Replication Concepts for Bioenergy Research Experiments

Michael D. Casler · Wilfred Vermerris · Richard A. Dixon

Abstract While there are some large and fundamental differences among disciplines related to the conversion of biomass to bioenergy, all scientific endeavors involve the use of biological feedstocks. As such, nearly every scientific experiment conducted in this area, regardless of the specific discipline, is subject to random variation, some of which is unpredictable and unidentifiable (i.e., pure random variation such as variation among plots in an experiment, individuals within a plot, or laboratory samples within an experimental unit) while some is predictable and identifiable (repeatable variation, such as spatial or temporal patterns within an experimental field, a glasshouse or growth chamber, or among laboratory containers). Identifying the scale and sources of this variation relative to the specific hypotheses of interest is a critical component of designing good experiments that generate meaningful and believable hypothesis tests and inference statements. Many bioenergy feedstock experiments are replicated at an incorrect scale, typically by sampling feedstocks to estimate laboratory error or by completely ignoring the errors associated with growing feedstocks in an agricultural area at a field or farmland (micro- or macro-region) scale. As such, actual random errors inherent in experimental materials are frequently underestimated, with unrealistically low standard errors of statistical parameters (e.g., means), leading to improper inferences. The examples and guidelines set forth in this paper and many of the references cited are intended to form the general policy and guidelines for replication of bioenergy feedstock experiments to be published in *BioEnergy Research*.

Abbreviations

SRWC	Short-rotation woody crops
QTL	Quantitative trait loci
IVDMD	In vitro dry matter digestibility
LSD	Least significant difference
HSD	Honestly significant difference (Tukey's LSD)
HPLC	High-performance liquid chromatography
PCR	Polymerase chain reaction
RT-PCR	Real-time PCR
RNAseq	High-throughput expression profiling (RNA sequencing)

M. D. Casler (✉)
U.S. Dairy Forage Research Center, USDA-ARS, 1925 Linden Dr.,
Madison, WI 53706-1108, USA
e-mail: michael.casler@ars.usda.gov

M. D. Casler
e-mail: mdcasler@wisc.edu

W. Vermerris
Department of Microbiology and Cell Science and UF Genetics
Institute, University of Florida, P.O. Box 103610,
Gainesville, FL 32610-3610, USA

R. A. Dixon
Department of Biological Sciences, University of North Texas, 1511
W. Sycamore, Denton, TX 76203-5017, USA

R. A. Dixon
US Department of Energy, BioEnergy Sciences Center (BESC), Oak
Ridge National Laboratory, Oak Ridge, TN 37831, USA

Introduction

Experimental design is a critical aspect of bioenergy research projects, one that is often overlooked or underestimated in value. Proper experimental design allows researchers to develop all of their desired inferences and can help to ensure that

treatments result in statistically significant results. Conversely, use of improper experimental design, poorly chosen designs, or ad hoc experiments which were not designed for the intended purpose will typically result in lack of significant differences between experimental configurations, treatments or genotypes, inability to develop all desired inferences, and resistance from reviewers and editors of refereed journals. The purpose of this review paper is to carefully present and describe the concept of proper experimental design, specifically as it relates to proper forms and degrees of replication, for a range of bioenergy research subjects that span the conversion pipeline from biomass to energy. Our approach is to present a minimal amount of theoretical framework, just sufficient to describe the fundamental approaches to solving problems in designing proper experiments, and to rely heavily on examples. Upon publication, this article will serve as the de facto editorial policy for replication in articles submitted to *BioEnergy Research*. As such, this paper is not a comprehensive literature review or treatment of this subject, but rather a set of guidelines and principles to guide researchers in making decisions that enhance the value and worth of their research in the eyes of their peers. For that reason, we have divided the arguments and discussions about replication into five categories: agronomic crops including both woody and herbaceous species, glasshouse and growth chamber experiments, benchtop and reactor-scale experiments, plant genetics, and biochemistry/molecular biology. These sections are clearly labeled to enable easy identification of the most relevant topics for a wide range of research projects.

Controllable Versus Non-controllable Sources of Variation

By definition, bioenergy research projects are based on biological materials, always a biological feedstock and often a cocktail of microorganisms or enzymes responsible for complete or partial conversion of biomass to energy. By their very nature, biological experiments contain numerous sources of variability, some of which are desirable and under the researcher's control, while others are undesirable and only partially (or not at all) under the researcher's control. Nearly every researcher who conducts critical comparative experiments wishes to make biological comparisons among the treatments that may impact some important biological measurement variable. Due to the two main sources of biological variability mentioned above, the key to success in making biological comparisons is to have a certain degree of statistical confidence that the biological comparison is meaningful. In other words, the difference between two treatment means must be sufficiently large that it overcomes the random, uncontrollable, or undesirable variability that exists within a biological system, i.e., that the signal-to-noise ratio exceeds some critical minimum level. Typically, we use standard test statistics, e.g., t tests, F tests, and χ^2 tests, to determine if

signal-to-noise ratios are sufficiently large to warrant the formation of inferential statements. We customarily use a significance level of 0.05 or 0.01, but this is not always optimal. Rather, the choice of an appropriate significance level should, optimally, be chosen using some type of risk assessment and risk management approach, which can be applied to any experimental situation [1].

A unique aspect of bioenergy research, particularly as published in this journal, is the diversity of scientific disciplines and methodologies that are represented under a single umbrella. From an editorial standpoint, that level of diversity is somewhat problematic because it includes diversity in the design, conduct, and analysis of experiments, to the point of including differences in terminology and definitions of what constitutes an acceptable vs. an unacceptable experimental design.

What Is an Experimental Unit and Why Is It Important?

Put very simply, an experimental unit is the smallest unit to which a treatment is applied. This definition represents a critical phase in the initial planning of an experiment, because it defines the scale at which the first and foremost level of replication should be planned. The term "treatment" is used herein as a general descriptor of the manipulations or applications that the researcher creates to form the basis for the experiment and the hypothesis tests to follow. Randomization is a critical aspect of the assignment of treatments to experimental units, ensuring that multiple experimental units of each treatment are independent of each other. Systematic designs, in which the treatments are not randomized, are not acceptable, as discussed later.

Every experiment has an experimental unit, whether the researcher recognizes this concept or not. In agronomic experiments, the experimental unit is typically a single plant or a cluster of plants in a "plot." In glasshouse experiments, the experimental unit is usually a single pot but sometimes is defined as a flat or tray containing several pots that are all treated alike. In benchtop research, the contents of a single reactor vessel generally comprise the experimental unit, whether that vessel is a fermentation vessel, a petri dish, or a flask that contains a microorganism or enzyme interacting with a substrate. Smaller experimental units are impossible for benchtop research, unless a vessel can somehow be subdivided into units that can be treated and measured independently of each other. Finally, in genetic research, there is a wide array of experimental units, depending on the goals of the experiment. For example, field studies of different cultivars or biomass crops will have a "plot" or cluster of plants as the experimental unit. Genetic variability studies may reduce the size of the experimental unit to the individual genotype or clone, provided that individual genotypes can be properly

replicated. Genomic studies, focused on the discovery of quantitative trait loci (QTL) can be designed with the allele as the treatment level, such that large population sizes in which many individuals will represent each treatment level, can substitute for traditional replication of the individual genotypes.

Once the experimental unit has been defined, the next step for the researcher is to determine how to replicate treatments that form the core of the experiment. Proper replication requires that, for each treatment, the following conditions must be met: (1) there must be multiple experimental units, repeated in time or space or both, (2) each experimental unit must receive the treatment, be allowed to express itself, and be measured independently from all other experimental units, throughout the entire course of the experiment, and (3) treatments must be randomized, not organized in a systematic or ordered manner. The requirement of independence is critical and absolute and is not open for debate [2]. Failure to replicate treatments at the proper scale, sometimes termed pseudo-replication [3], significantly impacts the inference that is possible from an experiment.

In many experimental situations, the experimental unit is too large to be fully utilized in the data collection phase and requires analysis of samples representing the experimental unit. All biomass quality evaluations are subject to this restriction, often with samples of <1 g of tissue having to represent many tons of biomass harvested from a switchgrass or woody-species plot, for example. Additional examples of this restriction include aliquots of reaction products in solution or samples of microorganisms from algal, fungal, or bacterial cultures. These situations all require a carefully designed sampling scheme to ensure that the final sample to be assayed or analyzed is representative of the entire experimental unit or a carefully defined portion thereof. Three examples of the latter approach are the use of stem segments from the lowest internode to represent switchgrass biomass [4], stem cores to represent maize biomass [5], or stem disks to represent woody biomass [6]. In each case, highly recalcitrant biomass is sampled from carefully defined tissue that is more homogeneous than that found in the whole biomass sample.

