GYs}^2) + (\sigma_e^2)}$$

[5.6]

Variance components for the GLY model are estimated from the balanced ANOVA of a set of trials repeated over locations and years. Variance components can also be estimated for unbalanced trials using the restricted maximum likelihood (REML) method.
Example: Modeling the effect of replication within and across sites and years on $H$.

Recall that Cooper et al. (1999) conducted a resource allocation study for the Thai rainfed lowland rice breeding program in northern Thailand. In this study over 1000 unselected breeding lines from 7 crosses were evaluated for 3 years at 8 sites. Variance components estimates were:

$$
\begin{align*}
\sigma^2_G &= 0.060 \pm 0.006 \\
\sigma^2_{GS} &= 0.003 \pm 0.006 \\
\sigma^2_{GY} &= 0.049 \pm 0.006 \\
\sigma^2_{GYS} &= 0.259 \pm 0.009 \\
\sigma^2_e &= 0.440 \pm 0.006 
\end{align*}
$$

Calculate $H$ for the following scenarios:

a. Testing in a 4-replicate trial at 1 site
b. Testing in 2-replicate trials at 5 sites in 1 year
c. Testing in 4-replicate trials at 5 sites in 1 year
d. Testing in 1-replicate trials at 5 sites in 2 years

The relationship between $H$ and the correlation of line means across trials

As noted above, $H$ is the expected correlation between estimates of cultivar means from independent sets of trials. In the example above, $H$ for a single-site trial with 4 replicates is 0.13. Thus, the expected correlation between line means estimated in independent 4-rep trials is only 0.13!

Upward bias of $H$ estimates derived from a single trial

Note that $H$ estimates for a single trial are biased upwards, because G effects from single trials are actually confounded by (mixed with) the genotype and G x E effects:
This means that the genotypic variance component estimated from a single trial is biased upward by the size of the GxE variance of the real TPE. Estimates of H from single trials are therefore severely inflated, and are not a good guide to predicting response to selection. Only variance components estimated from a series of trials repeated over sites and years within the TPE are useful for this purpose.

**The relationship between \( H \) and selection response**

\( H \) is closely related to the **response, \( R \),** that can be expected from selection. \( R \) is also affected by the **selection intensity** (proportion of the population selected) and the genetic variance in the population:

\[
R = i \sqrt{H \sigma_G} \quad \text{[5.7]}
\]

Where

\( i = \) the selection differential (difference between population mean and mean of the selected fraction) in phenotypic standard deviation sections;

\( \sqrt{H} = \) the square root of the broad-sense heritability of the selection section

\( \sigma_G = \) the square root of the genetic variance.
Eq. 5.7 can be used to model the effect of changes in resource allocation in a breeding program on selection response. If selection intensity remains constant, $R$ is proportional to $\sqrt{H}$.

**Example: using predicted $H$ to compare different resource allocation plans for breeding programs.**

Consider the Thai RL breeding program, with variance components as noted above. What is the predicted effect on $R$ of changing testing from a single 4 replicate trial to 5 2-replicate trials?

Let $H_{1,4}$ and $R_{1,4}$ be predicted $H$ and selection response, respectively, for testing at 1 site, with 4 reps. Let $H_{5,2}$ and $R_{5,2}$ be predicted $H$ for testing at 5 sites, with 2 reps/site. Assuming there is no change in selection differential, the increase in $R$ resulting from increased replication over sites is:

$$R_{5,2}/R_{1,4} = \sqrt{H_{5,2}/H_{1,4}}$$

Using the variance components presented above for the Thai program, estimate the proportional decrease in $R$ expected when testing is reduced from 5 sites, 2 reps per site to 1 site, 4 reps per site.
Summary

- $H$ measures the repeatability of yield trials.
- $H$ takes values ranging from 0 to 1, and is affected by replication
- $\sqrt{H}$ is proportionate to selection response.
- In multi-environment trials, repeatability and selection response are affected more by increasing the number of trials than the number of replicates per trial.
- It is rarely useful to include more than 3 replicates per trial in a trial series repeated over 3 locations or years. For multi-environment trials with 5 or more trials, 2 replicates per site are usually adequate. For METs consisting of 8 or more trials, it is often adequate to have 1 replicate per site.
- Estimates of $H$ are severely inflated when derived from a single trial.
- $H$ is the expected value of the correlation between sets of means derived in different trials.
- Selection response is proportional to $\sqrt{H}$.

References

MODULE 6

Correlations among traits: implications for screening

6.1 Introduction

Many important traits are positively or negatively correlated, because they are controlled by some of the same genes or because they are developmentally or structurally related. An example of a genetic correlation due to a common set of genes might be the association between grain zinc and iron content; varieties that accumulate high concentrations of one element usually also accumulate the other, probably because of a common uptake mechanism. An example of a structural association between traits is the relationship between biomass yield and grain yield; these traits are highly correlated simply because grain yield is a large component of biomass yield. Correlations between genotypic effects for different traits are called genetic correlations ($r_G$).

Breeders are concerned with genetic correlations because:

- They can cause undesired changes in traits that are important but that are not under direct selection. For example, selection for grain yield alone may result in increased height and growth duration, because these traits are often positively correlated with yield.

- Under some circumstances, it may be more effective to conduct indirect selection for grain yield or stress tolerance via selection for a correlated trait than to select directly.

- Analytical methods useful for measuring correlations among traits are also useful in describing the relationship between performance in the SE or screen and TPE. All selection in the SE for performance in the TPE is a form of indirect selection. To predict response in the TPE to selection in
the SE, the genetic correlation between performance in the selection and target must be known, at least roughly.

In this section, we will learn how to estimate genetic correlations, and how these estimates are used to predicting selection response.

