Chapter 2
Experimental Designs for Next Generation Phenotyping

Luiz Alexandre Peternelli and Marcos Deon Vilela de Resende

Abstract The increase in popularity of high-throughput genotyping in breeding programs is associated with recent advances in DNA sequencing technology and large decreases in genotyping costs. However, the limits of using genotyping for making predictions and, therefore, identifying potential candidate materials for selection thus reside in the quality of the phenotyping. High-throughput phenotyping technologies have been developed and implemented prior to planting and during cultivation. Much of this phenotyping has occurred in relatively small and restricted environments where many influential factors in the quality of phenotype can be adequately controlled. In many situations, however, it is necessary to perform phenotyping under field conditions. In this case, depending on the characteristic of interest to be collected, the influence of factors difficult to be controlled in such adverse conditions can cause the need for use of alternatives that can ensure a sufficiently accurate and precise phenotyping. In this sense, the science of Statistics contributes with an important role, either in the use of traditional basic concepts, in the planning of controlled experiments, or in modeling and developing appropriate analyzes. This chapter will discuss several experimental designs that can potentially be used for phenotyping under variable conditions, describing their various characteristics. Also it will address on topics related to the problem of obtaining accurate and precise phenotypic information, and the role of statistics in the success of this venture so fashionable today.

L.A. Peternelli (✉)
Universidade Federal de Viçosa, Viçosa, Brazil
e-mail: peternelli@ufv.br

M.D.V. de Resende
Embrapa Florestas, Universidade Federal de Viçosa, Viçosa, Brazil
e-mail: marcos.deon@ufv.br

R. Fritsche-Neto and A. Borém (eds.), *Phenomics*,
DOI 10.1007/978-3-319-13677-6_2

2.1 Introduction

Genotyping is becoming increasingly routine and more widely accepted in breeding programs. This increase in popularity is associated with recent advances in DNA sequencing technology and large decreases in genotyping costs. As high-throughput genotyping can be performed with satisfactory quality, the limits of using genotyping for making predictions and, therefore, identifying potential candidate materials for selection thus reside in the quality of the phenotyping (Lado et al. 2013). Genotyping is now highly mechanized and uniform across organisms, but phenotyping methods still vary by species, are laborious, and are sensitive to environmental variation (Cobb et al. 2013). The ideal situation would be a phenotypic characterization that does not have any errors and therefore reproduces the true population or individual phenotypic value, at least for the conditions under which it is measured. However, to obtain an accurate predictive model, the genetic differences among the materials and the experimental conditions that affect the precision of the phenotypic value should be taken into consideration.

Regarding field experiments in which the breeder will select the materials, detailed knowledge of the field conditions and the material being selected is essential for a successful breeding program. To obtain this information, high-throughput phenotyping technologies have been developed and implemented prior to planting and during cultivation. When possible, characterization of the experiment before planting provides better information on the heterogeneity of the field and therefore allows one to define experimental strategies for subsequent, more accurate phenotyping studies. In addition, the measurements made during crop growth seek to reduce the variance caused by any nongenetic factor to which the material may still be subjected (Cabrera-Bosquet et al. 2012; Masuka et al. 2012; Crossa et al. 2006).

If the researcher is unable to use advanced phenotyping technologies or can only use a limited aspect of these technologies, traditional methods can be applied to studies, including effective experimental designs that can capture a large portion of the field variance, as well as correction methods employed during modeling and data analysis. In this context, various strategies can be employed, including strategies for spatial analysis that involve modeling the covariance matrix of the errors and the polynomial functions of rows and columns for fitting spatial trends.

Genetic analysis of field materials has two aims: (i) to infer the genotypic values of the materials and (ii) to rank the genetic materials by their genotypic values. Clearly, there is no interest in estimating the phenotypic means of the genetic materials in the experiments aimed at estimating the genetic means, also known as the genotypic values. In other words, the researcher is interested in future means, when the materials are planted again on commercial farms after the selection process. When planted commercially, even when planted at the same site or in the same region as the experiment, the effects of blocks and plots and the random environmental effects will not be repeated. As these effects are included to an extent in the phenotypic means, they are not sufficient to draw conclusions concerning the

genotypic values of the genetic materials. Thus, utilizing phenotypic means for predicting results of subsequent studies is not desirable or recommended. On the contrary, the breeder is interested in the genotypic values free of environmental effects. These should be the values used for analyses of future outcomes (e.g., subsequent analyses based on molecular marker linkage, genomic selection, quantitative trait loci identification, and differential gene expression analysis by RNA-seq) based on genotyping data, thus allowing for better model predictions and ensuring conclusive results.

However, phenotyping via field experiments is generally associated with unbalanced data for several reasons, including plant and plot losses, unequal numbers of seeds and seedlings available for treatments, experimental networks with different numbers of replicates and different experimental designs, and non-evaluation of all combinations of genotypes and environments. In addition, when the automated collection of phenotypic data is impractical and a group of researchers analyzes the materials in the field, researcher bias can decrease accuracy. Thus, statistical models should include all of the sources of variation and noise to better "correct" the measured phenotypic values. Therefore, the optimal procedure for genetic analysis is restricted or residual maximum likelihood/best linear unbiased prediction, also generically called *mixed linear models*. Mixed-model data analysis allows for various sources of variation to be included in the model, without impeding analysis. In addition, these models seamlessly handle unbalanced data, leading to more precise estimations and predictions of genetic parameters and genetic values, respectively.