Ultimately, most bioenergy experiments end with some form of laboratory analysis. Bulking samples across different experimental units that represent the same treatment, i.e., creating composite field samples, is a common practice to save funds and time spent on laboratory analysis. In general, it is best to analyze samples representing individual experimental units, rather than bulk samples representing multiple experimental units. The argument used to justify bulking is that it "evens out" variation among experimental units, but estimation of this variation is critical for developing appropriate statistical inferences. In addition, bulking samples has an inherent risk of inadvertent sample mix up. Furthermore, bulking results in loss of information on individual experimental units,

including outliers caused by malfunctioning equipment or unexpected local environmental variation. The inclusion of such outliers in a bulk sample can affect the interpretation of the larger data set without the researcher's knowledge.

Scale and Form of Replication

Biological experiments can be replicated at many scales and, indeed, many experiments contain several different levels of replication. Scale is important, because it determines the level of inferences that are possible with respect to the treatments. Replication at the exact scale of the experimental unit, i.e., multiple, independent experimental units per treatment, provides an estimate of experimental error. It is only this estimate of experimental error, computed at the proper scale, which allows the researcher to adequately judge whether or not the treatment-to-treatment variability exceeds the random or uncontrollable variability by a sufficient degree to judge the treatment means to be different.

As a general principle, replication at scales larger than the experimental unit is conducted to allow the researcher to develop greater or broader inferences with regard to a range of experimental conditions under which the treatments are expected to cause different effects. For example, agronomic experiments can be replicated across multiple locations, allowing potentially broader inferences with respect to soil type, climate, and/or other environmental conditions under which treatment effects occur. Similarly, reactor or benchtop experiments can be conducted under a range of physical, chemical, and/or biological conditions or with different feedstocks.

Conversely, replication at scales smaller than the experimental unit occurs when the observational unit is smaller than the experimental unit. In many types of research, the experimental unit is too large for measurements to occur at that scale, e.g., large field plots (greater than 0.1 ha) or large reactor vessels such as pilot plants. In these situations, representative sampling of the experimental unit is a major problem. Frequently, multiple observations are taken on each experimental unit, e.g., random samples of plants or random aliquots of a reaction solution. Typically, all measurements taken at one point in time from an experimental unit can be averaged together to deliver the observation for that particular experimental unit. In this case, statistical analysis of data from such an experiment can be made on the basis of these average values per experimental units as the input data. Replication at scales smaller than the experimental unit is not a valid substitute for replication at the experimental unit scale, because residual variances tend to become smaller, in both expectation and practice, as the scale becomes smaller. Thus, using too small a scale for replication results in biased

hypothesis tests, i.e., p values that are unrealistically low (overoptimistic) and inflation of the true type I error rate [2, 3].

Example 1: failure to use the proper scale of replication leads to identification of treatment mean differences that are not real.

Consider the data in Table 1 for which there were three levels of replication: replicate switchgrass plots in the field (experimental units), replicate samples taken from each plot (sampling units), and laboratory replicates of in vitro dry matter digestibility, IVDMD (observational units). The experimental design was a randomized complete block with three replicates and experimental units of 3.6 m^2. The F ratio for treatments is inflated by using any error term that is based on replication at a scale smaller than the experimental unit, e.g., sampling error or laboratory error. The net result is a least significant difference (LSD) value that is unrealistically low, creating an inflated type I error rate and numerous false conclusions. In other words, failure to use the proper scale of replication would have led to numerous declarations of treatment differences that were not real.

General Guidelines for Agronomic Studies

Agronomic studies will often involve field experiments whereby variation in soil composition, fertility, temperature, and precipitation vary from location to location and from year to year. In order for data related to bioenergy production to be maximally relevant, it is therefore important not only to acquire data from multiple locations and/or multiple years but also to replicate at a given site.

Example 2: confounding factors—impact of manure and harvest time on maize biomass quality and productivity.

To illustrate the importance of replication, consider an experiment meant to quantify the effect of manure treatments and harvest time on quality and productivity of maize biomass. The researcher plans four treatments (two sources of manure in combination with two harvest dates, with harvesting to be accomplished using farm-scale equipment). Due to the logistical difficulties of creating multiple experimental units of these four treatments, each treatment is applied to only one section of a uniform and carefully defined field of maize (treatments T1 through T4 in Fig. 1, where each plant in the figure represents a unit of measurement, e.g., the specific field area to which the treatment is applied, divided into four sections or samples). The experimental unit and treatment are completely confounded with each other, i.e., the treatments are replicated but not at the proper scale to eliminate the confounding problem. How can the researcher eliminate this confounding effect?

Two mechanisms of replication are available: (1) collect multiple samples within each experimental unit or (2) repeat the entire experiment for two or more years. Both alternatives represent replication at a different scale than the treatments or experimental units. Option (1) fails to solve the fundamental problem that treatments and experimental units are completely confounded with each other, resulting in the problems pointed out in Example 1 (Table 1). Option (2) solves this problem by using years as a blocking or replication factor, only provided that the experiment is conducted on new and randomized plots in the second year. While the maize field may appear to be uniform, there is no guarantee that it is sufficiently uniform to implement the design in Fig. 1, which is based on the fundamental assumption that the only difference among the four

Table 1 Mixed models analysis of variance for an experiment with three levels of replication: three blocks with one experimental unit per treatment inside each block, two independent samples per experimental unit, and two independent laboratory determinations per sample

Source of variation	df	Mean square	Variance component associated with error term (random effect)	F ratio for treatments	p value for treatments	LSD (0.05) for comparing treatment means
Blocks	2	2485				
Treatments	4	781				
Experimental Error	8	147	20	5.31[a]	0.0219[a]	11[a]
Samples within Experimental Units	15	69	23	11.32[b]	0.0002[b]	7[b]
Determinations within Samples	30	23	23	33.96[c]	<0.0001[c]	4[s]

Data were in vitro dry matter digestibility (mg/g) of switchgrass biomass (Casler, 2000, unpublished data). The total number of laboratory determinations was 12 per treatment for each of three hypothesis testing scenarios (a, b, c)

[a] Correct (unbiased) values of F ratio and LSD calculated using experimental error mean square in this experiment with three experimental units per treatment, two samples per experimental unit, and two determinations per sample. The treatment means were 496, 502, 505, 514, and 515 mg/g

[b] Values of F ratio and LSD calculated using the mean square representing variation among samples within experimental units in this experiment

[c] Values of F ratio and LSD calculated using the mean square representing variation among determinations within samples in this experiment
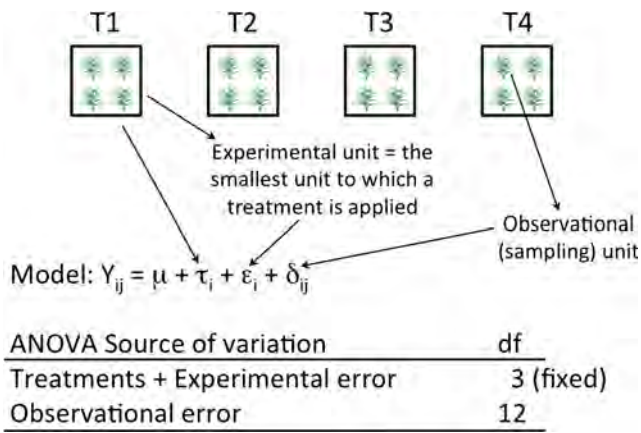
Fig. 1 Design example 1: Sixteen plants are assigned to four experimental units, each of which is assigned to one treatment level (T1 through T4, symbolized by the treatment effect, τ_i in the linear model). Treatments are confounded with experimental units because replication is conducted at a scale that does not match the treatment application

experimental units is due to the application of treatments. While this assumption may be correct, there is no way to know for certain. By definition, that is the nature of confounding—two or more items are inextricably intertwined, such that no amount of statistical analysis, wishful thinking, or discussion can separate them. The problem of confounding treatments with experimental units is illustrated in the anlysis of variance (ANOVA) source of variation that reflects the fact that treatment effects cannot be estimated as a pure and unconfounded factor (Fig. 1). Imagine the potential impact on the data collected if there is an underground gradient in fertility, soil type, moisture holding capacity, pH, or some other important factor that cannot be observed or was not measured and that the direction of this gradient is horizontal with respect to the fieldmap in Fig. 1.

Because statistical analysis software packages do not recognize the limitations of the experimental design, anyone can

devise hypothesis tests for this scenario and apply them to the data. Use of multiple samples within each experimental unit, represented by the observational (or sampling) units in Fig. 1, allows estimation of an error term and subsequent F tests or t tests, but this error term is biased downwards due to scale. Observational units occur over a smaller and more uniform area than experimental units, and they are correlated with each other, thus resulting in a smaller variance than would be expected for multiple independent experimental units of each treatment. Such a test results in overoptimistic p values and an unknown degree of inflation of the type I error rate.

Solution 1: an experimental design with replicated treatments.