6.2 Objectives

• Basic statistical results regarding linear models will be reviewed.
• Genetic and environmental correlations will be defined for traits measured on the same plot.
• A simple method for estimating genetic covariances between traits measured on the same plot will be presented.
• Genetic and environmental correlations will be defined for traits measured on different plots.
• A method for estimating the genetic correlation for line means across environments will be presented.
• Models for predicting correlated response to selection will be presented.
• An approach to determining whether a screening method is effective will be presented.
6.3 Variances, covariances, and correlations

*The product-moment correlation:*

For 2 variables, A and B, the product-moment correlation is:

\[ r = \frac{\sigma_{AB}}{\sigma_A \sigma_B} \] \hspace{1cm} [6.1]

*The variance of a sum*

If \( Y = A + B \), then

\[ \sigma_Y^2 = \sigma_A^2 + \sigma_B^2 + 2 \sigma_{AB} \] \hspace{1cm} [6.2]

6.4 Genetic covariances and correlations for traits measured on the same plot

If 2 different traits (say, height and yield) are measured on the same plot, both genotypic and environmental effects can contribute to the correlation between line means:

\[ Y_A = m_A + G_A + e_A \]
\[ Y_B = m_B + G_B + e_B \]

The genetic correlation is the correlation of the genotypic effects for the two traits:

\[ r_{G(AB)} = \frac{\sigma_{G(AB)}}{\sqrt{\sigma_{G(A)}^2 \sigma_{G(B)}^2}} \] \hspace{1cm} [6.3]
There is also an environmental correlation between plot residuals for different traits.

The phenotypic correlation is the correlation of the line or genotype means for the two traits:

\[
\begin{align*}
    r_{AB} &= \frac{\sigma_{P(AB)}}{\sqrt{(\sigma^2_{P(A)} \sigma^2_{P(B)})}} \\
    &= \frac{\sigma_{G(AB)} + \{\sigma_{E(AB)}/r\}}{\sqrt{(\sigma^2_{G(A)} + \sigma^2_{E(A)}/r)} \sqrt{(\sigma^2_{G(B)} + \sigma^2_{E(B)}/r)}} \\
    &= \frac{\sigma_{G(AB)} + \{\sigma_{E(AB)}/r\}}{\sqrt{(\sigma^2_{G(A)} + \sigma^2_{E(A)}/r)} \sqrt{(\sigma^2_{G(B)} + \sigma^2_{E(B)}/r)}} \quad [6.4]
\end{align*}
\]

Note that, as the number of replicates increases, \( r_p \) approaches \( r_G \). So phenotypic correlations are fairly good estimators of genetic correlations in well-replicated trials.

### 6.5 Estimating \( r_G \) for traits measured on the same plot

There is an easy way to estimate \( r_G \) with any software that performs ANOVA. The method relies on Eq. 11.2. To estimate \( r_G \), we need to estimate \( \sigma^2_{G(AB)} \), \( \sigma^2_{G(A)} \), and \( \sigma^2_{G(B)} \).

We have discussed estimation of \( \sigma^2_{G(A)} \) and \( \sigma^2_{G(B)} \) at length. To estimate \( \sigma^2_{G(AB)} \), we perform the following steps:

1. Add together the measurements A and B for each plot, giving the new combined variable a new name (say Y). This can be done with a spreadsheet.
2. Perform an ANOVA on the new combined variable, and then estimate the genetic variance component using the method described.
3. Re-arrange equation 6.2 to isolate the genetic covariance component:

\[ \sigma_{G(AB)} = \frac{\left(\sigma^2_Y - (\sigma^2_A + \sigma^2_B)\right)}{2}. \] \[6.5\]

**Example**

Calculate the genetic correlation between grain yield and harvest index in a set of 40 varieties tested in a 3-rep trial.

Step 1: For each plot, add the value of HI and GY as in the table below. Call the new variable HIGY

<table>
<thead>
<tr>
<th>Rep</th>
<th>Plot</th>
<th>Entry</th>
<th>GY</th>
<th>HI</th>
<th>GYHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>IC60080-46A</td>
<td>3.418</td>
<td>0.380</td>
<td>3.798</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>IC71524-44</td>
<td>3.345</td>
<td>0.332</td>
<td>3.677</td>
</tr>
</tbody>
</table>

Step 2: Do an ANOVA for HIGY

Step 3: Estimate the genetic variance components for HI, GY, and HIGY

Step 4: Use the results of Step 3 and Eq. 6.5 to estimate the genetic covariance

Step 5: Use the genetic covariance and variance components to estimate the genetic correlation.

**ANOVA for GY, HI, and HIGY**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS for HI</th>
<th>MS for GY</th>
<th>MS for GYHI</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rep</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>(\sigma^2_e + r \sigma^2_G)</td>
</tr>
<tr>
<td>Entry</td>
<td>39</td>
<td>0.0117</td>
<td>2.4377</td>
<td>2.7215</td>
<td>(\sigma^2_e)</td>
</tr>
<tr>
<td>Error</td>
<td>78</td>
<td>0.0022</td>
<td>0.1470</td>
<td>0.1537</td>
<td>(\sigma^2_e)</td>
</tr>
</tbody>
</table>

Calculate variance components, covariance components, and \(r_G\)
6.6 Genetic correlations for the same trait measured in different environments

Often, it is of interest to measure the genetic correlation for yield or another trait in measured in different environments. If this genetic correlation is high, the environments can be treated as part of 1 TPE, and it may be assumed that there is little GEI between them.

Assume that the two trials or environments are called A and B. The model for each site is:

\[ Y_A = m_A + G_A + e_A \]
\[ Y_B = m_B + G_B + e_B \]

If the entries are re-randomized for each site, the G’s are correlated, but the e’s are not. Any covariance across sites is the genetic covariance. As the number of reps within each site or group of environments increases, the line mean correlation (or phenotypic correlation) approaches 1.0.

*Note that the correlation between any 2 environments can’t exceed the repeatability (H) within the environments*

There is an easy way to estimate the genetic correlation across environments, which we will call \( r_{G'} \) to distinguish it from the genetic correlation within environments. To estimate it, we need to know:

- The line mean correlation across environments (\( r_p \))
- H within each of the environments being compared (say \( H_A \) and \( H_B \))

\[ r_{G'} = r_p / \sqrt{(H_A \times H_B)} \] [6.7]
**Example:**

Consider a TPE that we might wish to divide into 2 subregions, A and B. A set of 50 varieties is tested at 3 sites within each subregion. Means are estimated for the varieties over trials within subregions. The phenotypic correlation for means across subregions is calculated as 0.55. Line mean $H$ for means estimated over 3 trials is 0.7 for subregion A and 0.6 for subregion B.

\[ r_{G'} = r_p / \sqrt{(H_A \times H_B)} \]
\[ = 0.55 / \sqrt{(0.7 \times 0.6)} \]
\[ = 0.85 \]

Note that even though the phenotypic correlation across environments was quite low, the genotypic correlation was high. Phenotypic correlations are low because of the obscuring influence of random environmental “noise”.

