nature communications

Perspective

https://doi.org/10.1038/s41467-025-62616-x

How thoughtful experimental design can empower biologists in the omics era

Received: 6 July 2025

Accepted: 24 July 2025

Published online: 06 August 2025

Check for updates

Maggie R. Wagner [1,3] & Manuel Kleiner [2]

The modern biology toolbox continues to evolve, as cutting-edge molecular techniques complement some classic approaches and replace others. However, statistical literacy and experimental design remain critical to the success of any empirical research, regardless of which methods are used to collect data. This Perspective highlights common experimental design pitfalls and explains how to avoid them. We discuss principles of experimental design that are relevant for all biology research, along with special considerations for projects using -omics approaches. Established best practices for optimizing sample size, randomizing treatments, including positive and negative controls, and reducing noise (e.g., blocking and pooling) can empower researchers to conduct experiments that become useful contributions to the scientific record, even if they generate negative results. They also reduce the risk of introducing bias, drawing incorrect conclusions, or wasting effort and resources on experiments with low chances of success. Although experimental design strategies are often covered in undergraduate- and graduate-level courses and in textbooks, here we provide a succinct overview and highlight their relevance to modern biology research. This Perspective can be used in training of early-career scientists and as a refresher for seasoned scientists.

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of." – Ronald A. Fisher

The modern biology toolbox is larger than ever, as a widening array of cutting-edge molecular techniques supplements the classic approaches that still drive the field forward. Teams of researchers with complementary expertise may combine these tools in countless creative ways to produce novel research. Regardless of which methods are chosen for data collection, all empirical research projects share a common foundation: the principles of good experimental design. Far from eliminating the need for statistical literacy, -omics technologies make careful, sound experimental design more important than ever[1,2].

This Perspective was motivated by the observation that many biology projects are doomed to fail by experimental design errors that make rigorous inference impossible. The results of such errors can range from waste to, in the worst case, the introduction of misleading conclusions into the scientific literature, including those with clinical consequences. Our goal is to highlight common experimental design errors and how to avoid them. Although methods for analyzing challenging datasets are improving[3–6], even advanced statistical techniques cannot rescue a poorly designed experiment. For this reason, we focus on choices that must be made before an experiment or study is conducted, rather than on data analysis choices (however, if you are working on microbiomes, we highly recommend that you read this article in conjunction with Willis and Clausen[7], who highlight important points for planning and reporting on statistical analyses). In particular, we address four key elements of a well-designed experiment: adequate replication, inclusion of appropriate controls, noise reduction, and randomization.

First, we discuss how in -omics research, many errors arise because of the misconception that a large quantity of data (e.g., deep sequencing or the measurement of thousands of genes, molecules, or microbes)

[1]Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, USA. [2]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC, USA. [3]Kansas Biological Survey & Center for Ecological Research, Lawrence, KS, USA. e-mail: maggie.r.wagner@ku.edu

ensures precision and statistical validity. In reality, it is the number of biological replicates that matters. We also explain how to recognize and avoid the problem of pseudoreplication, and we introduce power analysis as a useful method for optimizing sample size. Second, we introduce several strategies (blocking, pooling, and covariates) for minimizing noise, or variation in the data caused by randomness or other unplanned factors. Third, we briefly review how missing positive and negative controls can compromise experimental results, and provide examples from both -omics and non-omics research. Finally, we describe two critical functions of experimental randomization: preventing the influence of confounding factors, and empowering researchers to rigorously test for interactions between two variables.

While these practices are covered in undergraduate and graduate level classes and in textbooks, as reviewers we—along with editors and some colleagues—have observed basic experimental design errors in submitted manuscripts. Indeed, the authors of this Perspective have also made and learned from some of these errors the hard way. This fact inspired this Perspective to provide a succinct overview of important experimental design principles that can be used in training of early-career scientists and as a refresher for seasoned biologists. We present examples from projects that use high-throughput DNA sequencing; however, unlike best methods for statistical analysis, these experimental design principles apply equally to any experiment regardless of the type of data being collected, including proteomics, metabolomics, and non-omics data. We also provide a list of additional resources on the topics discussed (Table 1) and practical steps for designing a rigorous experiment (Box 1).

Empowerment through replication

Many researchers intuitively understand that any individual data point might be an outlier or a fluke. As a result, we have very low confidence in any conclusion that was reached based on one isolated observation. With each additional data point that we collect, however, we get a better sense of how representative that first data point really was, and our confidence in our conclusion grows. Statistics empower us to measure and communicate that confidence in a way that other researchers will understand.

Why biological replication is more important than sequencing depth. Most biologists have heard that having more data will empower them to test their hypotheses. But what does it really mean to have more data? High-throughput technologies that generate millions to billions of DNA sequence reads, and counts of thousands of different genes or microbes, can create the illusion of a big dataset even if the *number of replicates*, or sample size, remains small. Although deeper sequencing per replicate can improve power in some cases, it is primarily the number of biological replicates that enables researchers to obtain clear answers to their questions.

To illustrate why this is, consider the hypothesis that two species of plants host different ratios of two microbial taxa in their roots. We can estimate this ratio for the two groups or populations of interest (i.e., all existing individuals of the two species) by collecting random samples from those populations. A sample size of 1 plant per species would be essentially useless, because we would have no way of knowing whether that plant is representative of the rest of its population, or instead is an anomaly. This is true regardless of the *amount* of data we have for that plant; whether it is based on 10^3 sequence reads or 10^7 sequence reads, it is still an observation of a single individual and cannot be extrapolated to make inferences about the population as a whole. Similarly, if we measure the abundances of thousands of microbes per plant, this same problem would apply to our estimates for each of those microbes. In contrast, measuring more plants per species will provide an increasingly better sense of how variable the trait of interest is in each population.