So, what options are available to solve this problem? We begin with the simplest and least problematic, from most researchers' point of view. Replication at the proper scale, resulting in unbiased estimates of treatment means and error variances, requires multiple and independent experimental units for each treatment. Figure 2 illustrates the simplest solution to this problem, using four treatments each replicated four times. The number of observations made is identical to the design in Fig. 1, but each of the four observations made on a single treatment were mutually independent, assuring that both problems, confounding and the improper scale of replication (Fig. 1), have been solved. In this particular case, only one sample or data point is collected within each experimental unit, confounding the experimental and observational units, but this usually is not a problem. If researchers are concerned about this problem, for example in very large experimental units that are subject to high levels of variability, then the design illustrated in Fig. 3 offers a solution. In this case, the design remains the same, with multiple independent experimental units per treatment, but the multiple samples or

Fig. 2 Design example 2: 16 plants are assigned to 16 experimental units, 4 of which are independently assigned to each treatment level (T1 through T4, symbolized by the treatment effect, τ_i in the linear model). This is an example of the completely randomized design
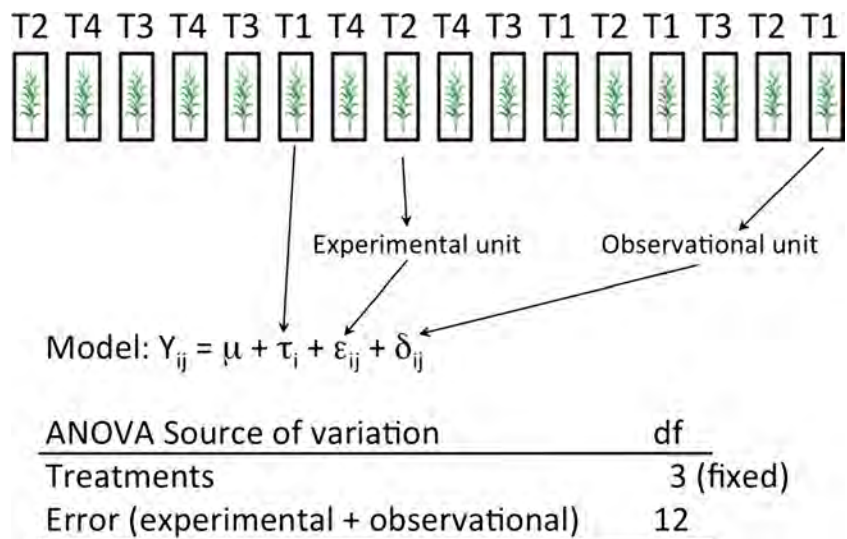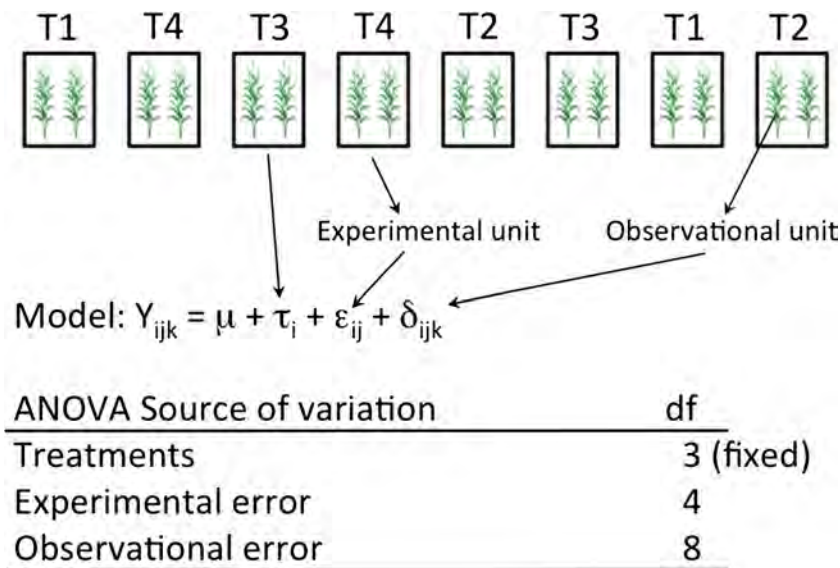
Fig. 3 Design example 3: Sixteen plants are assigned to eight experimental units, two of which are independently assigned to each treatment level (T1 through T4, symbolized by the treatment effect, τ_i in the linear model). Each experimental unit contains two observational, or sampling, units. This is another example of the completely randomized design. In this design, the researcher is using two independent samples from each experimental unit to provide a better estimate of the performance of each experimental unit, i.e., representative sampling, as compared with design example 2

observations within each experimental unit help the researcher to control variability that may be present within the plots. The ANOVA in Fig. 3 shows estimates of two random effects, representing variability at two scales (the observational unit and the experimental unit scale), but it is equally valid to compute the averages of all samples within each experimental unit, then to conduct the ANOVA on the experimental-unit average values. Experimental units in most agronomic research are established from seeds planted in rows, broadcast seeding, or from transplanted plants. Observations made on units smaller than the entire experimental unit can be on individual plants, on groups of plants within a quadrat or sampling frame, or from individual bales of biomass. Figures 2 and 3 both represent variations of the completely randomized design, the simplest of all experimental designs.

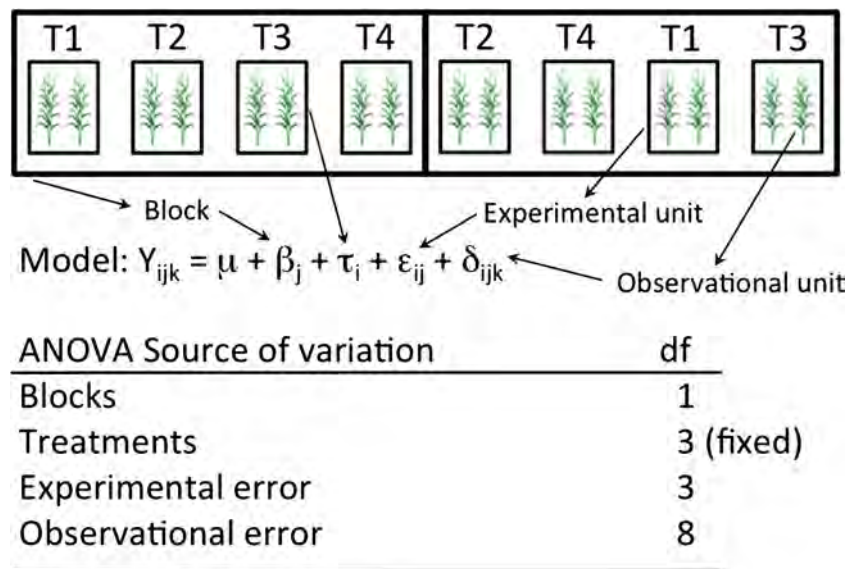Solution 2: replication over multiple years or locations.

Another option would be to repeat the experiment over multiple years. The easy way to accomplish this does not actually solve the problem of confounding within an individual year or growing season. If the treatments shown in Fig. 1 are repeated on the same experimental units for multiple years, as would be the case with repeated observations of a perennial plant, any environmental or unknown factors that are confounded with treatments remain as such for the duration of the experiment. In some cases, it may become worse with time, if its effects are cumulative. Even though multiple observations are made on each treatment, those observations are not independent of each other, e.g., they are correlated with

each other. Because they are correlated with each other, they do not provide an unbiased estimate of experimental error, as with the design in Fig. 1. However, multiple years could be used to solve this problem, as shown in Fig. 4. In this case, the block on the left, which contains four experimental units and four treatments is applied in 1 year, while the block on the right is applied to a new field of maize, independent of the previous years' field, in the second year. This represents an application of the randomized complete block design, in which each year acts as a "block" of experimental units. Of course, such a design would not require the use of multiple years, provided that sufficient land and labor are available to apply multiple experimental units of each treatment in a single year. The choice of completely randomized design vs. randomized complete block (to block or not to block) can be complex and is beyond the scope of this paper, having been effectively discussed in numerous other sources [2, 7, 8].