### 6.7 Predicting correlated response in a target trait resulting from selection for a secondary trait

The main reason for estimating a genetic correlation is to determine if we would have a greater response if we select for a secondary trait than for our target trait. Selecting for a secondary trait when our goal is to improve some other target trait is referred to as *indirect selection*. Indirect selection produces a *correlated response* in the target trait, if the target trait and the secondary trait are correlated. Correlated response in trait A to selection for trait B is predicted as:

\[ CR_A = k r_G \sqrt{H_B} \sigma_{G(A)} \]  
[6.8]
Remember the equation for direct response:

\[ R_A = k \sqrt{H_A} \sigma_{G(A)} \]

If \( k \) is the same for both trait A and trait B, we can determine from these 2 equations if direct or indirect selection is likely to be superior:

\[ \frac{CR_A}{R_A} = r_G \sqrt{H_B}/\sqrt{H_A} \quad [6.9] \]

In other words, indirect selection for a secondary trait will be superior if the heritability of that trait is high, and the correlation between the traits is close to 1.

Occasionally, breeders and physiologists wishing to select for improved performance under a particular environmental stress find it difficult to select directly for yield under that stress. An example of this situation is screening for drought tolerance. It can be difficult to evaluate breeding lines for drought tolerance, because drought occurs irregularly. Many researchers have tried to use secondary anatomical or physiological parameters like root-pulling resistance or root mass to assist in identifying drought-tolerant genotypes. A drought-tolerant genotype is one that produces a high grain yield under a particular type of drought tolerance. Therefore, for a secondary trait to be useful in screening, it must have a high genetic correlation (high \( r_G \)) with yield under stress and must be repeatably measurable (high \( H \)). For practical use in a breeding program, the secondary trait must also be inexpensive and easy to measure in large trials or nurseries.
Summary

- The *phenotypic correlation* ($r_p$) is the correlation of line means for different traits, or for the same trait in different environments.
- The *genotypic or genetic correlation* ($r_G$) is the correlation of genotypic effects free from confounding with the effect of plots or pots.
- Estimates of genetic correlations between traits, or between the same trait measured in different environments, are useful in determining the predictive power of a screen or a selection environment.
- Estimates of genetic correlations are also useful in deciding whether to select directly for a target trait or indirectly for a secondary or correlated trait.
- Genetic correlations can be estimated on different traits in the same experimental sections (plots), or on the same trait in different plots. The methods used for estimating these two types of genetic correlation are slightly different.
- If $r_G$ between the target trait and the secondary trait, or between the target trait measured in the SE and the TPE, is substantially less than 1, it is likely that direct selection will be more effective than indirect selection.
- If $H$ for the secondary trait is less than $H$ for the target trait, then direct selection will *always* be more effective than indirect selection.
Breeding program management

7.1 Choosing parents and managing a pedigree breeding program

Most breeders consider parent choice and the development of fixed lines in their pedigree nurseries as the most important aspect of their work. Experienced breeders choose parents carefully, and manage the development of fixed lines in a systematic way. There are only a few firm rules about parent choice, and many effective ways to manage a pedigree nursery. Pedigree nurseries should be managed to generate a high frequency of adapted lines with the quality, plant type, biotic and abiotic stress resistances, and duration preferred by farmers in the TPE. This section will consider the principles by which pedigree programs can be planned to achieve these goals. Later, we will observe in more detail the methods and systems used.

In this module, the following are presented:

1. strategies for choosing parents for particular objectives
2. the relative advantages of 2-way crosses, 3-way crosses, and backcrosses in generating breeding populations
3. the effect of F2 population size on the probability of recovering desired recombinants.
4. nomenclature by which crosses and derived generations are described.
5. benefits of pedigree versus bulk selection in early generation.
6. the kinds of traits that are most usefully selected for in pedigree breeding programs.
7. the factors involved in determining how many generations to continue pedigree selection before initiating replicated testing.

8. practical features of pedigree nursery design for controlling field variability

### 7.1.1 Strategies for choosing parents

Parents should be chosen carefully to fulfill the objectives of the breeding program.

**Use at least one adapted, farmer-preferred parent**

In an earlier section it was noted that successful breeding programs have very specific objectives, and often aim to develop a replacement for a currently popular variety. For example, a program may wish to replace a popular variety having one or more defects with one of similar quality and adaptation, but with improved resistance to diseases or shorter duration. It follows from this that crosses should generally be between locally adapted and preferred lines and donors carrying the traits that are to be improved in the current variety. Often, a backcross to the adapted parent is advisable, in order to increase the frequency of progeny lines with appropriate plant type and quality.

In general, the donor line should be as high-yielding, high-quality, and adapted as possible while still expressing the trait of interest. It is very difficult (but sometimes necessary!) to use parents that are agronomically poor or of unacceptable quality as donors and still recover useful recombinants. In other cases, a donor parent may have one useful trait but unadapted. In such a case, several cycles of “pre-breeding” are likely to be necessary before unadapted materials of this type can be used. Using non-adapted donors is especially difficult if the trait they contribute is controlled by several genes. If the trait is controlled by a single gene whose location is known, it can be incorporated into the adapted parent via marker-aided selection (MAS).
7.1.2 What kinds of crosses are most likely to produce superior lines?

Two-way, three-way, and double crosses

Breeders often formulate crosses to ensure that all the desirable characteristics they are seeking will be present in the resulting segregating population. This means that breeders mainly use 2-way crosses (crosses between 2 pure lines), 3-way crosses (crosses between pure lines and an F₁, also called topcrossovers), and double crosses (crosses between two F₁s). 3-way and double crosses are used when the breeder wishes to combine a complex set of traits, usually disease, pest, or stress resistances, that cannot be found in any two lines. Some varieties have been successfully developed from such crosses. However, all complex crosses, and even conventional 2-way crosses, have a major disadvantage in that they break apart favorable linkage blocks that have been painstakingly assembled through previous cycles of crossing and selection. The probability that the complex package of genes that makes an elite rice cultivar will occur by chance in the progeny of a double, 3-way, or even conventional 2-way cross is very low. This is why many breeders tend to use large F₂ populations for their initial selections.