To what extent does the amount of data per replicate matter? Deeper sequencing can modestly increase power to detect differential abundance or expression, but those gains quickly plateau after a moderate sequencing depth is achieved[3,8]. Extra sequencing is most beneficial for the detection of less-abundant features, such as rare microbes or low-expression transcripts, and features with high variance[8]. Projects that focus specifically on such features will require deeper sequencing than those that do not, or else may benefit from a more targeted approach. Finally, it is worth highlighting the related problem of treating -omics features (e.g., genes or microbial taxa) as the units of replication, as is common in gene set enrichment and pathway analyses. Such analyses describe a pattern within the existing dataset, but they are entirely uninformative about whether that pattern would hold in another group of replicates[9]. Instead, they only allow inference about whether that pattern would hold for a newly-observed feature in the already-measured group of replicates, which is often not the researcher's intended purpose.

Replication at the right level. Biological replicates are crucial to statistical inference precisely because they are randomly and independently selected to be representatives of their larger population. The failure to maintain independence among replicates is a common experimental error known as *pseudoreplication*[10] (Box 2). When experimental units are truly independent, no two of them are expected to be more similar to each other than any other two. Pseudoreplication becomes a problem when the incorrect unit of replication is used for a given statistical inference, which artificially inflates the sample size and leads to false positives and invalid conclusions (Fig. 1). In other words, not all data points are necessarily true replicates.

Although pseudoreplication is occasionally unpreventable (particularly in large-scale field studies), it can and should be anticipated and avoided whenever possible. The correct units of replication are those that can be randomly assigned to receive one of the treatment conditions that the experiment aims to compare. In experimental evolution, for instance, the replicates are random subsets of the starting population, assumed to be identical, each of which may be assigned to a different selective environment[11,12]. Failure to include enough independent sub-populations, or to keep them independent throughout the experiment (*e.g.*, by pooling replicates; Fig. 1C–D), will cause pseudoreplication of the evolutionary process of interest[13]. In some cases, however, mixed-effects modeling techniques can adequately account for the non-independence of replicates[10].

Optimizing sample size

If most measurements in a dataset are similar to each other, this indicates that we are measuring individuals from a *low-variance population* with respect to the dependent variable. In contrast, a wide range of trait values signals a *high-variance population*. This within-group variance (i.e., the variance within one population) is central to determining how many biological replicates are necessary to achieve a clear answer to a hypothesis test: when within-group variance is high relative to the *between-group variance*, more replicates are required to achieve a given level of confidence (Fig. 2A). However, when the budget for sequencing is fixed, then increasing the sample size is costly—not only because wet lab costs increase, but also because the amount of data per observation decreases. Too many replicates can waste time and money, while too few can waste an entire experiment. How can a biologist know ahead of time how many replicates are enough?

A flexible but underused solution to this problem—*power analysis*—has existed for nearly a century[14,15]. Power analysis is a method to calculate how many biological replicates are needed to detect a certain effect with a certain probability, if the effect exists (Fig. 2A). It has five components: (1) sample size, (2) the expected effect size, (3) the within-group variance, (4) false discovery rate, and (5) *statistical power*,

Table 1 | Additional resources for improving experimental design strategies

Type	Description	Ref.
General experimental design		
Book	Introductory-level textbook covering the basic principles of statistics, using examples from microbiology.	Biostatistics and Microbiology: A Survival Manual (2009)[51]
Book	Comprehensive, clearly-written introductory-level textbook that covers both theory and practice of designing experiments and analyzing data.	The Analysis of Biological Data (2019)[52]
Book	Introductory textbook: key definitions and boxes with in-depth discussions and take-home messages. Flow charts for design decision making. Covers some experimental aspects such as within-subject designs that are not well-covered in other resources.	Experimental Design for the Life Sciences (2017)[53]
Book	Provides experimental design basics and detailed explanations and examples on choosing models, checking model assumptions, variations of specific experimental design types e.g., incomplete block designs, and implementation examples in different softwares.	Design and Analysis of Experiments (2017)[54]
Book	This book features similar topics as the one in the previous line. However, it has additional models and also explains how to deal with missing observations.	Principles of Experimental Design (2008)[55]
Journal article	Explores the basics of experimental design in greater detail, including the setup and analysis of several commonly used blocking and randomization approaches.	Fundamentals of Experimental Design: Guidelines for Designing Successful Experiments (2015)[56]
Journal article	Covers randomization, blocking, pooling, and power analysis with an emphasis on proteomics experiments.	Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments (2009)[57]
Book	Provides theoretical and practical explanations of many types of experimental design, implemented in the R statistical coding language.	Optimal Experimental Design with R (2011)[58]
Journal article	Illustrates how a carefully planned experimental design (in this case, the use of specialized designs for optimizing bacterial growth) can solve experimental challenges.	Greater enhancement of *Bacillus subtilis* spore yields in submerged cultures by optimization of medium composition through statistical experimental designs (2010)[59]
Resource	A curated list of R packages that are useful for various aspects of experimental design.	CRAN Task View: Design of Experiments (DoE) & Analysis of Experimental Data (2025)[60]
Blocking and randomization		
Book	Accessible introductions to the formal "Design of Experiments" (DOE) methodology commonly used in industrial research.	Design of Experiments for Engineers and Scientists (2023)[61]
Journal article	A clear empirical demonstration of the benefits of randomized block designs for -omics studies.	Blocking and Randomization to Improve Molecular Biomarker Discovery (2014)[62]
Journal article	Explains why testing the effects of multiple factors simultaneously is preferable, from a pragmatic standpoint, to testing one factor at a time.	The Importance of Outcome Dynamics, Simple Geometry, and Pragmatic Statistical Arguments in Exposing Deficiencies of Experimental Design Strategies (1996)[1], Statistical Design of Experiments for Synthetic Biology (2021)[63]
Book chapter	A detailed overview of blocking methods and related concepts that can be applied to any type of experiment, from the field to the lab.	Blocking Principles for Biological Experiments. in Applied Statistics in Agricultural, Biological, and Environmental Sciences (2018)[64]
Improving signal-to-noise ratio		
Journal article	Illustrates an approach to use spatial information as covariates to reduce noise and enable detection of subtle biological effects in field experiments.	Increased signal-to-noise ratios within experimental field trials by regressing spatially distributed soil properties as principal components (2022)[34]
Journal article	Empirical comparison of two methods for reducing noise: blocking *vs.* inclusion of spatial covariates.	Can Spatial Modeling Substitute for Experimental Design in Agricultural Experiments? (2019)[65]
Replication		
Journal article	Explains the differences between biological units, experimental units, and observational units in lab-based research.	What exactly is 'N' in cell culture and animal experiments? (2018)[66]
Journal article	A balanced discussion of when pseudoreplication is or is not a problem, with practical suggestions for experimental design.	Using Biological Insight and Pragmatism When Thinking about Pseudoreplication. (2018)[10]
Journal article	Explains why gene set enrichment analysis and pathway analysis cannot be used to make inferences about what to expect in a new group of samples.	Analyzing gene expression data in terms of gene sets: methodological issues (2007)[9]
Book	Focuses on the design and analysis of experiments that require repeated or multiple observations from each experimental unit of replication.	Repeated Measures Design for Empirical Researchers (2015)[67]
Power analysis		
Tutorial	Clearly demonstrates how to conduct power analysis for a typical microbiome experiment using the R package micropower.	Tutorial on power analyses for microbiome analyses with micropower (2023)[68], Power analyses for microbiome studies with micropower (2020)[68]
Journal article	Describes approaches to power analysis for various types of hypothesis test that may be conducted using microbiome data; some can be generalized to any -omics data.	The rise to power of the microbiome: power and sample size calculation for microbiome studies (2022)[28]
Journal article	Describes a tool for power analysis that illustrates many complex aspects of -omics data, and its use for conducting a cost-benefit analysis.	scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies (2021)[69]