Large-Scale Agronomic or Engineering Experiments

As pointed out in example 2, certain types of treatment factors are routinely difficult to replicate at the proper scale. Examples include irrigation treatments, slurry or manure applications, planting dates, and soil types. Replication of treatments, such as irrigation and planting dates, can be achieved, although sometimes at great cost. Replication of "natural" treatments, such as soil type, is virtually impossible without some form of artificial environment, such as an aboveground or belowground soil-filled structure. For many of these factors, replication at the scale of the experimental unit is

Fig. 4 Design example 4: Sixteen plants are assigned to eight experimental units. Experimental units are grouped into two blocks, each of which contains a number of experimental units exactly equal to the number of treatment levels (T1 through T4, symbolized by the treatment effect, τ_i in the linear model). Each experimental unit contains two observational, or sampling, units. This is an example of the randomized complete block design, with sampling

Model: $Y_{ijk} = \mu + \beta_j + \tau_i + \varepsilon_{ij} + \delta_{ijk}$

ANOVA Source of variation	df
Blocks	1
Treatments	3 (fixed)
Experimental error	3
Observational error	8

impossible or extremely difficult due to logistical or equipment expenses, the need for larger plots than are available, or extreme labor requirements. Agronomists typically utilize split-plot designs to assist in solving this problem [7], but this design does not eliminate the need to replicate treatments at the whole-plot experimental unit scale. These obstacles should not be used as excuses to proceed with poorly designed experiments that result in limited or impossible-to-interpret inferences but should be used to drive creativity toward solving problems using imaginative and innovative, yet statistically valid, approaches. There is a rich body of literature that derives from biological scientists interacting with statisticians or biometricians to solve experimental design and measurement problems, some of which is cited in this paper.

Harvesting and storage research represent two specific research topics that are specifically problematic regarding the conflict between statistical inference and logistics of accomplishing the research. Seldom are these studies replicated at the proper scale to generate valid statistical inferences between field-based treatments. Harvesting studies typically involve the use of different harvest dates to create biomass crops with different moisture contents or different harvesting or swathing equipment that may impact either dry matter losses or crop quality. Typically, an experiment with three treatments will have only three field plots, patterned after Fig. 1. However, there is a fairly easy solution to this problem in which engineers could use the design in Fig. 2 by randomizing single passes of each treatment through the field, allowing for very precise estimation of confidence intervals and inferential statements at the field level. Likewise, biomass storage studies generally fail to replicate at the proper level. These studies typically involve different types of bales and storage treatments, many of which can be easily replicated and randomized, e.g., different sizes of bales, different shapes

of bales, and different bale coverings. Other types of storage treatments are difficult or impossible to replicate, e.g., inside or outside a barn, ground surface composition (concrete, grass, or gravel), or stacking methods.

Specific Considerations for Experiments with Perennials

Many of the obstacles and impediments to creating valid inferences for perennial crop research, including both herbaceous species and short-rotation woody crops (SRWC) are similar to those discussed above. Nevertheless, there are some obvious differences between annuals and perennials, created largely by differences in scale and longevity of experimental plots, with perennials often demanding research on a larger physical scale and conducted over a longer period of time than for herbaceous crops. Many perennials are genetically heterogeneous and may require considerable time to become established. In addition, individual plant mortality creates open spaces that can be handled in many different manners or simply ignored and considered part of the "treatment." Sometimes, experimental material, such as seed or rhizomes, are in short supply, prompting the need for unreplicated or partially replicated designs, such as discussed later in example 5. In addition, stand age is often an important factor to be studied, to gain an understanding of changes in biomass production over time. Creating an experimental design in which the fixed effect of stand age is not confounded with random weather effects is particularly difficult [9].

Due to these requirements, scientists working on perennials cannot respond as quickly and readily as those working on annuals to new initiatives or by creating new experiments every year as needs arise. As such, many critical comparative experiments designed to evaluate bioenergy production and sustainability characteristics of perennials are conducted on

existing stands that were created strictly for production purposes, creating potential limitations to research inferences. Pseudo-replication [3], replication at an improper scale, which is usually smaller than the actual experimental unit, can be a common problem in research on existing stands of perennials.

Example 3: Pseudo-replication in Production-scale Research

As an example, consider a new experiment to be conducted at an existing plantation of SRWC or in an area that includes numerous fields of a perennial energy grass. An existing research area might typically include multiple planting dates and multiple genotypes, clones, or cultivars planted over several years, with a small amount of new land converted to the biomass crop in each year. Considerable care must be taken when using this type of production land for bioenergy research. Figure 1 represents the basic scheme of the potential experimental area, in which experimental units are represented by the land used for energy-crop establishment in any given year and the treatments become different establishment years and genotypes. Simply using this experimental area to draw statistical and biological inferences about stand age or genotype is completely untenable, as described above in Example 2. Such an experimental area would have stand age, genotype, and perhaps a host of other environmental factors all confounded with each other to unknown degrees. The researchers could convince themselves that stand age and genotype are the dominant sources of variation, but many reviewers will justifiably not be convinced that this assumption is valid. It is far better for the research team to face this issue "head-on" at the design phase than after receiving unfavorable peer reviews from an international scientific journal.

Ideally, a new experiment would be designed, with replicated and randomized treatments, solely within each individual planting that is homogeneous for establishment date and genotype. As with the agronomic experiment designed above, proper randomization and replication of treatments to multiple independent experimental units within a homogeneous experimental site allow for clean and valid statistical and biological comparisons and hypothesis tests. Repetition of the experiment across different establishment dates and/or genotypes would allow the researchers to broaden the inferences to be derived from such an experiment. In some cases, this may require some compromise in the final size of the experimental unit, resulting in "plots" that are smaller than desired. For a discussion of this concept and approaches to evaluate its potential impact on experimental results, see the following sources [7, 10]. In most biological situations where a researcher is evaluating the costs and benefits of increasing the size of the experimental unit vs. the number of replicates, and there is no specific data to help guide an informed choice, the latter is generally better for both statistical precision and inference space [11, 12].

Two highly undesirable approaches to this problem would be (1) partially or completely confounding the new treatments with establishment dates and genotypes or (2) ignoring the multiple establishment dates and genotypes in the layout of the experimental treatments. The first of these, in the worst case, would simply take the existing problem and make it worse, by adding another confounded factor to the mix, e.g., a new genotype or a different planting density in the next establishment year. The second of these is problematic in a different way, because it introduces variability that may not be controllable to the new critical comparative experiments. There are sophisticated spatial analysis methods readily available to biological researchers [10, 13], but the proper use of experimental design, including proper randomization and replication, is always the preferred first step [7, 14]. Spatial analysis, no matter how sophisticated, should never be viewed as an alternative or excuse to ignore principles of proper randomization, replication, and blocking.

Alternatively, a highly desirable solution to this problem could be created by the introduction of a new research objective. For example, consider Fig. 1 in which the four "treatments" represent four different genotypes established in 1 year. Assuming that each field plot is sufficiently large, new critical comparative experiments, e.g., fertilization treatments or harvest timings, could be designed with proper replication and randomization within each of the field plots in Fig. 1. Such a study would be extremely powerful, allowing the development of inferences with regard to the critical comparative experiment within and across the various genotypes. The "genotype" effect is the only effect for which a valid statistical inference cannot be developed, due to the confounding and lack of replication.

Finally, one more alternative solution would involve the use of control-plot designs, which can be designed and tailored to fit almost any researcher's needs [7]. Control plots are frequently used in on-farm or farm-scale agronomic research as a mechanism to adjust observations on unreplicated treatments for spatial variation that typically exists within extremely large experimental areas. One highly efficient design is to assign one of the treatments as the control and to pair the control treatment with each other treatment in a paired-plot design.

Guidelines for Glasshouse and Growth Chamber Research

Glasshouse research is generally conducted at one or both of two scales: macro- or micro-environmental. Research at the micro-environmental scale is simple and straightforward, involving one set of environmental conditions and likely presenting little problem for most researchers. The required

environmental conditions for the glasshouse should be predetermined by the research team and arranged to be as uniform as possible throughout the physical space to be utilized. The experimental unit should be carefully defined, usually a single pot, containing one or more plants, or sometimes a tray or flat of cells. Once this definition is made, each treatment should be assigned to multiple and independent experimental units (replicates) and all experimental units should be randomized in an experimental design or pattern that is capable of dealing with spatial variation that may exist within the glasshouse environment. Except in very unusual cases, multiple plants within a pot cannot serve as independent replicates of the treatment, because they are not independent of each other and their variance will be an underestimate of the true variance among experimental units treated alike.

The completely randomized design, which involves a simple randomization of all experimental units in a single block, is a good idea only under one of two circumstances: (a) there is little likelihood of spatial variation, or microclimate effects, within the glasshouse or (b) the researchers plan to re-randomize or rearrange the experimental units periodically during the course of the experimental time period [15–17]. We strongly discourage option (b) for four reasons: (1) rearrangement of experimental units increases the workload and creates more opportunities for mistakes in randomization or damage to plants [17]; (2) successful application of this approach requires that experimental units spend similar amounts of time within each microclimate [15]; (3) there is no evidence that this approach will lead to greater precision than the use of formally randomized designs, because the rearrangement either eliminates or creates significant challenges in accounting for spatial or microclimate variation; and (4) there are numerous alternative experimental design options ranging in both complexity and expected effectiveness. Microclimate effects are common in glasshouse environments and these effects are generally predictable over extended periods of time [18, 19]. As such, systematic blocking designs will generally be more effective in accounting for microclimate variation than rearrangement approaches [15].