Backcrosses

Backcrossing is usually considered to be a strategy suitable mainly for introducing one or a few disease or insect resistance genes into an elite cultivar. Many breeders make few backcrosses in their main cultivar development programs, but there is strong evidence that backcross populations should be used more often breeding. Because the excellent performance and quality of elite pure-line cultivars is the result of very complex combinations of alleles at hundreds or even thousands of loci, breeders should use crossing strategies that keep these gene combinations largely intact, but that still allow the introduction of a few favorable genes. Backcross populations allow breeders to make incremental improvements in an elite, adapted variety used as a recurrent parent by introducing a relatively small
number of genes from a \textit{donor} parent with some characteristics that complement the elite variety’s deficiencies. BC_1 or BC_2 crosses can be used to generate segregating populations that have a high frequency of alleles and allele combinations from an adapted recurrent parent.

There are several applications of the backcrossing program approach. Several elite cultivars can be improved by retaining their favorable features while improving them in several important respects. For example, an elite variety can be improved after selection in the BC_1 and BC_2 generations for the disease resistance and early maturity of the donor and the quality of the recurrent parent. This high-quality, short-duration line will be highly suitable for drought-prone areas.

Another promising application of the backcross breeding approach is selection for drought tolerance in BC_2 populations in the Molecular Breeding Program (so-called because DNA markers are used to track the introgression of donor segments into the recurrent-parent background genotype). This is illustrated below:

<table>
<thead>
<tr>
<th>Cross</th>
<th>% recurrent parent</th>
<th>Plants crossed</th>
<th>Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR64 x donor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F_1 x IR64</td>
<td>50</td>
<td>1-5 crossed</td>
<td>None</td>
</tr>
<tr>
<td>BC_1 F_1 x IR64</td>
<td>75</td>
<td>20 crossed</td>
<td>None</td>
</tr>
<tr>
<td>BC_2 F_2 x IR64</td>
<td>87.5</td>
<td>20 selfed</td>
<td>None</td>
</tr>
<tr>
<td>BC_3 F_3 x</td>
<td>87.5</td>
<td>2000 selfed</td>
<td>Screened under</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>severe stress</td>
</tr>
</tbody>
</table>
BC₂F₂-derived F₃ lines selected from BC₂F₂ plants that produce seed under severe stress exhibit significantly improved yield under drought stress, but, because they are 87.5% derived from the recurrent parent, are usually similar to IR64 in yield and quality under non-stress conditions. The screening of BC₂-derived populations is somewhat difficult, because many crosses of the segregating BC₁F₁ to the donor are needed to ensure that all donor genes are represented in the BC₂F₂ population that undergoes selection. But much of the benefit of the scheme might still be derived if selection were initiated from the BC₁F₂ instead. Selection under stress in BC populations is a promising way to improve stress tolerance in elite backgrounds, while retaining most of the favorable features of the original recurrent parent.

**How many crosses to make?**

The number of crosses made by a breeding program will depend largely on the resources it has available. Breeders often feel they need to explore many crosses, but careful choice of parents can reduce the number of crosses needed. Several breeding programs have achieved success using a small number of very diverse crosses, in which very large F₂ populations were screened. An intermediate approach, where a moderate number (10-30) of carefully-chosen crosses is explored each year through the use of large F₂ or BC₁F₂ populations, is probably optimal for small breeding programs.

### 7.1.3 The effect of F₂ population size on recovering recombinants

If the parents differ by more than a few important genes, as is usually the case, large F₂ populations must be screened to ensure that desired recombinants are recovered. For every locus that is heterozygous in the F₁, there is a 75% chance that any F₂ plant will carry the desired allele, and a 25% chance that it will not. For any number of loci, n, segregating for desired major genes, the probability that any F₂ plant will carry all the
desired genes in the heterozygous or homozygous condition is approximately:

\[ P = (.75)^n \]

If there are 10 loci segregating for major genes affecting quality, duration, and pest resistances, then the probability that any \( F_2 \) plant will be carry the desired gene is about 0.056, or about 1 in 19. If, at the end of our pedigree line development process, we want to have 100 advanced lines carrying the desired traits, each descended from a different \( F_2 \) plant, available for replicated yield testing, we need about 19 \( x \) 100 \( F_2 \) plants from which to select. This corresponds well with the practical experience of many breeders, who usually recommend a minimum population size of about 2000 in the \( F_2 \). If a 3-way or double cross is used, population sizes need to be increased; for crosses of this type, \( F_2 \) populations of 3000-5000 plants should be screened, if possible.

7.1.4 Describing crosses and lines

Crosses

The “slash” nomenclature is useful for describing crosses. In this system, crosses are represented by the “/” symbol or “x”. The most recent cross made receives the highest number of slashes. Thus, if ICGS 34 is crossed to Chico, the cross is described as:

ICGS 34/Chico (or ICGS 34 x Chico)

If this \( F_1 \) is crossed to ICG7878, the resulting three-way cross is written:

ICGS 34/Chico//ICG7878 (or ICGS 34 x Chico xICG7878 x ICG 7878)

Backcrosses are denoted by a number indicating the number of doses of the recurrent parent, followed by an asterisk or
multiplication sign. Thus, if Fleur 11 is crossed to ICGS 34 and then backcrossed, the cross is described as:

Fleur 11/2*ICGS 34

**Lines**

It is important to track the development of lines carefully through a pedigree breeding program. New crosses should be given a unique identifier by which the population is tracked. For example, the cross RRB/ICGV 02801 carries the unique ICRISAT cross number ICGX 08400. In each generation after the F1, in which a line is extracted via single-plant selection, the selected plant number should be attached to the pedigree via a dash. For example, if 100 plants are selected in the F2 derived from ICGX 08400, the line derived from the 35th selected F2 line is denoted:

ICGX 08400-35

If the F2 is advanced in bulk, then the bulk population is denoted:

ICGX 08400-B

The 12th F3 plant selected from this bulk would give rise to a line described as:

ICGX 08400-B-12

If the line from ICGX 08400-B-12 is harvested in bulk, the resulting line is:

ICG 08400-B-12-B.
**Describing the homozygosity of a line**

The proportion of loci that were heterozygous in the F₁ and that are homozygous in Fₙ is simply \(1 - 0.5^{n-1}\).

E.g., in the F₄ generation, the proportion of homozygous loci is, on average, \(1 - 0.5^3\), or 0.875.

**Describing the homogeneity of a line**

The genetic homogeneity of a line, or the proportion of loci (relative to those that were heterozygous in the F₂) fixed in it, is determined by the level of homozygosity of the plant from which the line was selected. Thus, in an F₆ line derived from an F₅ plant, it is expected that, of the loci that were heterozygous in the F₁, the proportion that are fixed is \(1 - 0.5^4\), or \(1 - 0.0625 = 0.9375\).