Table 1 (continued) | Additional resources for improving experimental design strategies

Type	Description	Ref.
Journal article	Describes the use of intermediate/advanced mathematical tools to analyze microbiome and other high-dimensional data and to conduct power analysis using simulated -omics datasets while conserving covariance among features.	A pathway for multivariate analysis of ecological communities using copulas (2019)[70]
Journal article	Meta-analysis that assesses whether previous studies on obesity and the gut microbiome were sufficiently powered.	Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome (2016)[71]

BOX 1.
Putting it into practice

Thinking about statistics early in project development will minimize your need to use advanced statistical techniques later and will maximize your work's scientific value. It will certainly prevent you from spending time and money on an experiment that is doomed to fail. The steps outlined below describe how to work backwards from your planned analysis to ensure a properly designed experiment.

1. Think about your experimental goals. If you are doing hypothesis-based research, what kinds of observations would support or refute your hypothesis? Does your goal relate to an interaction between two or more independent (explanatory) variables? With that in mind, sketch a plan to generate the data you need to test your hypothesis; then, write out the statistical test you will use and explicitly define the units of replication for that test.
2. Create a mock dataset for your planned experiment: a spreadsheet with a row for each replicate, and a column for each *independent variable* (any variables that may influence the outcome). For now, include just 3 replicates per experimental group. If there are multiple independent variables, ensure that all necessary combinations of those variables are represented. Several mock dataset templates are provided in the Supplementary Information.
3. Add a column to hold randomly-generated numbers, which you can use later to randomize experimental units. Add columns to hold any covariates and nuisance variables (e.g., batches) that you will want to control for later.
4. Add columns to hold measurements (your *response* or *dependent variables*). For -omics projects these are often features such as ASVs, transcripts, or metabolites; they may also be higher-level properties of those features such as microbial phyla or gene ontology categories. Consider what form your measurements will take; for example, many types of -omics data are counts ranging from zero to large integers.
5. Run a practice statistical test using your mock dataset. This will ensure that all the variables you need to test your hypothesis are included in your dataset. As a bonus, you will be well prepared to quickly analyze your real data when you get it.
6. Anticipate the possible outcomes of this statistical test. What are alternative explanations for each of those possibilities? If there are changes you could make to the experiment or analysis to remove the ambiguity, return to Step 1 and revise accordingly.
7. Optimize your sample size. Use your domain expertise to define the minimum effect size that you would consider biologically valuable, and conduct a power analysis for your chosen statistical test. Then, revisit your mock dataset and add more rows as needed to increase the sample size.
8. Use the literature and your domain expertise to determine what positive and negative controls are necessary for correct interpretation of your expected results.
9. Share your completed plan with colleagues and ask for their feedback. If they were reviewing this work, would they be convinced by the results? Do they think the plan is feasible? If not, what improvements would they suggest?
10. Consider submitting your finalized plan for peer review at a journal that publishes Registered Reports (Box 3).

or the probability that a false null hypothesis will be successfully rejected. By defining four of these, a researcher can calculate the fifth.