Currently, the development of effective phenotyping methods requires multidisciplinary collaboration involving biologists, agronomists, computer scientists, engineers, and statisticians (Cobb et al. 2013). The level of expertise required is related to the use and development of equipment for automated and efficient data collection, the definitions of the variable of interest to be collected, appropriate field conditions for plant growth and analysis, volume of data to be collected, stored and analyzed and, finally, the planning of experiments to better control for systematic variations. For data analysis, the wide availability of computer resources (software and computational power) has facilitated the work of statisticians during experimental planning. In the past, it was common to have restrictions for implementing various experimental and field data collection designs because the theoretical knowledge and available computational power were limiting factors. Now, it is possible to obtain more accurate means (or effects) for complex experimental designs in the context of mixed linear models, therefore ensuring greater effectiveness of subsequent analyses that require sufficiently accurate phenotypic values.

In addition to mixed models, Bayesian analyses have facilitated data analysis and have increasingly ensured that one can obtain adjusted data with the desired quality. Bayesian analysis provides more precise estimates of variance components, genetic parameters, genetic values, and genetic gains, in addition to allowing for accurate analyses of samples with finite sizes. The informational richness provided by this approach allows for the determination of point estimates and probability intervals for the posterior distributions of the parameters. The great advantage is

that it is a modeling approach whereby, via the prior distributions of the effects and model parameters, a researcher can incorporate future knowledge regarding the problem in question. More details on this data analysis approach can be found in the literature (Silva et al. 2013; Resende 2002).

Although there is consensus in the numerous publications that address the importance of increasing the accuracy of phenotyping in field studies, little has been noted regarding the experimental designs that would be the most appropriate for experiments in which large-scale phenotyping is desired. This chapter will discuss several experimental designs that can potentially be used for this purpose, describing their various characteristics and results to the extent that the reader can implement them satisfactorily.

2.2 Basic Principles of the Experiments

Experiments differ among studies. However, all experiments are guided by several basic principles established at the beginning of the twentieth century by Fisher in several of his publications (Fisher 1926, 1935). The use of these principles (replication, randomization, and local control) is necessary to obtain valid conclusions.

The principle of replication consists of applying the same treatment to several plots within the same experiment for estimating the experimental error, or residual variance.

The principle of randomization provides all of the experimental units the same chance of receiving any of the treatments, thus preventing one of the treatments from being systematically favored or disfavored by external factors. A great benefit of randomization is to provide reliability for the estimates of the experimental error of the means for the treatments. By allowing the experimental error to be validly estimated, this principle ensures the use of significance tests (e.g., comparisons of treatment means) by making the experimental errors independent.

Finally, local control is a commonly applied principle, but it is not obligatory because experiments can be conducted without it. The goal of local control (or blocking) is to divide a heterogeneous environment into homogenous sub-environments. Treatments are distributed within the sub-environment, making the experimental design more efficient by reducing experimental error.

2.3 Experimental Design

There are no explicit citations for the experimental designs most commonly applied for large-scale phenotyping. Several studies (Araus and Cairns 2014; Fiorani and Schurr 2013; Cobb et al. 2013; Poorter et al. 2012) have noted the importance of organizing experiments according to an experimental design that allows for

increasing the accuracy of the phenotypic information, but few studies name these designs (Lado et al. 2013; Auer and Doerge 2010).

Because the main interest of the researcher when evaluating various phenotypic characteristics is to better characterize the material under analysis, by destructive means or not, it is expected that collecting the most accurate data is of utmost importance. In this context, the term *accuracy* should be well understood. Figure 2.1 illustrates the concepts of accuracy and precision.

Accuracy is defined as the correlation between the true genotypic value and the value estimated from the genotypic and phenotypic data from the experiments. An accurate estimator has a small difference between the true and estimated values, that is, it has a small mean squared error (MSE). An optimal estimation/prediction method should minimize the MSE, given by MSE = bias^2 + precision = bias^2 + PEV, where PEV is the prediction error variance. Thus, a minimum MSE estimator has little or no bias and high precision (low PEV). With no bias, MSE = PEV.

The concepts of bias, precision, and accuracy are illustrated in Fig. 2.1. High accuracy (the capacity to hit the target) is a combination of high precision (low variance in the various attempts; i.e., low PEV) and low bias (mean of the various attempts equal to the prediction target). Thus, accuracy is the ability to identify the truth, and precision is the ability to always obtain the same answer but not necessarily the truth.

Designs recognized as having potential to improve the effectiveness (less prediction variance) of phenotyping in field experiments include the randomized complete block design (RCBD), the augmented block design (ABD), and the incomplete block design (IBD), with their possible variations.

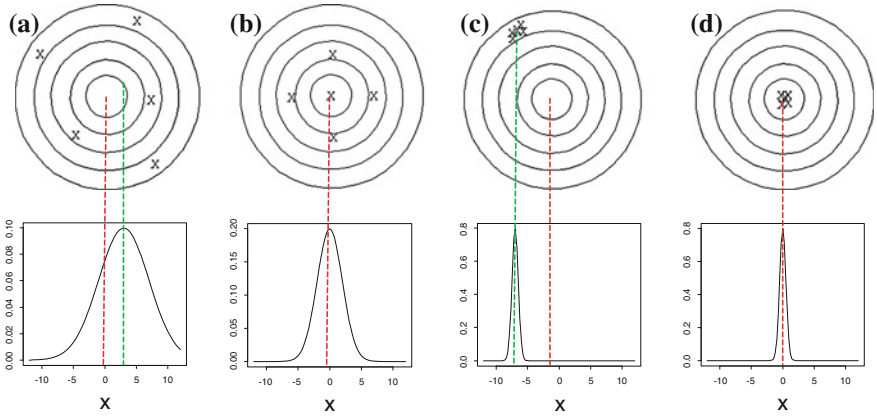

Fig. 2.1 Illustration of the concepts of accuracy, precision, and bias. **a** High bias, low precision ⇒ low accuracy; **b** low bias, low precision ⇒ low accuracy; **c** high bias, high precision ⇒ low accuracy; **d** low bias, high precision ⇒ high accuracy. The *vertical red line* shows the true value (target value). The *vertical green line* shows the prediction bias. The shape of the *curve* shows the precision: a curve more concentrated at the mean implies higher precision (low PEV), while a curve less concentrated at the mean implies less precision (high PEV). PEV, prediction error variance