Blocking designs generally provide the optimal solution for glasshouses or growth chambers that possess some microclimate effects [7]. Blocking designs can be grouped into two categories: one-way blocking (randomized complete block, lattice, and balanced incomplete block designs) or two-way blocking (Latin squares, Youden squares, Trojan squares, lattice squares, and row-column designs). Not all blocking designs and patterns are effective [20]; as such, experiments with many small blocks, such as incomplete blocking designs should be used in combination with spatial statistical methods to identify patterns of microclimate effects that can be used as a basis for designing future experiments [18, 19]. This is a highly empirical process that may involve several iterations or repetitions to identify consistent patterns [18–20].

Edmondson [18] suggested that using uniformity trials, in which all pots represent a single uniform treatment, is "folly" and we agree that such an experiment is unnecessary given the ability of blocking designs and spatial statistics to parse out sources of variability on a fine scale, especially when the design involves many small blocks [2, 7, 14].

Macro-environmental glasshouse or growth chamber experiments create experimental design and statistical analysis problems, because replication is seldom conducted at the proper scale. Multiple glasshouses or growth chambers are often used to create highly controlled and differential environmental conditions, for which hypothesis testing and/or quantitative differences are of interest. Most of these experiments are designed as shown in Fig. 1, with the glasshouse or chamber as the experimental unit and replication conducted only within the experimental unit, at some type of sampling unit level. These conditions may form the entire experiment or serve as one factor in a multifactor experiment. In the case where they form a single factor experiment, the statistically viable options are (1) to double the number of chambers, allowing true replication of the macro-environmental treatments, or (2) to repeat the experiment over time, re-randomizing the environmental conditions (treatments) to the various glasshouses or growth chambers to ensure that the error term is estimated at the proper scale, i.e., that each treatment is applied to multiple and independent experimental units (glasshouses or chambers). Use of multiple plants, pots, trays, or flats within each house or chamber does not allow the researcher to estimate the proper error term, underestimating the true variation among experimental units treated alike (Fig. 1; Table 1).

> Example 4: a glasshouse study on the effect of macro-environmental variation.

Recently, *BioEnergy Research* published a paper that used a novel and very economical approach to solve this problem. Arnoult et al. [21] designed a two-factor experiment with four macro-environmental treatments, defined by all combinations of two daylengths and two temperature regimes, each assigned to one glasshouse. This is almost identical to the design shown in Fig. 1, differing only in the number of treatments and replicates within each macro-environment. Within each glasshouse, the second factor consisted of eight different genotypes of Miscanthus. Genotypes were established in a completely randomized design with ten replicates within each glasshouse (80 plants and pots within each glasshouse). Without replication of the environmental treatments over time, comparisons among genotypes provide the only valid statistical hypothesis tests. The authors were partly interested in making comparisons among genotypes for mean performance across environments, but their principal goal was to determine if genotypes differed in their response to the environmental factors (light

and photoperiod). The authors designed three independent-sample t tests to evaluate the environmental sensitivity of each genotype: (1) long vs. short days, (2) warm vs. cool temperature regime, and (3) the daylength×temperature interaction. Each of these represents an interaction contrast within the larger genotype×environment interaction. Each t test was applied to the raw data of each genotype, creating a total of 24 hypothesis tests that focused on the simple effects of each environmental response within each genotype. Comparison of these effects across the eight genotypes provided an assessment of their relative environmental responses.

Guidelines for Benchtop and Reactor-Scale Research

Reactor-scale research consists of biomass conversion studies in batch or continuous-flow reaction vessels. The scale can vary from small and relatively inexpensive reactors present in multiples, to large and expensive pilot- or commercial-scale reactors, available as a single unit. Benchtop research may take on a wider range of disciplinary activities, e.g., germination tests, growth of fungal or bacterial pathogens, insect preference or survivorship studies, etc. The principles discussed below apply equally to any type of critical comparative experiment conducted on biological material that is subject to random and unexplained sources of variation.

It is common for descriptions of benchtop or reactor-scale research to contain sparse or no reference to experimental design, generally referencing the use of laboratory duplicates or triplicates as the only reference to replication. Our recommendation to researchers in these disciplines is to include a thorough description of the experimental design, preferably using standard terminology classically used in statistical textbooks (e.g., [2, 8]). Ideally, the treatments should be randomized and replicated so that any variation that is introduced into the experiment, associated with changes in reaction conditions over time or with an inability to stop every reaction at exactly the same time, can be controlled or removed from the fixed effects of treatments. For example, if there is some disturbance or variation that is associated with timing of stoppage or sampling of reaction products, the worst-case scenario would be to have triplicate samples of each treatment that are grouped together, sampled consecutively, and analyzed consecutively (e.g., Fig. 1). In this case, the undesirable time-scale variation is completely confounded with desirable treatment variation. If such a situation has the potential to occur, Fig. 4 represents the best-case scenario, allowing estimation and removal of a "block" or "time" effect in the ANOVA. Such a design can be applied even in the case of a single large-scale reactor, simply by applying the treatments in Fig. 4 in a randomized sequence over time. In this case, the authors should describe their laboratory duplicates as having been arranged in a randomized complete block in time. Figures 2 and 3 represent

intermediate-level solutions to this potential problem, using the completely randomized design to arrange duplicates or triplicates of the treatments in a manner that reduces the potential for confounding and bias to treatment effects. Regardless of the specific design, randomization is an essential insurance policy against undesirable consequences of a disturbance or unexpected source of variation [22].

Another common omission in the reactor-scale experiments is a description of the feedstock and its preparation. This stems from the fact that these kinds of experiments typically rely on very large, industrial-scale samples, and it is often assumed that the feedstock composition for a given source (e.g., wheat straw, corn stover, pine wood chips, etc.) can be treated as a constant. As many readers of *BioEnergy Research* will appreciate, this is rarely an accurate assumption. As a consequence, lack of details on the origin of the feedstock makes independent replication very challenging, if not impossible. We therefore encourage authors to provide as much detail on their samples in the experimental section of the manuscript, including the location where the feedstock was collected, the age of the material (in units of time or physiological stages), any genotypic information (especially when commercially produced feedstocks were used), and the manner in which the samples were handled after collection (drying, grinding, storage conditions). Lastly, especially when a single feedstock sample is used, or when comparisons are made between two species, each represented by a single source, the authors need to be careful to not overstate the conclusions and leave room for the possibility that observed differences may not occur in the same way when different samples of these feedstocks are used. For example, if two feedstocks were used in an experiment, care must be taken in the conclusions not to attribute feedstock differences to only one of many potentially confounding factors, i.e. the numerous factors that may or may not have differed between the two feedstocks and their environments, such as species, harvest date and stage, soil type, climate, etc.

Lack of replication is also a common challenge in reactor-scale research involving anaerobic digestion. These studies typically rely on a sample collected from a sewage treatment facility. Experiments are conducted with this sample to evaluate, for example, methane production following anaerobic digestion of a number of different feedstocks, or a common feedstock that has been prepared following different pretreatments. It is common to report on methane yields over a period of days or weeks, and then compare the relative merits of certain feedstocks or treatments. When conclusions are based on a single anaerobic digestion experiment, the conclusions only apply to the conditions involving this particular sample of sludge. Given that sewage sludge is a microcosm of bacteria, it is unlikely to be static. The microbial community on the day the sample was collected may be different 2 days or 2 weeks later, when the sewage composition changes, for

example, due to seasonal fluctuations. In addition, conducting laboratory experiments based on a small sample of sludge has the risk of random changes in the microbial community, analogous to the concept of genetic drift. As a consequence, it would be preferable, at the very minimum, to replicate experiments involving sewage sludge by including parallel incubations with subsamples of the sludge. Even better, though, would be to collect sludge samples on different days or from different sewage treatment plants in order to extend the conclusions beyond one specific sample of sludge. Samples of sewage sludge collected on different dates or from different sewage treatment plants should not be bulked or homogenized with each other, but should be applied to feedstocks in independent reactor vessels as a mechanism to assess feedstock conversion under potentially different reaction conditions. Considerable care should be taken to match the replication and experimental unit designation to the question to be answered or the principal hypothesis to be tested. For example, if the "treatments" include multiple genotypes of a feedstock, individual field replicates of the genotypes should not be bulked, but processed independently as an additional "treatment" factor for the fermentations. Additional critical information for fermentation studies includes the duration of the "starvation" period and the standard substrate (with known methane yield) to be used in the fermentation experiment [23].

Guidelines for Genetics Research

Genetic research is highly diverse, in scope and purpose, as well as in the statistical methodology that is used to develop genetic inferences. Many genetic field studies are based on agronomic crops or SRWC, so the fundamental principles that apply to those experiments were discussed in those two sections of this paper. Perhaps the most fundamental distinctions are: (1) genetic experiments tend to be large with many treatments and (2) there are two scenarios in which a case can be made not to replicate genetic experiments or to replicate them in a less traditional manner. With regard to experiment size, genetic experiments generally demand more sophisticated experimental designs and spatial analysis methods to develop precise estimates of genotype or family means. Incomplete block designs tend to be the rule, rather than the exception [7], often combined with spatial analysis methods [24]. This topic is discussed in great detail in numerous review articles and books, e.g., [7, 24].