Usually, both the effect size and within-group variance are unknown because the data have not yet been collected. The choice of effect size for a power analysis is not always obvious: researchers must decide a priori what magnitude of effect should be considered biologically important. Acceptable solutions to this problem include expectations based on the results of small pilot experiments, using values from comparable published studies or meta-analyses, and reasoning from first principles. For example, a biologist planning to test for differential gene transcription may define the minimum interesting effect size as a 2-fold change in transcript abundance, based on a published study showing that transcripts stochastically fluctuate up to 1.5-fold in a similar system. In this scenario, the stochastic 1.5-fold fluctuations in transcript abundance suggest a reasonable within-group variance for the power analysis. As another example, a bioengineer may only be interested in mutations that increase cellulolytic enzyme activity by at least 0.3 IU/mL relative to the wild type, because that is known to be the minimum necessary for a newly designed bioreactor. To determine the within-group variance for a power analysis, the bioengineer could conduct a small pilot study to measure enzyme activity in wild-type colonies. As a final example, if a preliminary PerMANOVA for a very small amount of metabolomics data (say, $N = 2$ per group) from a pilot study estimated that $R^2 = 0.41$, then a researcher could use that value as the target effect size for which he/she aims to obtain statistical support.

Freely available software facilitates power analysis for basic statistical tests of normally-distributed variables, including correlation, regression, *t*-tests, chi-squared tests, and ANOVA[16–19]. Power analysis for -omics studies, however, is more complex for several reasons. -Omics data comprise counts that are not normally distributed and may contain many zeroes, and some features may be inherently correlated with each other; these properties must be modeled appropriately for a power analysis to be useful[20]. The large number of features often requires adjustment of *P*-values to correct for inflated Type I error rates, which in turn decreases power. Furthermore, statistical power varies among features because they differ not only in within-group variance, but also in their overall abundance. In general, more replicates are required to detect changes in low-abundance features than in high-abundance ones[8]. Statistical power also varies among different alpha- and beta-diversity metrics[21] that are commonly used in amplicon-sequencing analysis of microbiomes. For power

BOX 2.

Examples of inconclusive and/or incorrect conclusions due to experimental design errors

Pseudoreplication. Lack of clarity during experimental design about the actual experimental unit can lead to incorrect assumptions about the number of replicates available for statistical testing. A highly-cited 2012 study[72], for example, concluded that the transfer of intestinal microbiota from pregnant women in their third trimester into germ-free mice induces greater adiposity and inflammation as compared to microbiota from women in the first trimester. To properly support this conclusion, the necessary experimental unit is the microbiota from a pregnant woman in her first or third trimester, requiring the microbiota of several women in each group to be tested separately for proper replication. The authors, however, pooled the microbiota of five women per condition and used this as the inoculum for six germ-free mice, resulting in N = 1 per condition (only one third-trimester inoculum and one first-trimester inoculum). The mice are in this case observational units, but not experimental units. Therefore, the broad conclusions drawn about effects of first versus third trimester microbiomes on mouse phenotypes are not statistically valid. This incorrect use of the observational unit (mouse) as a replicate instead of the experimental unit (individual human-derived inocula) has been widespread in experiments transferring human microbiomes into germ-free mice[73].

Lack of appropriate controls. While it is difficult to show post hoc that conclusions of a study are actually false, the absence of appropriate controls can put results and conclusions into serious doubt. (1) A 2019 study[74] compared microbiome composition of rats fed two different sources of dietary protein (casein versus chicken) using metagenomics. The authors found that *Lactococcus lactis* was significantly higher in rats fed the casein diet, however, the experiment did not have controls to test for potential contamination of food sources with microbial DNA. It has been shown previously that casein contains high amounts of *L. lactis* DNA and protein[75,76], and while we cannot be certain that this applies to the rat diet study as well, due to the absence of controls it is likely that the enrichment of *L. lactis* is a false positive result. (2) A 2016 study[77] used 16S rRNA gene sequencing to investigate the bacterial communities associated with the roots of maize planted in sterile sand and two soils. The plants grown in sterile sand shared >20 OTUs (microbial taxa) with the plants grown in natural soils, and from this the authors concluded that transmission via the seed must be a major source of bacteria found in maize plants, as the bacteria in the sterile sand could not have come from the soil. However, no negative controls were sequenced, raising the possibility that contamination introduced during sample processing such as the "kit-ome"[78] was the true source of the shared OTUs. This puts the conclusions relating to the presence of microbial taxa in sterile plants into question. (3) A major controversy in the microbiome field in the last decade was the question of whether the placentas of healthy pregnant women contain a microbiome. Initially several studies presented evidence for a placental microbiome based on amplicon sequencing of microbial marker genes. However, all of these studies lacked appropriate positive and negative controls and did not account for potential contamination during sample processing. Ultimately, upon inclusion of proper controls, the presence of a placental microbiome was solidly refuted by many studies[79]. The controversy led to many improvements in microbiome sequencing procedures.

analysis of multivariate tests (e.g., PERMANOVA), which incorporate information from all features simultaneously to make broad comparisons between groups, simulations are necessary to account for patterns of co-variation among features. In such cases, pilot data or other comparable data are crucial for accurate power analysis.

Fortunately, tools are available to estimate power for common analyses used in proteomics[22], RNA-seq[3,20,23,24], and microbiome studies[21,25,26]. Recent reviews[27,28] demonstrate this process for various forms of amplicon-sequencing data, including taxon abundances, alpha- and beta-diversity metrics, and taxon presence/absence. They also consider several types of hypothesis tests, including comparisons between groups and correlations with continuous predictors. Although power analysis may seem daunting at first, it is an investment with excellent returns. This skill empowers biologists to budget more accurately, write more convincing grant proposals, and minimize their risk of wasting effort on experiments that cannot generate conclusive results.

Empowerment through noise reduction

As explained above, statistical power is positively related to the sample size and negatively related to the within-group variance. Thus, we can increase power either by including more biological replicates or by reducing within-group variance. Because our budgets do not allow us to increase replication indefinitely, it is useful to consider: What methods exist to increase power by decreasing within-group variance?