**Nomenclature to characterize both the homozygosity and homogeneity of a line**

To completely describe the genetic structure of a line, it is necessary to know two things:

- The number of generations of inbreeding
- The inbreeding generation in which the line was established (that is, when it was derived from a single-plant selection).

An F₇ line derived from an F₆ plant is described as an F₆:7 (or \(F_6\)-derived \(F_7\)) line. Such a line is both highly homozygous and highly homogeneous. An F₇ line derived from an F₃ plant by bulking the intermediate generations has the same level of homozygosity, but is much less homogeneous, because the F₃ plant from which it was derived was still heterozygous at many loci.
7.1.5 Pedigree versus bulk selection

Most groundnut breeders use pedigree selection from the F2 generation onward. However, it may be more efficient to inbreed without selection, either in a bulk plot or a rapid-generation-advance facility, until plants are relatively homozygous, allowing homogeneous lines to be established. If bulk generation advance is used in the F2 or later, care must be taken to maintain a large population, so that genetic variability is retained through the inbreeding process. It is generally efficient to advance about 2000 plants per generation by harvesting a single pod per plant, and bulking the harvested pods. Mild selection for pod type, duration, and freedom from leaf disease is often applied in retaining plants for the bulk.

7.1.6 What traits should be the focus of pedigree breeding?

Pedigree breeding is mainly based on visual selection, rather than on agronomic measurements. Therefore, selection should focus on highly heritable traits that are easily visually identified in a small plot or controlled-environment screen. These include, but are not limited to:

- Maturity and flowering date
- Plant type
- Pod size and shape
- Disease resistance

Selection for low-heritability agronomic traits like pod yield, is difficult to do effectively in pedigree nurseries because of the large effect of environmental heterogeneity on these traits. The objective of a pedigree breeding program is to generate a large set of uniform lines that are acceptable in terms of the highly heritable traits listed above. This set is the raw material for selection for yield in replicated trials.
7.1.7 For how many generations should pedigree selection continue?

There is considerable variation among breeders of pure line crops in the number of generations used for pedigree selection. Varieties have been released from lines as early as F2-derived, and as late as F10-derived or later, but most rice lines are probably derived from single F5, F6, or F7 plants. There are two issues to consider when deciding the generation in which to cease pedigree selection and begin replicated testing:

a. The degree of genetic variance left within lines upon which to perform selection;

b. The phenotypic uniformity of lines.

In terms of genetic variability remaining within lines, most is exhausted by the F4 generation. An F4 plant is homozygous at 87.5% of the loci that were heterozygous in the F1. Thus, 87.5% of the genetic variation in a breeding population is found among F4-derived lines not within them. In terms of increasing response to selection, there is little benefit from carrying pedigree breeding on past the F4 or F5.

The main reason to continue pedigree selection to advanced generations (F6 or later) is to achieve adequate visual uniformity of lines that may be segregating for plant type duration, grain type, or other easily-observed traits. Even a small number of segregants for these traits can make a line look ragged and ununiform. Both farmers and breeders like a “flat-top” uniform appearance in new varieties, although farmers are often able to accept considerable variation within landraces. Lines established from F6 or F7 plants are likely to be highly visually uniform.

**Exercise**

Construct a table for lines established from the F2 through the F7 generation indicating the proportion of segregating loci expected within the line.
7.1.8 Controlling field variability in pedigree nurseries

Repeated checks and gridding

Pedigree nurseries can be laid out in any way that permits convenient rating, scoring, and selection. Repeated check varieties should be included at regular intervals for visual reference, permitting the effects of field variability to be assessed, but it is important not to devote too much field space to checks. Generally, a check plot after every 10-20 pedigree lines should be adequate.

Field variation can affect even highly heritable traits like height and days to flowering. In addition to the use of repeated checks, field variation can be controlled by gridded selection, or the practice of making selections only among nearby pedigree lines.

Family selection

Field variability can also be controlled by separating the sublines selected from the same line in a previous generation and planting them in different blocks in the field. This can be done with F$_4$ lines within F$_3$ families, or F$_5$ lines within F$_4$ families. The individual lines within families can be treated as replicates of the same family. This is because closely related sister lines share most of their genes; much more genetic variation occurs among than within families derived from the F$_3$ generation onward. For example, two-sister F$_4$ lines derived from the same F$_3$ plant have a coefficient of coparentage (the probability that an allele drawn from 1 line will be identical to that from the other by descent) of 0.75. For F$_5$ lines from the same F$_4$ plant, the coefficient is 0.875. If F$_{4.5}$ or F$_{5.6}$ sister lines are handled as replicates of the same family (that is, planted in different blocks) the advanced pedigree nursery can then be treated as an observational yield trial. Only the best lines within the best families are selected.
Summary

- Most successful crosses have at least 1 high-quality, adapted parent that is preferred by farmers in the TPE. It is often difficult to use unadapted donor parents directly to develop varieties. Pre-breeding may be needed.

- BC₁- or BC₂-derived populations may be more efficient in terms of generating high-quality, high-yield lines than two-way, three-way, or double crosses because they leave adapted gene blocks in the elite recurrent parents intact.

- At least 2000 plants should be screened in the F₂ (more, if 3-way or double crosses are used.

- A standard pedigree description system should be used, to facilitate exchange of information with other breeding programs.

- Bulk inbreeding in the F₂ and F₃ generations is used by some breeders to inexpensively produce relatively uniform lines in which to begin pedigree selection.

- Pedigree selection should focus only on highly heritable, easily-scored traits. The goal is to produce a large sample of adapted lines of appropriate quality to submit to replicated yield testing.

- Pedigree selection through the F₆ or F₇ generation may be needed to produce visually uniform lines, but there is little genetic variability among F₆ sister lines derived from the same F₅ plant.

- Regularly occurring checks should be included in pedigree nurseries, but their frequency should not exceed 10% of the total number of plots.