ABD is most commonly used in the initial steps of a breeding program when there is still a substantial amount of material to be analyzed and, mainly, when there is little propagation material (and thus, the replication of treatments is difficult or impossible). One advantage of ABD is the ease of establishing the experiments, which is particularly useful for sugarcane breeding programs, for example (Souza et al. 2006; Peternelli et al. 2009). RCBD, in turn, is more commonly used in the later stages of breeding programs when, in addition to possessing sufficient propagation material to perform several replicates, more reliable conclusions concerning the analyzed treatments are desired. In contrast, RCBD is unviable when the number of treatments is large. Under this scenario, the block will be very large and will likely encompass heterogeneous conditions, thus limiting the efficiency. IBD, in turn, is employed when the block size is smaller than the number of treatments. If the researcher purposefully creates the blocks according to the number of treatments, the blocks may become very large, which will result in environmental heterogeneity within the block, thus leading to high prediction error. For this reason, IBD is preferred when homogeneity within the block is needed. This homogeneity can be guaranteed, for example, when the block to be homogeneous could only contain 20 plots, but the researcher must phenotype more than 20 different materials (treatments).

Several other aspects regarding these designs will be discussed. Theoretical and practical details of the analysis of these designs can be found in Resende (2007), Faraway (2005), Ramalho et al. (2005), Hinkelmann and Kempthorne (1994, 2005), Barbin (2003), Storck et al. (2000), Steel et al. (1997), Scott and Milliken (1993), Cochran and Cox (1992), Banzato and Kronka (1989), and Cox (1958).

2.3.1 Randomized Complete Block Design

RCBD is the most widely used of all of the experimental designs. It is suitable when there is complete homogeneity in the experimental conditions. In this case, the experimental area or material is divided into blocks (or groups), maintaining homogeneity within each block, and each block contains at least one replicate of each treatment distributed randomly within each block (Fig. 2.2).

Replicate 1		
1	4	7
2	5	8
3	6	9

Replicate 2		
1	2	3
4	5	6
7	8	9

Fig. 2.2 Layout of an experiment employing a RCBD with nine treatments. There are two replicates in this arrangement (often called blocks). Treatments are numbered *1–9*. In an RCBD, if one wants to add control treatments, the controls are allocated to new plots within each replicate. Within each replicate (or block), the treatments are allocated randomly

In experiments with this design, the blocks should be defined in a layout that confers homogeneity to each block. Theoretically, it does not matter if the experimental conditions in one block differ from the experimental condition of another block because these differences do not cause treatment × block interactions. This lack of interaction means that comparisons between pairs of treatments, for example, are not affected by the block in which these treatments are established.

It is important to emphasize that the use of an RCBD when it is not necessary results in a loss of efficiency and a decrease in the precision in the experiment. However, in general, it is necessary to divide the experimental area into homogeneous blocks that contain the treatments. Thus, this type of design is widely used for field conditions.

2.3.2 Augmented Block Design

ABD has been employed in various phases of breeding programs (e.g., for sugarcane). Initially proposed by Federer (1956), an ABD allows for genotypes to be analyzed without using replicates; only the controls are replicated (Fig. 2.3).

The experimental error can be estimated from the controls. This design is a type of IBD and is commonly called Federer blocks, in honor of its creator. This design is unbalanced and nonorthogonal. Thus, it should be analyzed using a mixed-model method.

The establishment of an ABD is very simple. It starts similarly to an RCBD with controls. However, the treatments, or new materials, are distributed among these blocks but not replicated between blocks. The statistical analysis of this design entails a fit of the "effects" attributed to each treatment, corresponding to a penalization of the treatments allocated to the best blocks and a bonus for the treatments located in the worst blocks.

In certain instances, two replicates are included when enough material is available to obtain a better estimate for the effects of each treatment, thus doubling the material requirements and operational costs of the program and reducing the area available for other goals or reducing the number of clones available in the area. However, by keeping the size of the experimental area constant, numerous studies have demonstrated that this practice of doubling the ABD does not necessarily

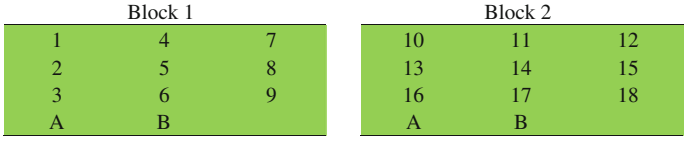

Fig. 2.3 Layout of an ABD, with 18 treatments and two controls. There are two blocks in the arrangement. *A* and *B* are the controls. Treatments are numbered *1–18*. All of the treatments and controls are randomly distributed across the blocks

result in gains in estimates of treatment effects (Peternelli et al. 2009). The greater difficulty is in defining the material that should be used as a control, thus providing the estimate of the experimental error. It is possible that this estimate is influenced by the choice of controls. Therefore, the researcher should use this design with care.