A classic way to minimize within-group variance is to remove as many variables as possible[29]. Common examples include using a single strain instead of multiple; strictly controlling the lab environment; and using only male host animals[30]. Such practices minimize noise, or variance contributed by unplanned, unmeasured variables. However, they come with a drawback: the loss of generalizability[30–32]. If a mutation's phenotype cannot be detected in the face of minor environmental variation, is it likely to be relevant in nature? If an interesting function occurs only in a lab strain, is it important for the species in general? Such limitations should always be considered when interpreting results.

Another technique to reduce within-group variance is *blocking*, the physical and/or temporal arrangement of replicates into sets based on a known or suspected source of noise (Fig. 2B). For instance, clinical trials may block by sex, so that results will be based on comparisons of males to males and females to females. Crucially, all experimental treatments must be randomly assigned within each block. Blocking is also useful for large experiments with more replicates than can be measured at once; as long as the replicates within each block are treated identically, sources of noise such as multiple observers can be controlled. The most powerful form of blocking is *paired design*, in which experimental units are kept in pairs from the beginning. One unit per pair is randomly assigned to the treatment group, the other to the control; the difference between units is then calculated directly, automatically accounting for sources of noise that are shared within the pair (Fig. 2B, bottom panel). A possible downside of highly-structured blocking designs is that they can complicate the re-use of the data to answer questions other than the one for which the design was optimized, whereas a simple randomized design is highly flexible but not optimized for any particular purpose.

When sources of noise are known beforehand, they ideally should be measured during the experiment so that they can be used as *covariates*. A covariate is any independent variable that is not the focus of an analysis but might influence the dependent variable. When using regression, ANOVA, or similar methods, one or more covariates may be included in the statistical model to control for their effects on the variable of interest (Fig. 2C). For instance, controlling for replicates'

BOX 3.

Registered Reports can empower biologists

The steps outlined in this Perspective (Box 1) can be used informally to optimize research plans before acting on them. However, journals can formalize them as an alternative to the standard publishing model[80]. *Registered Reports* are a mode of publication in which the decision to accept or reject a manuscript is made before results are known; the peer review process considers only the impact of the project and the soundness of the research plan. When a journal accepts a Registered Report, it commits to publishing the resulting manuscript regardless of the outcome, as long as the authors have followed the peer-reviewed plan, met reasonable quality standards, and interpreted their results fairly.

This publishing model recognizes that far from being inconclusive, negative results from well-designed, adequately-powered studies are valuable additions to the scientific literature. In contrast, positive results from poorly-designed or poorly-analyzed experiments are less valuable than they seem, or even harmful. Even careful studies can generate false positives due to random chance[81]. Nevertheless, authors, editors, and readers alike typically find positive results more exciting than negative results. The consequence is a severe over-representation of positive results in the scientific literature[82,83]. This publication bias can generate inaccurate conclusions[84], mislead practitioners[85], incentivize questionable research practices[81,86], and waste time and resources[87,88].

Registered Reports will benefit not only the scientific enterprise in general[89], but also individual researchers. First, authors of Registered Reports receive valuable expert feedback on their experimental and analytical plans. Second, Registered Reports prevent the waste of resources on doomed experiments, which empowers PIs to optimize their budgets and protects the time of all project participants. Third, they free authors from the pressure to obtain positive results. This especially benefits early-career researchers whose professional trajectories depend on publishing in high-quality journals.

To be sure, Registered Reports would bring challenges–for instance, managing the timeline from Report acceptance to publication, if delays arise due to funding uncertainty, personnel turnover, or technical difficulties. Another perceived challenge is the restriction on conducting analyses other than those prescribed in the accepted Report. Indeed, the freedom to explore a rich dataset can lead to new insights and the discovery of unexpected patterns. This is an important part of the scientific process. A well-designed Registered Reports policy will allow authors to include such analyses in their publications, given that they are labeled as exploratory. Authors can immediately follow up on such results by submitting a new Registered Report for review. Overall, the benefits of Registered Reports outweigh the downsides; in any case, most journals that offer them do so as an option, not as a mandate.

spatial locations can dramatically reduce noise in field studies[33,34]. Similarly, dozens of covariates related to lifestyle and medical history contribute to the variation in human fecal microbiome composition[35]. The authors showed that by controlling for just three of these covariates, the minimum sample size needed to detect a microbiome contrast between lean and obese individuals would decrease from 865 to 535 individuals per group.

Finally, *pooling*–combining two or more replicates from the same group prior to measurement–can reduce within-group variance and the influence of outliers[36]. For example, pooling RNA can empower biologists to detect differential gene expression from fewer replicates, reducing costs of library preparation and sequencing[37]. This approach is especially helpful for detecting features that are low-abundance and/or have large within-group variance. Its main drawbacks are that it reduces sample size and results in the loss of information about specific individuals, which may be necessary for unambiguously connecting one dataset to another or linking the response variable to covariates. In fact, excessive or unnecessary pooling is a common experimental design error that can eliminate replication (Fig. 1C–D; Box 2), but is easily avoided by remembering that the pools themselves are the biological replicates.