- Selection among families of closely-related lines can be used in advanced pedigree generations. Individual lines in a family can be treated as replicates, and planted in different blocks to overcome the effect of field variability. This design combines the features of an advanced pedigree nursery and an observational yield trial.
Breeding improved varieties relies on access to suitable parental germplasm. However, gaining access to suitable parental germplasm is becoming increasingly politicized and legally controlled, subject to developing international agreements as well as to national legislation. How do we determine whether we have the right to use a particular line? In legal jargon, do we have the necessary “Freedom to operate” (FTO)? Who owns the germplasm? How do we respect the rights of the owner? What are those rights? What are we allowed to do with it? Can we donate it to others? Can we use it for breeding and research? Can we sell it to others? Can we claim it as ours? Can we protect it from others? How do we ensure that we do only what we are allowed? How do we assure others that we do only what we are allowed?

These questions are difficult issues. Resolution of the issues is a matter of policy and human rights more than of science. Yet we must understand the issues and comply with the associated policy and law, so that we can achieve our objective of poverty alleviation without contravening human rights.

This module has three principal learning objectives:

1. To learn about changing concepts and developing international agreements regarding intellectual property and germplasm exchange, in order to understand the historical background to germplasm exchange, the problems that have been encountered, the increasing importance attached to intellectual property (IP), and the need for the new regime.
2. To understand the principles of germplasm exchange with IRRI, including an outline of IRRI policy and its use of Material Transfer Agreements.

3. To learn in a practical way how to exchange germplasm with IRRI.

### 8.1 Changing concepts and international agreements

Modern practice in all commercial areas (buying and selling some kind of property – a DVD, for example) is based on separating Tangible Property (TP) from Intellectual Property (IP). The tangible property is the physical thing that we can touch: the DVD itself. The intellectual property is the knowledge and know-how required develop and find out how to produce and market a product – in our example, that could be everything involved in producing a film on the DVD.

This separation of tangible and physical property is necessary to support the heavy investment that modern industries make in producing an affordable product. A film company might invest US$100 million in developing a film, and release it on a DVD worth US$1. The company somehow needs to recoup its US$100 million investment, by selling the DVDs at a price that reflects its investment in making the film on top of the physical cost of the DVD. Yet if anyone else copies that DVD for commercial sale without having invested in making the film, they could obviously easily outcompete the original company.

The solution in trading almost any product is to sell the tangible property but not the associated IP. If you buy a DVD, it is illegal to make and sell a copy. If you buy a packet of rice in the supermarket, you buy the rice for the purpose it is sold – i.e. to cook a meal; but you do not buy the right to use the logo on the packaging. In many countries, when a farmer buys a sack of seeds from a seed company, those seeds can be used
to produce and sell a crop but cannot be retained for further breeding.

The separation of tangible and intellectual property is a major change from traditional practices. Traditionally, there was no distinction between the tangible and intellectual property components of a product that was bought, bartered or exchanged. The owner of the product was allowed to do anything with it.

Use of plant genetic resources has lagged behind the rest of the industry. Until recently, when plant breeders obtained a sample of crop germplasm, they expected to be free to do anything they like with it. Doubts about the acceptability of this doing started to develop during the 1970’s and 1980’s. There was a transition period when the plant breeding sector apparently expected to be able to protect the IP on its improved varieties while having free access to farmers’ germplasm.

Doubts culminated in the ground-breaking Convention On Biological Diversity (CBD) in 1993. This was a key moment in history, redefining internationally accepted concepts relating to the conservation and exploitation of biodiversity. The CBD is now almost universally accepted: 188 sovereign nations are Party to the CBD, leaving only a very few exceptions: USA, Iraq, Somalia, East Timor and a few very small nations. The CBD has 3 key components:

1. Each nation has sovereignty over its own biodiversity
2. Each nation has a right to an equitable share of benefits arising from exploitation
3. Each nation has associated responsibility to conserve its biodiversity

Although the CBD encouraged germplasm exchange, at the same time it raised the required level of negotiation and
authorization from the individual scientist to the government. This had a very negative effect on germplasm exchange.

The International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) was negotiated as a direct response to the CBD, to facilitate access to crop genetic resources in harmony with the CBD through an efficient mutually agreed multilateral system of access and benefit sharing. The ITPGRFA has 2 key components:

1. Each Party should facilitate international germplasm exchange with other Parties, for a specified list of crops that are important for food security and for which countries are highly interdependent.

2. It provides a system to enforce the equitable sharing of benefits, both financial and otherwise, through a central fund providing for obligatory payment to the country of origin.

Thus the Treaty directly addresses concerns that lead to current problems.

### 8.2 Implications for germplasm exchange with ICRISAT

ICRISAT’s fundamental mission is to alleviate poverty and improve the quality of life of farmers and consumers, by sustainable improvement in the semi-arid tropics. In support of this mission, ICRISAT seeks to continue the simple free international exchange of germplasm. At the same time we recognize the critical importance of conforming with all relevant international agreements and national legislation, to ensure that we respect the rights of the IP owners of any germplasm that we use.

To do this, we have to ensure that germplasm is used only legally, and not misused or stolen. This is done through a
Material Transfer Agreement (MTA) for every transfer of germplasm into or out of ICRISAT (with one exception). An MTA sets legally binding terms and conditions that specify how the recipient may use it. The MTA serves to ensure that ICRISAT and its partners know what they can and cannot do with the germplasm; to ensure that we all comply with all relevant national and international law; and to demonstrate to others that we comply with all relevant national and international law.

We need to do this in both directions: sending germplasm to ICRISAT, and receiving germplasm from ICRISAT. For both cases we need to ensure that the sender has the authority to send, and we need to know the conditions under which they can be sent. For both directions of germplasm transfer, we can identify different classes of germplasm that are handled under different rules. The key to efficient exchange of germplasm, therefore, understands the different categories of germplasm and what we can do with them.

Groundnut germplasm may be sent to ICRISAT under one of three groups of conditions:

1. Under the ITPGRFA. This is for germplasm that is in the public domain and under the management of a Party to the ITPGRFA. In this case, the country concerned is required to facilitate access under the multilateral system of the ITPGRFA.

2. Under the CBD. This is for traditional unimproved varieties or wild relatives to be sent from a country that is Party to the CBD but not the ITPGRFA and is either the country of origin of the germplasm or legally obtained the germplasm from country of origin under CBD (i.e. after the CBD came into force). In this case, the CBD encourages the country concerned is required to facilitate access under terms that are set by the CBD authorities in the government of the donor.
3. Germplasm not governed by CBD or ITPGRFA. This can be germplasm from a country that is not party to ITPGRFA or CBD, or for privately-owned germplasm (e.g. commercial varieties with PVP, improved lines and other breeding materials, germplasm from an ITPGRFA Party not in public domain, or, in some countries, wild relatives or traditional varieties in situ), or germplasm outside its country of origin, in a non-ITPGRFA country obtained from country of origin before CBD. In this country, the breeder, institution or farmer concerned has full to decide whether and how to transfer the germplasm to IRRI, without prior approval of the government.