2.3.3 Incomplete Block Design

As mentioned previously, the heterogeneity within very large blocks will lead to a larger experimental error, which makes the phenotypic estimates of interest less precise by reducing the precision of the experiment. In the IBD design, the blocks are smaller, leading to less environmental heterogeneity within the blocks. The theory behind the planning and use of this design is extremely complex. Below, we briefly describe several concepts and peculiarities underlying IBD. However, there are several other important concepts and details of the analysis that must be addressed and are important to note. Valuable references on this topic are cited at the end of this chapter.

IBD designs can be classified into two categories: *resolvable designs*, in which the blocks can be grouped into replicates, and *nonresolvable designs*, in which the blocks cannot be grouped into replicates. Resolvable designs are preferred because analyses can be performed, when necessary and possible, using the completely randomized block design.

For the explanations below, the following definitions apply: v = number of treatments, k = size of the blocks or number of plots within each block, b = number of blocks, and r = number of replicates in the experiment.

Suppose that there are r replicates and v treatments. Additionally, suppose that within each replicate there are b blocks, each of size k. Figure 2.4 provides an example of this design.

In this layout, the blocks can be grouped into treatment replicates (*resolvable design*). Some authors (Williams and Matheson 1994) call this a *generalized lattice design*.

A balanced lattice square design is the most efficient IBD design if the aim is to compare two treatments. In this design, $v = k^2$, which may restrict its use in practice. To be balanced, $r = k + 1$. The high efficiency of the balanced IBD design is attributed to all of the treatment pairs occurring in at least one block of the experiment. However, for all of the treatment pairs to occur together at least once, a large number of

	Replicate 1			**Replicate 2**		
Block	1	2	3	1	2	3
	1	4	7	1	2	3
	2	5	8	4	5	6
	3	6	9	7	8	9

Fig. 2.4 Layout of an experiment using an IBD with nine treatments. There are three blocks (b = 3), two replicates (r = 2), nine treatments (v = 9) and the blocks have a size k = 3

replicates is needed ($r = k + 1$), which could prevent the implementation of a balanced IBD in practice. In this case, a partially balanced lattice square can be established. This design is obtained by considering a number of the $k + 1$ replicates from a balanced design. Therefore, another more practical design that can be implemented in the field is the alpha-lattice design (Patterson and Williams 1976), in which the v treatments are arranged in b blocks of size k, such that $v = ks$, for $s > 1$.

An advantage of alpha-lattice over the lattice square is the ability to use a large number of v values. That is, the relationship $v = ks$ in the alpha-lattice is less restrictive.

2.4 Modeling and Appropriate Analyses

There are various types of analyses for incomplete block experiments: (a) intrablock analysis, in which comparisons are only made between plots in the same block to estimate the treatment effects; and (b) analysis with recovery of interblock information, in which comparisons between blocks are also used to estimate treatment effects. Because it provides more precise results, the latter type of analysis is used by most computer programs in the context of mixed-model analyses.

If the researcher has additional available information that can contribute to a better fit (correction) of the phenotypic values collected in the field, additional analyses can be incorporated into the design model. Several examples are discussed in the following sections.

2.4.1 Covariance Analysis

If supplementary information that can somehow predict the performance of the experimental units is available, which would be the case when several variables are collected in the experiment, it is sometimes possible to estimate the extent to which the observations of interest were influenced by the variations in these supplementary measurements. The aim of these analyses would then be to adjust the mean response of each treatment to remove the experimental error from this external source, the covariate. Thus, the variance from the supplementary variable is removed from the experimental error, without having to include this variable in the experimental design. In summary, the usefulness is the removal of the experimental error that arises from external sources of variation, which would be impractical or very expensive to control for using more refined techniques. A typical covariate is the stand of plants per plot, which varies between plots and, therefore, should be controlled for during analysis and not by design.

Aulchenko et al. (2007) proposed fitting the model $y = Xb + Zg + e$, which yields $\hat{e} = y - X\hat{b} - Z\hat{g}$ after fitting, where g is a vector of polygenic effects. The model $\hat{e} = 1u + Wm_i + e$ is then fit to the residuals ($\hat{e}$) to identify the significant

markers (m_i). This analysis seeks to capture only the effects associated with Mendelian segregation, which arise only from the linkage disequilibrium between markers and genes. Thus, this approach is applicable to genome-wide association studies (GWAS) and genome-wide selection (GWS) of advanced generations, as opposed to the training population. Conversely, the fitting of the model $y = Xb + e$ is applicable to GWS in the current generation and in the short term (a few generations after the training population) and contains both the genetic effects from Mendelian segregation and those explained by genealogy (which are contained in the residuals $\hat{e} = y - X\hat{b}$), which are used for genomic analyses. In both models, the data in y are adjusted for the effects of the covariates in $\hat{b}$.