Empowerment through inclusion of appropriate controls

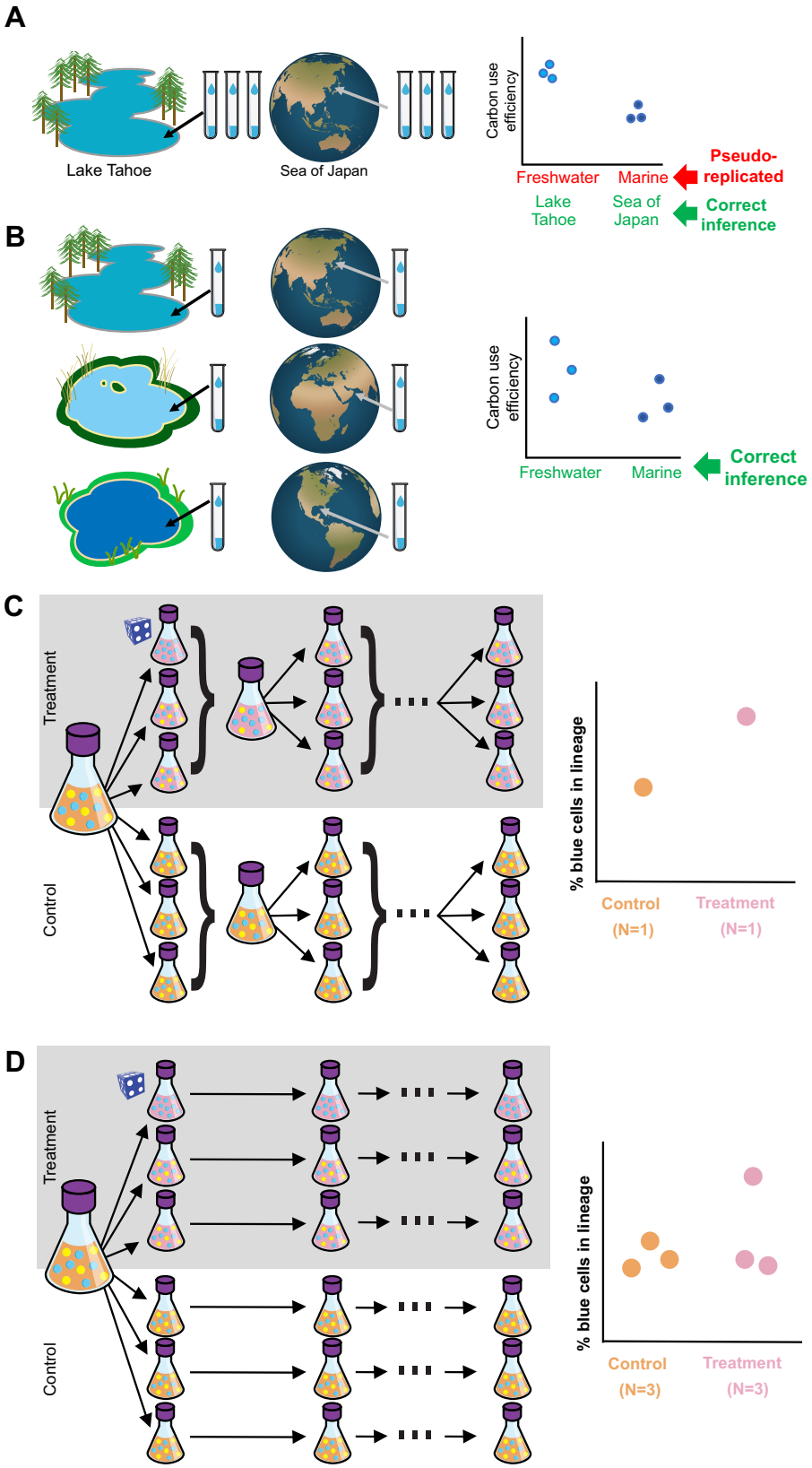
Key to any experimental design is the inclusion of appropriate positive and negative controls to strengthen conclusions and enable correct interpretation of the experimental results (Box 2). Positive controls can confirm that experimental treatments and measurement approaches work as expected. Positive controls can, for example, be samples with known properties that are carried through a procedure from start to end alongside the experimental samples. For microbiome-related protocols, mock community samples can often be a good positive control[38]. In addition to serving as positive controls, spike-ins can serve as internal calibration and normalization standards[39,40]. Negative controls allow detection of artifacts introduced during the experiment or measurement. Negative controls can be samples without the addition of the organism/analyte/treatment of interest that are carried alongside the experimental samples through all experimental steps. They often reveal artifacts caused by the matrix/medium in which samples are embedded; for example, to analyze secreted proteins in the culture supernatant of a bacterium, measuring the proteins in non-inoculated culture medium would be a critical negative control. In microbiome and metagenomics studies, negative controls are particularly crucial when working with low-biomass samples because contaminants from reagents can become prominent features in the resulting datasets[41].

Empowerment through randomized design

The random arrangement of biological replicates with regards to space, time, and experimental treatments is a crucial research practice for several reasons.

Randomization protects against confounding variables. For a given level of replication, the researcher must decide how to distribute those replicates in time and space. The importance of *randomization*—the arrangement of biological replicates in a random manner with respect to the variables being tested—has long been appreciated in ecology, clinical research, and other fields where external influences are impossible to control. Even in relatively homogenous lab settings, failure to randomize can lead to ambiguous or misleading results. This is because randomization minimizes the possibility that unplanned, unmeasured variables will be *confounded* with the treatment groups[42]. Unlike experimental noise—random deviations that decrease power by increasing variance—confounding variables cause a subset of replicates to deviate systematically from the others in a way that is unrelated to the intended experimental design. Thus, they cause biased results. Fortunately, confounding variables that are structured in time and/or space can be controlled through randomization (Fig. 3). Even for complex experimental designs, randomization can be easily achieved by sorting a spreadsheet based on a randomly generated number to assign the position of each replicate. In a *fully randomized design*, all replicates in an experiment are randomized as a single group. In structured designs