Traditional field experiments designed to evaluate large numbers of clones, genotypes, lines, or families are increasingly moving toward modified replication structures that give the genetic research more flexibility, broader inference, and greater precision for phenotypic data. Most genetic studies, particularly those that involve attempts to develop genotype-phenotype (marker-trait) associations generally have three common characteristics: (1) hundreds or thousands of genotypes or families, (2) a need or desire to evaluate genotypes in multiple environments, and (3) extremely limited amounts of land, seed, or vegetative propagules for replicating genotypes. As such, many breeders and geneticists are utilizing resource-allocation theory [7] to minimize or eliminate traditional replication in favor of evaluating all genotypes or families at multiple locations [25]. For example, if there is sufficient seed or vegetative tissue to replicate 500 genotypes four times each, resource-allocation theory and empirical results, ignoring costs, nearly always point to the use of four locations with one replicate at each location as the best choice [7].

For studies that are purely based on phenotype, such as selection nurseries or candidate-cultivar evaluations, lack of traditional replication for individual genotypes can be difficult for some peer reviewers to accept. However, one can make a case for this scenario in genotype-phenotype association studies that are designed to detect quantitative trait loci (QTL) for biomass traits. In these studies, the individual genotype or family is no longer the unit of evaluation, i.e., we generally do not care which genotype or family is best or worst. Rather, the unit of evaluation is the allele, i.e., for each locus, we need to know which allele is best and we need to estimate that allele's effect on phenotype. Thus, a simple randomization of 500 genotypes within a fairly uniform field area will result in a valid randomization of the two alleles at any locus. Even if there is some error involved in measuring the phenotype of each plant, the number of plants that possess each allele becomes the actual form of replication for the alleles, i.e., the number of genotypes or families (replication of alleles) is now the critical value with respect to the power of the experiment.

> Example 5: a control-plot design for a large number of unreplicated genotypes.

In either of the cases above, we strongly recommend that any experiment designed to evaluate a large number of genotypes and families without benefit of traditional replication at the experimental unit level include a mechanism to control and adjust for spatial variation, which is inevitable and inexorable in most large field studies. Figure 5 provides an example of a grid design of 421 genotypes, only one of which is replicated, i.e., the "control" genotype. This design could be applied to any type of genotype or family structure, provided that one genotype, line, or family can be chosen for a massive amount of replication, in this case 80 plots at each location. The design shown in Fig. 5 would be replicated at the desired number of locations and the data analysis would involve the use of spatial statistical methods to adjust each plant's phenotypic observation for spatial variation in rows, columns, and squares [24]. This design belongs to the family of partially replicated or augmented designs, which are highly flexible with regard to size, dimension, and ratio of test plots to control
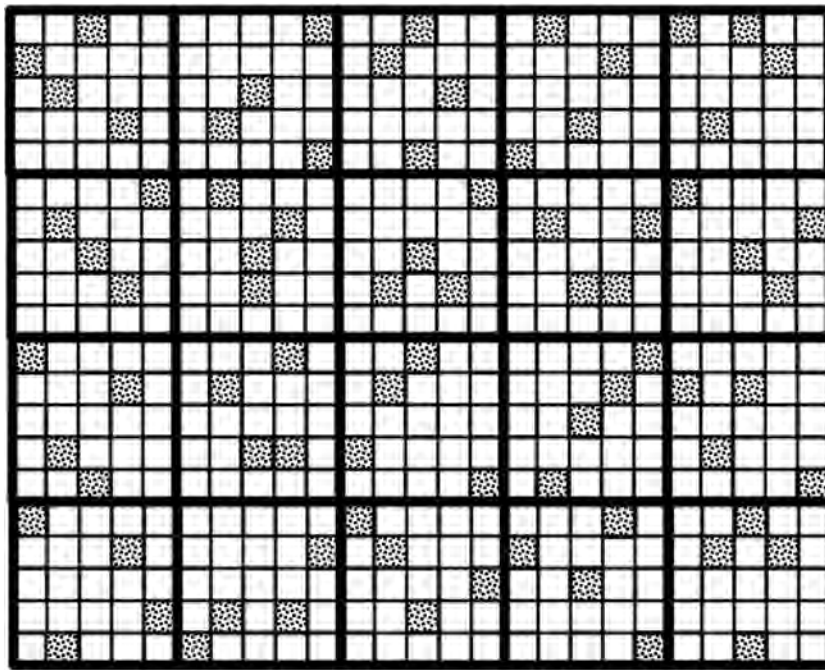
Fig. 5 Design example 5: 420 genotypes, lines, or families are evaluated in a grid design with one replicate or experimental unit per genotype at each location. In this example, 21 genotypes are randomly assigned to each of 20 groups or grids (*open squares*) and each grid contains four independent replicate plots of a single check genotype (*filled squares*). The entire design is repeated at the desired number of locations, with re-randomization at each location. Statistical data analysis would involve the use of spatial statistical methods to adjust each plant's phenotypic observation for spatial variation in *rows*, *columns*, and *squares*. This type of design is highly flexible with regard to size, dimension, and ratio of test plots to control plots

plots [7, 26, 27]. Such a design provides two levels of error control for estimating allele effects: (1) spatial adjustment to enhance precision and accuracy of phenotypic data and (2) large population size to provide adequate precision for estimation of allele effects.

Guidelines for Biochemical and Molecular Biological Experiments

Biochemical analyses in the test tube generally conform to the benchtop or reactor type of study described above. A good example is the determination of kinetic constants for two recombinant enzymes, one being the wild-type enzyme and the other a mutated form (e.g., a cellulase designed for enhanced activity). As described above, the worst-case scenario would be to have triplicate samples of each treatment that are grouped together, sampled consecutively, and analyzed consecutively. If the enzyme assay was performed, for example, using high-performance liquid chromatography (HPLC) to quantify product formation, each sample should be analyzed more than once on the HPLC to determine the analytical variation. In most cases, this is smaller than biological variation (but see below). There is also a need in such an experiment for additional replication at the level of the enzyme preparations themselves, to ensure that apparent differences in kinetics between the wild-type and mutant enzyme are not the result of different efficiencies of enzyme extraction, different times from enzyme extraction to assay, or any other differences in the physical environment in which the samples were handled. The method used for determination of enzyme activity can also influence the experimental design. For example, analysis by HPLC often involves extraction from the aqueous phase in which the enzyme assay was performed to an organic phase that is injected onto the column. The efficiency of this extraction process (determined by the partition coefficient of the reaction product between the aqueous and organic phases) can impact the size of the analytical variation, with low efficiencies leading to higher variations. Such efficiencies should be measured, and if possible controlled for, for example by inclusion of an internal standard that acts as a proxy for the product of interest (acceptable but not optimal), or, if the product of the enzyme reaction is radiolabeled or heavy-isotope labeled, by spiking the sample with unlabeled product, or by spiking with labeled product if the product is unlabeled, prior to separation and determination of product levels. Similar principles apply to the quantification of protein levels through mass spectrometric approaches [28].

The determination of gene transcript levels presents a number of issues as regards to both replication and interpretation. Three methods are now generally used: RT-PCR for low- to medium-throughput analyses, and DNA microarray or

RNAseq for high-throughput whole-genome level analyses. An important issue with PCR approaches and one that often causes problems for authors is whether the PCR method is truly quantitative. For example, simple staining of DNA gels with ethidium bromide does not give a linear response and suffers from problems with saturation, such that this method should never be used for making quantitative inferences. Staining DNA with SYBR green/SYBR gold gives linear responses over a greater concentration range. True quantitative PCR approaches, which exploit the proportional relationship between the quantity of the initial template and the quantity of the PCR products in the exponential phase of amplification (i.e., in real time), present the same problems of replication as found for enzyme assays, namely the need for sufficient analytical and technical replicates to ensure statistical significance. Additional statistical elements inherent in the method itself also require consideration for data analysis and quality control of real-time PCR experiments [29].

Microarray analysis presents a different type of challenge. In the older two-channel printed microarray design, the printing process itself introduces errors and large variance; in some respects, a spotted microarray resembles the previously described field in which differences in response (here DNA binding) might be confounded by differences in the microenvironment in different parts of the field (here the chip). This is less of a factor in the Affymetrix chip design. Statistical analysis of microarrays is a whole discipline in itself, in which normalization of gene expression and controlling for false positives are the major factors of concern. Authors reporting any type of microarray experiment should, at a minimum, adhere to the conventions described in [30]. The types of considerations necessary to overcome errors inherent in the spotted microarrays have been discussed [31], and an example of the types of corrections needed in analysis of Affymetrix arrays can be found in [32].