Germplasm may be sent from ICRISAT under one of four groups of conditions:

1. In trust germplasm, which is held in the genebank at ICRISAT and made freely available to anyone who agrees to the MTA prepared by the FAO Commission on Genetic Resources for Food and Agriculture. Basically this MTA says the recipient cannot claim IP on the germplasm. This germplasm is now being distributed under the MLS of the ITPGRFA using an interim MTA. By the end of 2006, a new MTA was established by the Governing Body of the ITPGRFA. The new MTA is similar to the old in terms of preventing the recipient from claiming IP, but will have additional text defining the mechanism of benefit sharing.

2. Germplasm bred ICRISAT or jointly by ICRISAT and its partners. In this case, ICRISAT and its partners have the authority to determine the conditions of access.

3. Germplasm developed by non-ICRISAT scientists. In this case, the donor has the authority to determine the conditions of access.

4. GM germplasm. This is regarded as a special case, needing special-purpose MTAs for each transfer.
8.3 Procedures for germplasm exchange

A germplasm is distributed along with one of the following MTAs: MTA for FAO designated germplasm (genebank materials); MTA for ICRISAT-developed seeds, and MTA for non-ICRISAT seeds. In the same way, all seeds going to ICRISAT should have a MTA. Thus, the first step in germplasm exchange is for the seed importer to accept the terms and conditions of the MTA associated with seeds.

The next step involves the plant quarantine rules and regulations of importing and exporting countries. The seed importer should provide the following documents/information to the exporter: import permit (if required by the importing country), seed treatment allowed; shipping instructions, and customs declaration requirements (if needed).

After seed shipment, the exporter should provide the importer the following shipping details: date sent, shipping company and tracking number. Importer should acknowledge receipt of seeds.

Summary

- Breeders now have to be exceptionally careful to ensure they have the right to use germplasm in their breeding programmes, and to ensure that they use the material legally.

- All international germplasm exchanges, with the exception of restoring germplasm to the country of origin, must now be accompanied by a Material Transfer Agreement (MTA) appropriate to the germplasm.

- We identify three basic classes of germplasm requiring different procedures for germplasm exchange:
  - Germplasm governed by the Multilateral System (MLS) of the International Treaty on Plant Genetic Resources
for Food and Agriculture (ITPGRFA), exchanged using a standard MTA developed especially for the MLS;

- Germplasm governed by the Convention on Biological Diversity, exchanged using a non-standard MTA that must be negotiated each time between the governments of the donor and the recipient

- Germplasm over which the breeder or farmer holds Intellectual Property Rights of a form that give the breeder or farmer authority to determine the conditions of exchange without involvement of governmental officials.

- Germplasm exchange should follow the plant quarantine rules and regulations of importing and exporting countries.

References

Convention on Biological Diversity (CBD):

- Home page: http://www.biodiv.org

International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA):

- List of Parties: http://www.fao.org/Legal/TREATIES/033s-e.htm
MODULE 9

Marker-assisted breeding for cultivar development

The development of DNA (or molecular) markers has irreversibly changed the disciplines of plant genetics and plant breeding. While there are several applications of DNA markers in breeding, the most promising for cultivar development is called marker-assisted selection (MAS). MAS refers to the use of DNA markers that are tightly-linked to target loci as a substitute for or to assist phenotypic screening. By determining the allele of a DNA marker, plants that possess particular genes or quantitative trait loci (QTLs) may be identified based on their genotype rather than their phenotype.

9.1 Advantages of marker-assisted selection

Marker-assisted selection may greatly increase the efficiency and effectiveness for breeding compared to conventional breeding. The fundamental advantages of MAS compared to conventional phenotypic selection are:

- Simpler compared to phenotypic screening
- Selection may be carried out at seedling stage
- Single plants may be selected with high reliability.

These advantages may translate into (1) greater efficiency or (2) accelerated line development in breeding programs. For example, time and labour savings may arise from the substitution of difficult or time-consuming field trials (that need to be conducted at particular times of year or at specific locations, or are technically complicated) with DNA marker tests. Furthermore, selection based on DNA markers may be more reliable due to the influence of environmental factors on field trials. In some cases, using DNA markers may be more
cost effective than the screening for the target trait. Another benefit from using MAS is that the total number of lines that need to be tested may be reduced. Since many lines can be discarded after MAS at an early generation, this permits a more effective breeding design.

The greater efficiency of target trait selection may enable certain traits to be ‘fast-tracked’, since specific genotypes can be easily identified and selected. Moreover, ‘background’ markers may also be used to accelerate the recovery of recurrent parents during marker-assisted backcrossing.

### 9.2 Importance of QTL mapping for MAS

The identification of genes and quantitative trait loci (QTLs) and DNA markers that are linked to them is accomplished via QTL mapping experiments. QTL mapping thus represents the foundation of the development of markers for MAS. Previously, it was generally assumed that markers could be directly used in MAS. However, there are many factors that influence the accuracy of QTL mapping such as population size and type, level of replication of phenotypic data, environmental effects and genotyping errors. These factors are particularly important for more complex quantitative traits with many QTLs each with relatively small effects (e.g. drought tolerance, yield). Therefore, in recent years it has become widely-accepted that QTL confirmation, validation and/or additional marker testing steps may be required after QTL mapping and prior to MAS. These steps may include:

- **Marker conversion** - may be required such that the marker genotyping method is technically simpler for MAS or so that the reliability is improved.
- **QTL confirmation** – testing the accuracy of results from the primary QTL mapping study
- **QTL validation** - generally refers to the verification that a QTL is effective in different genetic backgrounds
• Marker validation – testing the level of polymorphism of most tightly-linked markers within a narrow window (say 5 - 10 cM) spanning a target locus and also testing the reliability of markers to predict phenotype.