https://doi.org/10.1038/s41467-025-62616-x

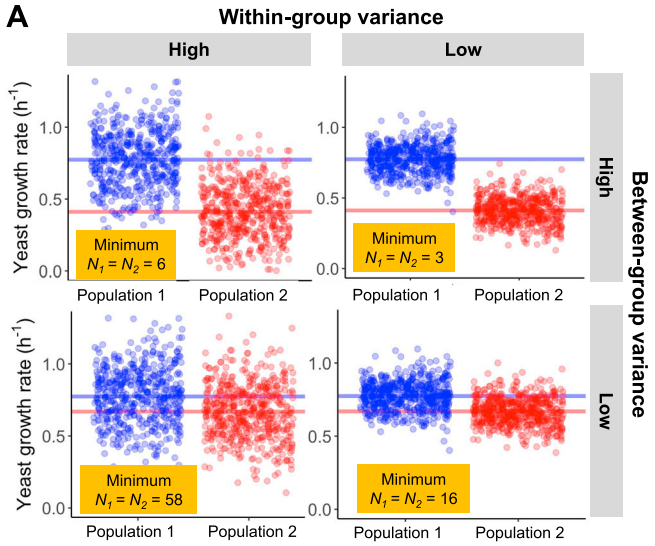


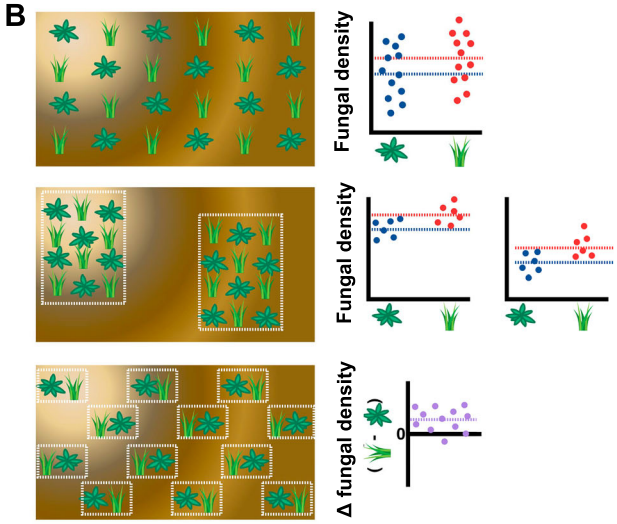
such as an experiment that employs blocking (see "Empowerment through noise reduction"), however, the best approach is to randomize the replicates within each block, independently of the other blocks.

Projects using -omics techniques are particularly vulnerable to *batch effects*, a common type of confounding variable with the potential to invalidate an experiment[1,43–46] (Fig. 3C). When not all replicates can be processed simultaneously, they must be divided into batches that often vary in their exposure to not only chronological factors but also variation among reagent lots, sequencing runs, and other technical factors. While batch effects can be minimized through careful control of

https://doi.org/10.1038/s41467-025-62616-x

Fig. 1 | Valid experimentation requires independence of all replicates. A To test whether carbon use efficiency differs between freshwater and marine microbial communities, a researcher collects three vials from Lake Tahoe and three from the Sea of Japan. This design is pseudoreplicated because the vials from the same location are not independent of each other; they are expected to be more similar to each other than they would be to other vials randomly sampled from the same population (i.e., freshwater or marine). However, they could be used to test the narrower question of whether carbon use efficiency differs between Lake Tahoe and the Sea of Japan, assuming that the vials were randomly drawn from each body of water. **B** In contrast to the design in panel A, collection of one vial from each of three randomly-selected freshwater bodies and three randomly-selected saltwater bodies enables a valid test of the original hypothesis. Alternatively, each replicate could be the composite or average of three sub-samples per location; this would be an example of pooling to improve the signal:noise ratio. **C** Pseudoreplication during experimental evolution can lead to false conclusions. Each time that the flasks are pooled within each treatment prior to passaging, independence among replicates is eliminated. As a result, a stochastic event that arises in a single replicate lineage (symbolized by the blue die) can spread to influence the other lineages, so that the stochastic event is confounded with the treatment itself. **D** In contrast, by maintaining strict independence of the replicate lineages, the influence of the stochastic event is isolated to a single replicate. The researcher can confidently rule out the possibility that a stochastic event has systematically influenced one of the treatment groups.





C

Model 1: `Fungal density ~ Species + residual error`
Model 2: `Fungal density ~ Species + (Block) + residual error`
Model 3: `Fungal density ~ Species + Soil.P + Soil.N + residual error`

Fig. 2 | Experimental design strategies for optimizing sample size and improving signal:noise ratio. Statistical power depends on both the between-group variance (the signal or effect size) and the within-group variance (the noise). Smaller effect sizes require larger sample sizes to detect, especially when noise is high. **A** Points show trait values of individuals comprising two populations (used in the statistical sense of the word); horizontal lines indicate the true mean trait value for each population. Thus, the distance between horizontal lines is the effect size. The populations could be two different species of yeast; the same species of yeast growing in two experimental conditions; the wild type and mutant genotypes; etc. To estimate the difference between populations, the researcher can only feasibly measure a subset (i.e., sample) of individuals from each population. The yellow boxes report the minimum sample size per group needed to provide an 80% chance of detecting the difference using a *t*-test, as determined using power analysis. **B** Blocking reduces the noise contributed by unmeasured, external factors (e.g., soil quality in a field experiment with the goal of comparing fungal colonization in the roots of two plant species). Soil quality is represented by the background color in each panel. Top: without blocking, soil quality influences the dependent variable for each replicate in an unpredictable way, creating high within-group variance. Middle: spatial arrangement of replicates into two blocks allows estimation of the difference between species while accounting for the fact that trait values are expected to differ between the blocks on average. Bottom: in a paired design, each block contains one replicate from each group, allowing the difference in fungal colonization to be calculated directly for each block and tested through a powerful one-sample or paired *t*-test. **C** Three statistical models that could be used in an ANOVA framework to test for the difference in fungal density between plant species, as illustrated in panel B. Relative to Model 1, the within-group variance for each plant species will be reduced in both Model 2 and Model 3. For Model 2, this is accomplished using blocking; a portion of the variance in fungal density can be attributed to environmental differences between the blocks (although these may be unknown and/or unmeasured) and therefore removed from the within-group variances. For Model 3, it is accomplished by including covariates; a portion of the variance in fungal density can be attributed to the concentrations of N and P in the soil near each plant and therefore removed from the within-group variances. Note that for Model 3, the covariates need to be measured for each experimental unit (i.e., plant), rather than each block; in fact, it is most useful when blocking is not an option.

experimental conditions, they are difficult to avoid entirely. Although some tools are available to cleanse datasets of batch-related patterns[47], the biological effect cannot be disentangled from the batch effect if the two are severely confounded. Therefore, it is always wise to randomize replicates among analytical batches (Fig. 3C).