2.4.2 Spatial Analysis

The researcher will often want to conduct his or her experiment in a new area and, therefore, does not have in-depth knowledge of the spatial heterogeneity of the site where the experiment is being implemented. Thus, when the heterogeneity is unknown a priori, the definition of blocks becomes arbitrary, which can result in strong heterogeneity within blocks, thus causing a decrease in the efficiency of the chosen design. When the program does not have the technology and resources required for the high-resolution collection of data on spatial variation in variables at the study site, one alternative is to randomly allocate plots of a single plant in the experimental field and then control for environmental heterogeneity by using covariance analysis to correlate a covariate with the studied variable (Papadakis method: Papadakis 1984) or by using regional or spatial variables (geostatistical methods). Potentially, a posteriori fitting of the environmental gradients in progeny tests may significantly increase the effectiveness of the selection of genetic parameters. Thus, establishing randomized plots of a plant (completely randomized design) is important. However, Gilmour (2000) advises that a posteriori blocking should not be based solely on the statistical significance of arbitrary contrasts. The researcher should identify the physical and environmental causes that lead to a given type of blocking. If the number of treatments and partitions allows the researcher to use a certain, efficient experimental design, they can reduce the need for a posteriori fitting techniques (e.g., spatial analysis; Resende 2002).

2.4.3 Polynomial Functions for Rows and Columns for Fitting Spatial Trends

This method is based on the procedure proposed by Federer et al. (2001), which basically involves the selection of a polynomial function for the rows and columns that refer to the coordinates of the experimental plots to better absorb the random

variations inherent in the data, according to the model for the design implemented. The mean values associated with each treatment will thus be corrected by this function, providing adjusted values that are used in subsequent analyses.

2.5 Important Considerations

2.5.1 More on Accuracy and the Number of Replicates

Figure 2.5 illustrates the accuracy of the data collected for a given individual as a function of the number of replicates of that individual (pure line or clone; i.e., absent of genetic variability but with environmental variability). If the trait follows a normal distribution with a mean $\mu = 10$ and $\sigma^2 = 4$, sampling replicates (e.g., $r = 1$, 2, 3, 4, 5) produce the plots shown in Fig. 2.5. When the environmental variability can be removed by blocking, the precision of the estimate (even with only one replicate) is much larger; that is, the curve will be more concentrated around the true value μ.

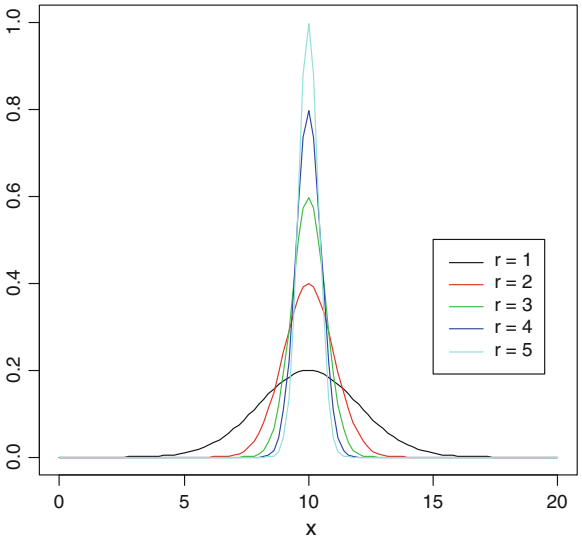

Fig. 2.5 Illustration of the range of values for trait X that can be obtained from various numbers of replicates of the experimental material. In this case, we are assuming $X \sim$ Normal (10, 4), which exhibits a CV = 20 %. Thus, with five replicates of the experimental material, it would be practically impossible to obtain a mean value greater than 12, but sampling only one replicate ($r = 1$) would likely yield values from 5 to 15. It is assumed that there is no genetic variability

2.5.2 Plot Size and the Number of Replicates

Several studies have confirmed that designs with a small number of plants per plot and numerous replicates are more efficient than those with numerous plants per plots and a small number of replicates (de Resende 2002). This relative superiority comes from the following: (a) the higher precision in the comparisons between treatments because of the greater number of replicates for a fixed-size experimental area; (b) the greater selective accuracy because of the greater number of replicates in a fixed-sized experimental area; (c) the greater individual heritability in the block because of the creation of more homogeneous blocks; (d) the lower overestimation (from any genotype × environment interaction) of heritability and genetic gain in a site because of the greater number of replicates analyzed (which may represent various environments); and (e) the smaller size and greater homogeneity of the block, reducing the need for spatial analysis of the experiments because local control is more effective.

As will be discussed, the plot should be considered the observational unit for data collection. For example, in sugarcane, the concept of plots of one plant should be interpreted as one furrow per plot for situations when all plants of a furrow are combined together in a composite sample to proceed with analyses. In this case, the individual heritability is defined as the heritability of a furrow and the number of replicates is determined as a function of the magnitude of this heritability at the furrow level.

The determination of sample sizes (in terms of numbers of replicates) for the estimation and prediction of various practical genetic breeding scenarios are explained by Resende (2002). To determine sample size, the criteria chosen was the maximization of the selective accuracy (the correlation between the true and estimated genetic values) as the number of replicates was increased (Table 2.1).

For example, with a heritability of 40 %, an accuracy of 90 % can be obtained with approximately seven replicates per clone. These are the recommended numbers per site. When there is a considerable genotype × environment interaction and a large planting area, experiments should be repeated in other sites before selection to minimize the adverse effects of the genotype × environment interaction.

When a researcher is conducting experiments with families, the genetic variability within the family contributes to the complexity of the problem. Thus, one must know the number of genotypes representing the families under study to determine the appropriate plot size and the number of replicates. In sugarcane breeding, for example, recommendations in the literature vary from 16 to 150 plants per family. In addition, the recommendations vary greatly depending on the parameter to be estimated and the type of trait to be analyzed (Peternelli et al. 2012; Leite et al. 2006, 2009).