Next-generation sequencing (RNAseq) is now becoming the method of choice for transcript profiling, particularly for those species with sequenced or partially sequenced genomes (e.g., *Populus* or *Panicum virgatum*). Because of the still significant cost of sequencing, and the enormous amount of data generated, observational studies with no biological replication are common, but are clearly prone to misinterpretation. The reader is referred to [33] for a discussion of the best experimental designs for meaningful comparisons of RNAseq datasets; the principles are fundamentally similar to those described above for field and greenhouse plot design.

Determining the Appropriate or Necessary Number of Replicates

The power of the statistical test is the probability of rejecting a null hypothesis that is, in fact, false. In other words, power reflects the capability of the researcher to detect true differences among treatment means—more powerful experiments result in a higher likelihood that true differences will be detected. Increases in power can be achieved principally by increasing the number of replicates and, secondarily, by changing the replication structure. Both of these topics have been treated in great detail elsewhere (e.g., [7, 24]) and will be presented briefly here.

Example 6: determine the number of replicates required to achieve desired power

Consider any type of bioenergy experiment with two levels of replication: true replication of treatments at the proper scale and sampling of experimental units that are too large to measure in toto. Such a design applies to a wide array of experiments, e.g., bales of biomass as the experimental unit and cores within bales as the sampling unit, agronomic plots as the experimental unit and plants within plots as the sampling unit, or reactors as the experimental unit and aliquots of reaction products as the sampling unit. Figure 6 shows the estimated power of hypothesis tests for 50 potential replication scenarios with a range of experimental units from $r=3$ to 12 replicates and sampling units from $s=3$ to 20 samples per experimental unit. These computations require input data in the form of estimates of error variances obtained from prior experiments and a desired detection limit, i.e., the difference
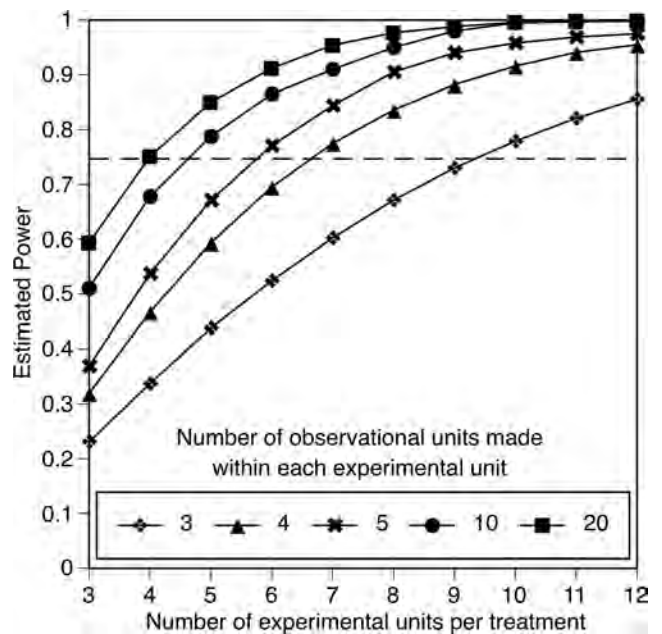

Fig. 6 Estimated power of a hypothesis test designed to detect a treatment difference of 5 % of the mean with a type I error rate of 0.05, experimental and sampling errors, respectively, of 5 and 10 % of the mean, and varying numbers of experimental units ($r=3$ to 12) and observational units ($s=3$ to 20). The *dashed line* represents power=0.75 and illustrates that different replication and sampling scenarios can be created to achieve the same result

between treatment means that the researcher desires to detect. From this particular scenario (variances of 5 and 10 % of the mean and a detection limit of 5 % of the mean), it is clear that there are several options that could achieve the desired result: $r=4$ and $s=20$, $r=5$ and $s=10$, $r=6$ and $s=5$, $r=7$ and $s=4$, or $r=10$ and $s=3$. Typically, these activities lead most researchers to the conclusion that most experiments are underpowered with regard to desired detection levels. Nevertheless, they provide a critical mechanism to assist researchers in designing better, more powerful, experiments, utilizing nothing more than information already obtained from prior experiments. Additional examples can be found in other publications, e.g., [7, 24].

Data Analysis and Presentation of Results

Significant thought should be given to the most logical and sensible data presentation for readers to interpret, understand, and cite the work. Presentation of means and standard errors alone is not acceptable for any data in which the discussion infers treatment differences. Readers cannot ascertain that these differences are statistically significant or repeatable based on means and standard deviations alone. Consequently, appropriate statistical tests need to be performed, and enough detail needs to be provided in the manuscript to enable independent replication of the data analysis. A comment in the Materials and Methods section that the statistical analysis was performed with a particular software package but without any further details is, therefore, not acceptable. The most common statistical procedure is analysis of variance (ANOVA) as based on fixed effects linear models or standard mixed effects linear models with diagonal variance/covariance matrices for random effects. Modern mixed linear models analysis based on restricted maximum likelihood (REML) approaches has become more common and readily available (Fig. 7). Authors should report the following fundamental information regarding the data analysis, as appropriate:

- List all effects that were originally fitted in the model.
- Specify fixed versus random effects, including justification.
- Identify any repeated measures effects in the model.
- Identify the criteria used to select the final model.
- When applicable, identify any and all covariance structures used in the final model, specifically including repeated measures autocorrelation structures, use of spatial analysis, and use of heterogeneous variance structures.
- Specify distributions fit to data and the link functions utilized, if appropriate.
- Specify if and how overdispersion was accommodated in a generalized linear mixed models approach, if necessary.
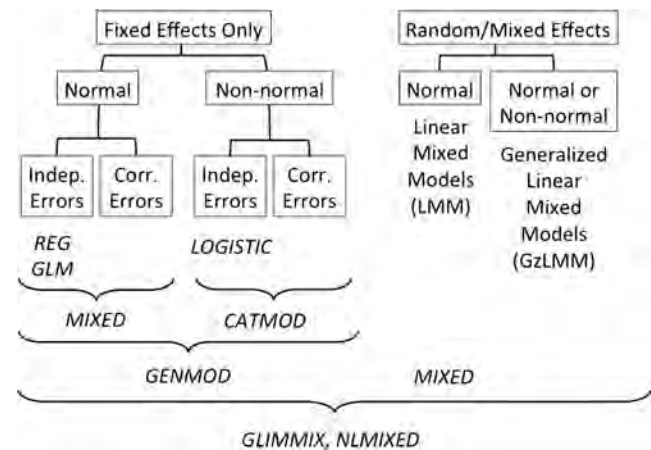

Fig. 7 Hierarchical diagram of SAS procedures (*in italics*) that are appropriate for various model assumptions and restrictions involving fixed vs. mixed or random effects at the highest level, normal vs. non-normal data at the second level, and independent vs. correlated errors at the third level. The diagram shows that generalized linear mixed models procedures that can handle the most complex cases are also the most generalizable to a wide range of model assumptions and restrictions, as well as hypothesis testing scenarios

Methods for data analysis have changed radically within the past 10 years. The classical ANOVA assumptions of normally distributed data, variance homogeneity, and uncorrelated errors are no longer required because of the development of sophisticated methods of analysis that can model a wide range of distributions and error structures. This is summarized in Fig. 7 and further explained in the next section for the interested reader.

For fixed effects of qualitative factors in which the specific levels are of interest, authors should consider the following options, in order of preference:

- Specific, pre-planned comparisons among treatment means, e.g., contrasts.
- Comparisons of each treatment to a control treatment, using Dunnett's test.
- *Ad hoc* comparisons of each treatment mean to each other treatment mean using a multiple comparison procedure, e.g., LSD, HSD, etc. It is most desirable to use a test statistic that adjusts for multiple non-independent comparisons, such as Tukey's test (HSD) or a Bonferroni correction, or a simulation approach.

For fixed effects of quantitative factors, e.g., "rate"- or "time"-related factors, scatter graphs or line graphs are generally the most informative method of presentation, accompanied by some type of regression analysis. Presentation of results for quantitative factors can be effectively accomplished in a tabular or bar graph form if the number of levels is very small, e.g., two or three, and if accompanied by appropriate statistical analysis to clarify the nature of the response. The legend of the table or figure needs to mention the number of

replicates used to calculate the standard deviation. Two excellent review papers clarify many of these general principles—despite the age of these papers these are timeless principles [34, 35].

Analysis of Large and Complex Data Sets

Proper statistical analysis is critical, especially when dealing with large and complex data sets that required substantial effort and money to generate. In most of these instances, statisticians will be involved in the design of the experiments and the analysis of the data, and *BioEnergy Research* encourages the participation of the statistician(s) in the writing of the manuscript, and hence their inclusion as co-author(s).