9.3 MAS schemes in plant breeding

9.3.1 Marker-assisted backcrossing

There are three levels of selection in which markers may be applied in backcross breeding. In the first level, markers may be used to screen for the target trait, which may be useful for traits that have laborious phenotypic screening procedures or recessive alleles. The second level of selection involves selecting backcross progeny with the target gene and tightly-linked flanking markers in order to minimize linkage drag. We refer to this as ‘recombinant selection’. The third level of MAB involves selecting backcross progeny (that have already been selected for the target trait) with ‘background’ markers. In other words, markers can be used to select against the donor genome, which may accelerate the recovery of the recurrent parent genome. With conventional backcrossing, it takes a minimum of five to six generations to recover the recurrent parent. Data from simulation studies suggests that at least two but possibly three or even four backcross generations can be saved by using markers.
9.3.2 Marker-assisted pyramiding

Pyramiding is the process of simultaneously combining multiple genes/QTLs together into a single genotype. This is possible through conventional breeding but extremely difficult or impossible at early generations. Using conventional phenotypic selection, individual plants must be phenotypically screened for all traits tested. Therefore, it may be very difficult to assess plants from certain population types (e.g. F₂) or for traits with destructive bioassays. DNA markers may facilitate selection because DNA marker assays are non-destructive and markers for multiple specific genes/QTLs can be tested using a single DNA sample without phenotyping. The most widespread application for pyramiding has been for combining multiple disease resistance genes in order to develop durable disease resistance.
9.3.3 Early generation marker-assisted selection

One of the most intuitive stages to use markers to select plants is at an early generation (especially \(F_2\) or \(F_3\)). The main advantage is that many plants with unwanted gene combinations, especially those that lack essential disease resistance traits and plant height, can be simply discarded. This has important consequences in the later stages of the breeding program because the evaluation for other traits can be more efficiently and cheaply designed for fewer breeding lines (especially in terms of field space).

Figure 9.2. Marker-assisted pyramiding of two disease resistance genes. Note that homozygotes can be selected from the \(F_2\) population.
9.4 Current obstacles for the adoption of MAS

There are many barriers to the adoption of MAS in plant breeding. Currently, one of the most important barriers for MAS is the prohibitive cost. Although there are only a small number of reports analyzing the economics of MAS versus conventional breeding in the literature, the cost-effectiveness of using MAS compared to conventional plant breeding varies considerably between studies. Two additional factors need to be considered for cost-analysis: (1) the equipment and consumables required to establish and maintain a marker.

**Figure 9.3.** Early generation selection scheme (proposed by Ribaut & Betran (1999). Note that many lines can be discarded in an early generation which permits the evaluation of fewer lines in later generations.
lab is considerable; and (2) there is a large initial cost in the development of markers which is seldom reported. For marker-assisted backcrossing, the initial cost of using markers would be more expensive compared to conventional breeding in the short term however time savings could lead to an accelerated variety release which could translate into greater profits in the medium to long term.

Another important factor obstructing the successful application of markers for line development is the low reliability of markers to determine phenotype. This is often attributable to the ‘thoroughness’ of the primary QTL mapping study. Even QTLs that are detected with high LOD scores and explain a large proportion of the phenotype may be affected by sampling bias (especially in small populations), and therefore may not be useful for MAS. Furthermore, the effect of a QTL may depend on the genetic background. This emphasizes the importance of testing the QTL effects and the reliability of markers (i.e. QTL/marker validation) before MAS is undertaken.

Finally the level of integration between molecular geneticists and plant breeders (and scientists from other disciplines) may not be adequate to ensure that markers are effectively applied for line development.

### 9.5 Future of MAS in breeding

Despite the relatively small adoption of markers in groundnut breeding to date, there will be a greater level of adoption within the next decade and beyond. Factors that should lead to a greater adoption of MAS in groundnut include:

- establishment of facilities for marker genotyping and staff training within research institutes in different countries
- currently available (and constantly increasing) data on genes/QTLs controlling traits and the identification of tightly-linked markers
development of effective strategies for using markers in breeding

establishment and curation of public databases for QTL/marker data

available resource for generating new markers from DNA sequence data arising from groundnut genome sequencing and research in functional genomics.

It is also critical that future endeavors in MAS are based upon lessons that have been learnt from past successes and (especially) failures in using MAS. Further optimization of marker genotyping methods in terms of cost-effectiveness and a greater level of integration between molecular and conventional breeding (especially in designing efficient and cost-effective strategies) represent the main challenges for the greater adoption and impact of MAS on rice breeding in the near future.

9.6 Suggested reading


About ICRISAT

The International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) is a nonprofit, non-political organization that does innovative agricultural research and capacity building for sustainable development with a wide array of partners across the globe. ICRISAT’s mission is to help empower 600 million poor people to overcome hunger, poverty and a degraded environment in the dry tropics through better agriculture. ICRISAT belongs to the Alliance of Future Harvest Centers of the Consultative Group on International Agricultural Research (CGIAR).

Contact Information

ICRISAT-Patancheru
(Headquarters)
Patancheru 502 324
Andhra Pradesh, India
Tel  +91 40 30713071
Fax  +91 40 30713074
icrisat@cgiar.org

Liaison Office
CG Centers Block
NASC Complex
Dev Prakash Shastri Marg
New Delhi 110 012, India
Tel  +91 11 32472306 to 08
Fax  +91 11 26841294

ICRISAT-Nairobi
(Regional hub ESA)
PO Box 39063, Nairobi, Kenya
Tel  +254 20 7224550
Fax  +254 20 7224001
icrisat-nairobi@cgiar.org

ICRISAT-Niamey
PO Box 12404
Niamey, Niger (Via Paris)
Tel  +227 722529, 722725
Fax  +227 734329
icrisatnsc@cgiar.org

ICRISAT-Bamako
(Regional hub WCA)
BP 320
Bamako, Mali
Tel  +223 20 223375
Fax  +223 20 228683
icrisat-w-mali@cgiar.org

ICRISAT-Bulawayo
Matapos Research Station
PO Box 776
Bulawayo, Zimbabwe
Tel  +263 83 8311 to 15
Fax  +263 83 8253/8307
icrisatzw@cgiar.org

ICRISAT-Lilongwe
Chitedze Agricultural Research Station
PO Box 1096
Lilongwe, Malawi
Tel  +265 1 707297/071/067/057
Fax  +254 1 707298
icrisat-malawi@cgiar.org

ICRISAT-Maputo
c/o INIA, Av. das FPLM No 2698
Caixa Postal 1906
Maputo, Mozambique
Tel  +258 21 461657
Fax  +258 21 461581
icrisatmoz@paninfra.com

Visit us at www.icrisat.org