A general approach for choosing sample size uses the confidence interval (CI; a 95 % CI is considered appropriate) for the sample mean ($\bar{y}$) of a normally distributed population. In this case, $CI = \bar{y} \pm 1.96\ s(\bar{y})$, where $s(\bar{y}) = (\hat{\sigma}^2/n)^{1/2}$ = the standard error of the mean. Thus, one can set a tolerance error (δ) in the estimate of the mean,

Table 2.1 Adequate number (N) of replicates per clone, in clonal tests, as a function of individual heritability (in one plot) (h_g^2), broadly speaking, to obtain an accuracy ($r_{\hat{g}g}$) of 90 and 95 %

h_g^2	N for $r_{\hat{g}g}$ = 90 %	N for $r_{\hat{g}g}$ = 95 %	h_g^2	N for $r_{\hat{g}g}$ = 90 %	N for $r_{\hat{g}g}$ = 95 %
0.05	81	176	0.40	7	14
0.10	39	84	0.45	6	12
0.15	25	53	0.50	5	10
0.20	18	38	0.60	3	7
0.25	13	28	0.70	2	4
0.30	10	21	0.80	2	3
0.35	8	17	0.90	1	2

given by $\delta = 1.96\ \mathrm{s}(\bar{y}) = (\hat{\sigma}^2/\mathrm{n})^{1/2}$. From this expression, $\mathrm{n} = (1.96^2\hat{\sigma}^2)/\delta^2$, which is the adequate sample size for an error tolerance of δ. The error tolerance is chosen by the researcher. If σ^2 is unknown, the estimated $\hat{\sigma}^2$ and t value (1.96) from Student's distribution is used in place of the z-score of the normal distribution.

Thus, to determine n, the δ for the estimate of the mean must be chosen and there must be an estimated $\hat{\sigma}^2$ for the phenotypic variability of the population. The error δ can be specified as a percentage of the mean (e.g., 10 %). In this case, δ is given by $0.10\bar{y}$.

Thus:

$$n = \frac{1.96^2\hat{\sigma}^2}{(0.10\bar{y})^2} = \frac{1.96^2\hat{\sigma}^2}{0.10^2\bar{y}^2} = \frac{1.96^2}{0.10^2}\left(\frac{\hat{\sigma}}{\bar{y}}\right)^2 = \frac{1.96^2}{0.10^2}\mathrm{CV}^2;$$

where CV is the coefficient of variation of the trait in the population (Resende 2007). Using this approach, only an estimate or prior knowledge of the individual phenotypic CV in the population is required: the larger the CV, the larger the required sample size (Fig. 2.6).

For binomial variables, using the approximation to the normal distribution, the same expression for n can be solved by replacing $\hat{\sigma}^2$ with $p(1 - p)$ and $\bar{y}$ with p,

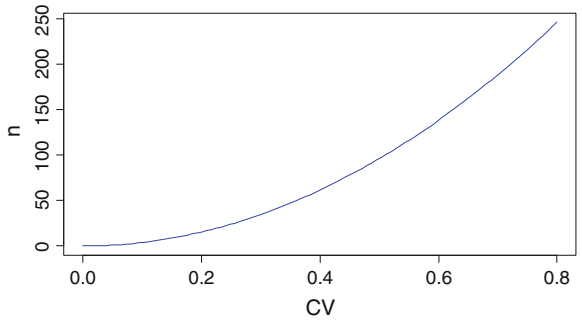

Fig. 2.6 Sample size (n) as a function of the phenotypic coefficient of variance (CV) of the trait in the population

where p refers to the observed proportion of the phenotypic class defined as "success." In the absence of information on p, one can use $p = 0.5$, which guarantees the largest variance possible.

2.5.3 Genetic Sampling and Effective Population Size

The genetic representativeness or effective size of a family is relevant to phenotyping in two aspects: (i) determining the size of the family in the experiment and (ii) obtaining adequate genetic representativeness of populations.

The effective size of full-sib families is given by $N_{ef} = (2n)/(n+1)$, where n is the number of individuals per family. The values of Ne for various values of n are listed in Table 2.2. This table also provides the results for half-sib and S1 families, which are discussed later.

Table 2.2 lists the number of individuals per family necessary to achieve a specific percent of the maximum N_{ef} of the family. An sample size of 100

Table 2.2 Effective size (N_{ef}) and the fractions of the maximum effective size (N_{efmax}) of a full-sib, half-sib, and S1 family as a function of the number (n) of individuals sampled per family

n	Full-sib		Half-sib		S1	
	N_{ef}	Fraction of N_{efmax}	N_{ef}	Fraction of N_{efmax}	N_{ef}	Fraction of N_{efmax}
1	1	0.500	1	0.250	0.670	0.670
5	1.667	0.833	2.5	0.625	0.909	0.909
7	1.750	0.875	2.8	0.700	0.933	0.933
10	1.818	0.910	3.1	0.775	0.952	0.952
12	1.846	0.923	3.2	0.800	0.960	0.960
15	1.875	0.938	3.3	0.825	0.968	0.968
18	1.895	0.947	3.4	0.850	0.973	0.973
20	1.905	0.952	3.5	0.875	0.976	0.976
25	1.923	0.962	3.6	0.90	0.980	0.980
30	1.935	0.968	3.64	0.91	0.984	0.984
40	1.951	0.976	3.72	0.93	0.987	0.987
50	1.961	0.980	3.77	0.94	**0.990**	**0.990**
60	1.967	0.984	3.88	0.97	0.992	0.992
100	**1.980**	**0.990**	3.88	0.97	0.995	0.995
150	1.987	0.993	3.92	0.98	0.996	0.996
200	1.990	0.995	3.94	0.985	0.997	0.997
250	1.992	0.996	3.95	0.988	0.998	0.998
300	1.993	0.997	**3.96**	**0.990**	0.998	0.998
∞	2.000	1.000	4.00	1.00	1.00	1.00

Source Adapted from Resende and Barbosa (2005)

individuals will encompass 99 % of the maximum representativeness of the full-sib family. Therefore, increasing the sample size above 100 contributes almost nothing to increasing the representativeness for a family.