This section describes the recently developed analysis methods that are especially useful for large, complex, or non-normal data sets. Generalized linear mixed models (GzLMM) approaches afford the user numerous options to model a wide range of distributions commonly found in biological data (e.g., binary, binomial, negative binomial, exponential, Poisson, etc.). The GzLMM family provides researchers with both ANOVA and regression frameworks for specific distributions without the need to alter the data by transformation to meet assumptions of normality [36]. In addition, both linear mixed models and GzLMM offer the user a wide range of both spatial and temporal autocorrelation structures, as well as the opportunity to model variance heterogeneity at several levels within the experiment. Non-parametric procedures, such as the Kruskal–Wallis or Mann–Whitney test are also valid options for simple experimental designs with relatively simple hypothesis tests, and situations in which their specific assumptions can be met, e.g., symmetry for Kruskal–Wallis and similar distributions for Mann–Whitney.

Using the Statistical Analysis System [36] as a model, Fig. 7 shows eight different software procedures that are appropriate for specific statistical analysis situations involving fixed vs. mixed or random effects, normal vs. non-normal data, and independent vs. correlated errors. Proc REG is strictly a regression-based method that is appropriate in only a very limited set of circumstances. Proc GLM, Proc MIXED, Proc GENMOD, and Proc GLIMMIX are all ANOVA-based methods that are designed to solve various problems and failed assumptions within the data, with Proc GLIMMIX having the greatest flexibility and problem-solving power. Proc CATMOD and Proc LOGISTIC are methods to be used on categorical data. Proc NLMIXED is a generalized mixed model approach to be used in non-linear modeling of data. Of course, most of these procedures are offered by several other statistical packages and we do not promote SAS over any of these; we simply use SAS as the model to illustrate this general principle.

In the more modern of the above methods, models are selected based on Akaike's or Bayesian Information Criteria,

as opposed to classical F tests or t tests in ANOVA, REG, or GLM procedures. One of these criteria is used to choose models with the appropriate number of parameters, all necessary error terms or random effects, and the appropriate variance/covariance structures.

Summary and Conclusions

In this paper, we have described the general principles for proper replication across a wide array of bioenergy research topics and disciplines. This paper is intended to serve as the foundation for editorial policy regarding scale and form of replication for future studies to be published in *BioEnergy Research*. While there is flexibility in form and degree of replication for most experiments, the general principles established in this paper include the following points.

- The most important scale for replication is at the individual treatment level. Treatments should be repeated in multiple and independent experimental units.
- Additional levels of replication can be applied to an experiment at higher scales, lower scales, or both.
- A higher scale of replication can serve to replace replication at the scale of the individual treatment, if the experiment is properly designed to ensure that each treatment is applied to multiple and independent experimental units.
- Authors should adhere to the following guideline in preparing information for materials and methods: provide sufficient information so that a colleague with similar skills could repeat your experiment and analysis. This includes providing certain minimal information regarding the experimental material, the treatments, the experimental design, and the statistical data analysis.

References

1. Carmer SG, Walker WM (1988) Significance from the statistician's viewpoint. J Prod Agric 1:27–33
2. Quinn GP, Keough MJ (2002) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge
3. Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. Ecol Mono 54:187–211
4. Casler MD, Vogel KP (2014) Selection for biomass yield in upland, lowland, and hybrid switchgrass. Crop Sci 54:626–636
5. Muttoni G, Johnson JM, Santoro N, Rhiner CJ, von Mogel KJH, Kaeppler SM, de Leon N (2012) A high-throughput core sampling device for the evaluation of maize stalk composition. Biotech Biofuels 5:27. doi:10.1186/1754-6834-5-27
6. Headlee WL, Zalesny RS Jr, Hall RB, Bauer EO, Bender B, Birr BA, Miller RO, Randall JA, Wiese AH (2013) Specific gravity of hybrid poplars in the north-central region, USA: within-tree variability and site×genotype effects. Forests 4:251–269
7. Casler MD (2013) Fundamentals of experimental design: guidelines for designing successful experiments. Agron J 105. doi:10.2134/agronj2013.0114
8. Steel RGC, Torrie JH, Dickey DA (1996) Principles and procedures in statistics, 3rd edn. McGraw-Hill, New York
9. Casler MD (1999) Repeated measures vs. repeated plantings in perennial forage grass trials: an empirical analysis of precision and accuracy. Euphytica 105:33–42
10. Williams ER, Matheson AC, Harwood CE (2002) Experimental design and analysis for tree improvement. CSIRO Publishing, Collingwood, 214pp
11. Lin CS, Binns MR (1984) Working rules for determining the plot size and numbers of plots per block in field experiments. J Agric Sci 103:11–15
12. Lin CS, Binns MR (1986) Relative efficiency of two randomized block designs having different plot sizes and numbers of replications and plots per block. Agron J 78:531–534
13. Fisher MM, Getis A (eds) (2010) Handbook of applied spatial analysis: Software tools, methods, and applications. Springer, New York, 811pp
14. Stroup WW (2002) Power analysis based on spatial effects mixed models: a tool for comparing designs and analysis strategies in the presence of spatial variability. J Agric Biol Environ Stat 7:491–511
15. Brien CJ, Berger B, Rabie H, Tester M (2013) Accounting for variation in designing greenhouse experiments with special reference to greenhouses containing plants on conveyor systems. Plant Methods 9:5 http://www.plantmethods.com/content/9/1/5
16. Hardy EM, Blumenthal DM (2008) An efficient and inexpensive system for greenhouse pot rotation. HortSci 43:965–966
17. Kempthorne O (1957) 126. Query: arrangements of pots in greenhouse experiments. Biometrics 13:235–237
18. Edmondson RN (1989) Glasshouse design for repeatedly harvested crops. Biometrics 45:301–307
19. Geurtal EA, Elkins CB (1996) Spatial variability of photosynthetically active radiation in a greenhouse. J Am Soc Hort Sci 121:321–325
20. Wallihan EF, Garber MJ (1971) Efficiency of glasshouse pot experiments on rotating versus stationary benches. Plant Physiol 48:789–791
21. Arnoult S, Quillet MC, Brancourt-Hulmel M (2014) Miscanthus clones display large variation in floral biology and different environmental sensitivities useful for breeding. BioEnergy Res 7:430–441
22. Piepho H-P, Möhring J, Williams ER (2013) Why randomize agricultural experiments? J Agron Crop Sci 199:374–383
23. Rath J, Heuwinkel H, Hermann A (2013) Specific biogas yield of maize can be predicted by the interaction of four biochemical constituents. BioEnergy Res 6:939–952
24. Gbur EE, Stroup WW, McCarter KS, Durham S, Young LJ, Christman M, West M, Kramer M (2012) Analysis of generalized linear mixed models in the agricultural and natural resources sciences. ASA-CSSA-SSSA, Madison
25. McCann LC, Bethke PC, Casler MD, Simon PW (2012) Allocation of experimental resources used in potato breeding to minimize the variance of genotype mean chip color and tuber composition. Crop Sci 52:1475–1481
26. Smith AB, Lim P, Cullis BR (2006) The design and analysis of multiphase plant breeding experiments. J Agric Sci 144:393–409
27. Williams E, Piepho H-P, Whitaker D (2011) Augmented p-rep designs. Biom J 1:19–27
28. Shuford CM, Sederoff RR, Chiang VL, Muddiman DC (2012) Peptide production and decay rates affect the quantitative accuracy of protein cleavage isotope dilution mass spectrometry (PC-IDMS). Mol Cell Proteomics 11:814–823
29. Yuan JS, Reed A, Chen F, Stewart CN Jr (2006) Statistical analysis of real-time PCR data. BMC Bioinforma 7:85
30. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert CJJ, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FCP, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans S, Schulze KS, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME): toward standards for microarray data. Nat Genet 29:365–371
31. Brown CS, Goodwin PC, Sorger PK (2000) Image metrics in the statistical analysis of DNA microarray data. Proc Natl Acad Sci U S A 98:8944–8949
32. Naoumkina M, Farag MA, Sumner LW, Tang Y, Liu C-J, Dixon RA (2007) Different mechanisms for phytoalexin induction by pathogen and wound signals in *Medicago truncatula*. Proc Natl Acad Sci U S A 104:17909–17915
33. Auer PL, Doerge RW (2010) Statistical design and analysis of RNA sequencing data. Genetics 185:405–416
34. Carmer SG, Walker WM (1982) Baby bear's dilemma: a statistical tale. Agron J 74:122–124
35. Chew V (1976) Comparing treatment means: a compendium. HortSci 11:348–357
36. Stroup WW (2014) Rethinking the analysis of non-normal data in plant and soil science. Agron J doi:10.2134/agronj2013.034237. SAS (2014) Statistical Analysis System. Cary, NC. http://support.sas.com/documentation/