For half-sib families, the effective size of a family (N_{ef}) is given by $N_{ef} = (4n)/(n+3)$, where n is the number of individuals per family. For half-sib families, 300 individuals provide 99 % of the maximum representativeness of a family (Table 2.2). Because the ideal crossing design for selecting parents and clones within families assumes that three crosses are performed per parent, three full-sib families are associated with each parent. Thus, by adopting a family size of 100 for full-sib families, we obtain a size of exactly 300 for each half-sib family.

The effective size of an S1 family is given by $N_{ef} = (n)/(n+0.5)$ and the maximum equals 1, when n goes to infinity. However, with n = 1, the N_{ef} is already equal to 0.67. With n = 50, N_{ef} is already 0.99—that is, 99 % of the maximum N_{ef} (Table 2.2). One can say that the probability of adding an effectively different individual is less than 1 for each 100 individuals added after n = 50 (or exactly 0.67 for the first 100 after 50). Nevertheless, there would not be sufficient precision for including only this individual in the selection among 150 S1 individuals. Thus, it is believed that 50 individuals is an adequate size for selection in S1 families. The number n = 50 for S1 families is comparable to the numbers n = 100 and n = 300 for the progeny of full and half sibs, respectively. In other words, these numbers (50, 100, and 300) provide 99 % of the maximum representativeness of S1, full-sib, and half-sib families, respectively, and therefore would be adequate sizes of progeny for selection within said families.

In conclusion, 100 individuals per full-sib family and 300 per half-sib family are adequate sample sizes. The 100 individuals from each full-sib family can be divided into two or three environments at multiple sites.

2.5.4 Number of Experimental Sites

The appropriate number of experimental sites can be determined from the selection efficiency (E_f) for the mean of several environments (ℓ) relative to the selection in only one environment aiming to obtain gains in the mean of ℓ sites. This efficiency can be inferred (for heritability, at the level of the mean, similar and tending to 1 in various environments, similar to well-designed clonal tests) by the expression $E_f = [\ell/[1+(\ell-1)r_{gg}]]^{1/2}$, where, r_{gg} is the genetic correlation involving the performance of the germplasm in the environment (Table 2.3; Resende 2002).

The results shown in Table 2.3 demonstrate that when the genetic correlation is equal to or greater than 0.70, the gain in efficiency from analyzing more than one experimental site is less than 10 %. If the genetic correlation is greater than 0.80, the gain in efficiency is less than 5 %. Conversely, using three sites instead of two sites

Table 2.3 Efficiency (in terms of genetic gain in the mean of the sites) of using ℓ sites instead of one site for assessing genetic material, for various values of the genetic correlation (r_{gg}) involving the performance of the germplasm in the environment

r_{gg}	ℓ	E	r_{gg}	ℓ	E
0.90	2	1.03	0.55	2	1.14
	3	1.04		3	1.20
0.80	2	1.05	0.50	2	1.15
	3	1.07		3	1.22
0.70	2	1.08	0.40	2	1.20
	3	1.12		3	1.29
				4	1.35
0.60	2	1.12	0.30	2	1.24
	3	1.17		3	1.37
				4	1.45
				5	1.51

is recommended only when the correlation (estimated using three or more sites) is less than 0.5. Using four sites would be advantageous when the correlation is less than 0.40. Resende (2002) presented other approaches for various selection strategies for which the ideal numbers of sites are defined. The interested reader should refer to this reference.

The appropriate number of experimental sites for a fixed total number of individuals depends on the heritability of the trait and the intraclass genetic correlation across sites. Setting $n\ell$ as the total number of individuals per accession, where n refers to the number of individuals per site, and comparing the analysis of the $n\ell$ individuals in one environment or in several environments, the efficiency of selection based on various sites compared to selection based on one single site is given by

$$E = \left[\frac{1 + (n\ell - 1)\hat{h}_i^2}{1 + (n-1)\hat{h}_i^2 + n(\ell - 1)\hat{r}_{gg}\hat{h}_i^2}\right]^{1/2}$$

where $\hat{h}_i^2$ is the estimated individual heritability within the site.

For example, for $h^2 = 0.20$, using 30 individuals per family will provide accuracy on the order of 90–95 % for the selection of individuals for propagation by seeds or clones for vegetative propagation (Resende 2002). For a total fixed number of individuals assessed, the author noted that it is advantageous (a gain of at least 6 %) to use four, three, and two sites for correlations with magnitudes of 0.30, 0.50, and 0.70, respectively. This result demonstrates that the interaction can be minimized without devoting additional resources but simply by dividing a large experiment across several sites.

References

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19(1):51–61

Auer PL, Doerge RW (2010) Statistical design and analysis of RNA sequencing data. Genetics 185:405–416

Aulchenko YS, Koning D, Haley C (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177:577–585

Banzato DA, Kronka SN (1989) Experimentação agrícola. FUNEP, Jaboticabal, 247 pp

Barbin D (2003) Planejamento e análise estatística de experimentos agronômicos. Midas, Arapongas, 208 pp

Cabrera-Bosquet LJ, Crossa J, von Zitzewitz MD, Serret J, Araus L (2012) High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. J Integr Plant Biol 54:312–320

Cobb JN, Declerck G, Greenbrg A, Clark R, McCouch S (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. Theor Appl Genet 126:867–887

Cochran WG, Cox GM (1992) Experimental designs, 2nd edn. Wiley, New York, 611 pp

Cox DR (1958) Planning of experiments. Wiley, New York, 308 pp

Crossa J, Burgueño J, Cornelius PL, McLaren G, Trethowan R et al (2006) Modeling genotype·environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Sci 46:1722–1733

Faraway JJ (2005) Linear models with R. Chapman & Hall/CRC, New York, 229 pp

Federer WT (1956) Augmented (hoonuiaku) designs. Hawaian Planters' Rec 55:191–208 (Aica)

Federer WT, Reynolds M, Crossa J (2001) Combining results from augmented designs over sites. Agron J 93:389–395

Fiorani F, Schurr U (2013) Future scenarios for plant phenotyping. Ann Rev Plant Biol 64:267–291

Fisher RA (1926) The arrangement of field experiments. J Ministry Agric Great Brit 33:503–513

Fisher RA (1935) The design of experiments, 2nd edn. Oliver & Boyd, Edinburgh

Gilmour AR (2000) Post blocking gone too far! Recovery of information and spatial analysis in field experiments. Biometrics 56:944–946

Hinkelmann K, Kempthorne O (1994) Design and analysis of experiments—volume I: introduction to experimental design. Wiley, New York, 495 pp

Hinkelmann K, Kempthorne O (2005) Design and analysis of experiments—volume II: advanced experimental design. Wiley, New York 780 pp

Lado B, Matus I, Rodriguez A, Inostroza L, Poland J, Belzile F, del Pozo A, Quincke M, Castro M, von Zitzewitz J (2013) Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. G3 3:2105–2114

Leite MSO, Peternelli LA, Barbosa MHP (2006) Effects of plot size on the estimation of genetic parameters in sugarcane families. Crop Breed Appl Biotech 6(1):40–46

Leite MSO, Peternelli LA, Barbosa MHP, Cecon PR, Cruz CD (2009) Sample size for full-sib family evaluation in sugarcane. Pesquisa Agropecuária Bras 44:562–1574

Masuka BJ, Araus L, Das B, Sonder K, Cairns JE (2012) Phenotyping for abiotic stress tolerance in maize. J Integr Plant Biol 54:238–249

Papadakis J (1984) Advances in the analysis of field experiments. Communicationes dÁcademie dÁthenes 59:326–342

Patterson HD, Williams ER (1976) A new class of resolvable block designs. Biometrika 63:83–92

Peternelli LA, Souza EFM, Barbosa MHP, Carvalho MP (2009) Delineamentos aumentados no melhoramento de plantas em condições de restrições de recursos. Ciência Rural 39:2425–2430 (UFSM-Impresso)

Peternelli LA, Resende MDV, Mendes TO (2012) Experimentação e análise estatística em cana-de-açúcar. In: Santos F, Borém A, Caldas C (eds) Cana-de-açúcar: bioenergia, açúcar e etanol —Tecnologias e perspectivas, 2nd edn. Editora Folha de Viçosa Ltda., Viçosa, pp 333–353

Poorter H, Fiorani F, Stitt M, Schurr U, Finck A, Gibon Y, Usadel B, Munns R, Atkin OK, Tardieu F, Pons TL (2012) The art of growing plants for experimental purposes: a practical guide for the plant biologist. Funct Plant Biol 39:821–838

Ramalho MAP, Ferreira DF, Oliveira AC (2005) Experimentação em genética e melhoramento de plantas. UFLA, Lavras, 300 pp

Resende MDV (2002) Genética biométrica e estatística no melhoramento de plantas perenes. Embrapa Informação Tecnológica, Brasília, 975 pp

Resende MDV (2007) Matemática e estatística na análise de experimentos e no melhoramento genético. Embrapa Florestas, Colombo, 560 pp

Resende MDV, Barbosa MHP (2005) Melhoramento genético de plantas de propagação assexuada. Embrapa Florestas, Colombo, 130 pp

Scott RA, Milliken GA (1993) A SAS program for analyzing augmented randomized complete-block designs. Crop Sci 33:865–867

Silva MAG, Peternelli LA, Nascimento M, da Silva FL (2013) Modelos mistos na seleção de famílias de cana-de-açúcar aparentadas sob o enfoque clássico e bayesiano. Revista Brasileira de Biometria 31:1–12

Souza EFM, Peternelli LA, Barbosa MHP (2006) Designs and model effects definitions in the initial stage of a plant breeding program. Pesq Agropec Bras 41(3):369–375 (Brasília)

Steel RGD, Torrie JH, Dickey DA (1997) Principles and procedures of statistics: a biometrical approach, 3rd edn. McGraw-Hill Companies, New York, 666 pp

Storck L, Garcia DC, Lopes SJ, Estefanel V (2000) Experimentação vegetal. In: Santa Maria RS (ed) da Universidade Federal de Santa Maria, 199 pp

Williams ER, Matheson AC (1994) Experimental design and analysis for use in tree improvement. CSIRO Information Services, East Melbourne, 174 pp

http://www.springer.com/978-3-319-13676-9

Phenomics
How Next-Generation Phenotyping is Revolutionizing Plant Breeding
Fritsche-Neto, R.; Borém, A. (Eds.)
2015, VIII, 142 p. 50 illus., 39 illus. in color., Hardcover
ISBN: 978-3-319-13676-9