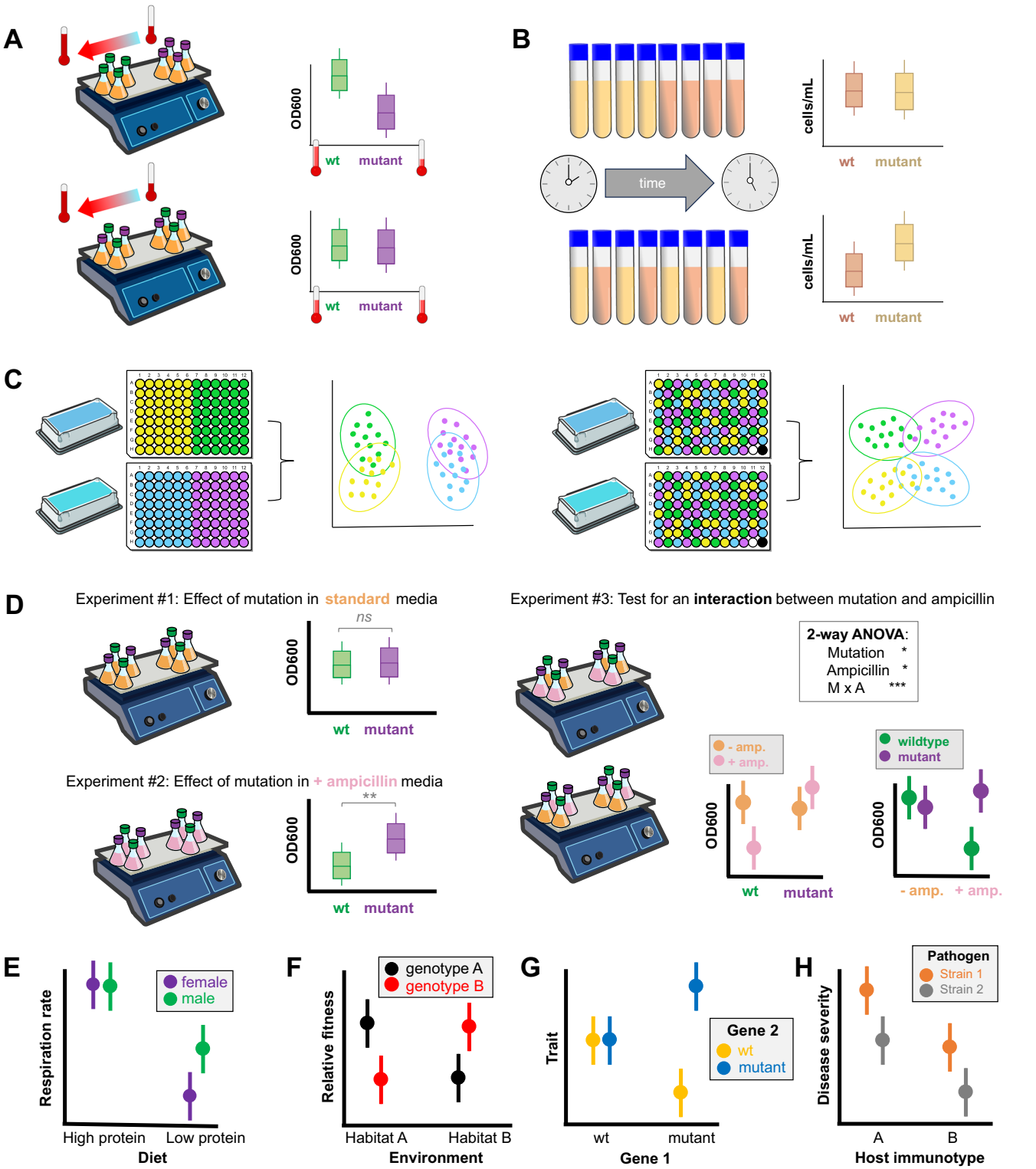
Randomization allows testing for interactions

Finally, randomization is necessary to rigorously test for *interactions* between experimental factors (Fig. 3D–H). An interaction is present if one independent variable moderates the effect of another independent variable on the dependent variable. Examples of interactions include epistasis between genes, temperature influencing the phenotypic effect of a mutation, and host genotypes differing in susceptibility to a pathogen.

Many biologists are interested in such context-dependency, but proper experimental design is crucial for testing interactions rigorously. It can be tempting to conduct a series of simple trials that alter one variable at a time and then compare the results of those trials; however, this approach is invalid for testing interactions[48]. Instead, when multiple independent variables are of interest, the biological replicates must be randomized with respect to all of them, simultaneously (Fig. 3D).

Occasionally full randomization is impossible and the experiment may require a *split-plot* design, where one variable is randomly assigned

https://doi.org/10.1038/s41467-025-62616-x



to groups of replicates rather than individual replicates. For instance, tubes cannot be randomized with respect to temperature within a single incubator; therefore, they must be distributed across multiple incubators, each of which is randomly assigned to a temperature treatment. The main challenge of split-plot designs is that two plots can differ from each other in ways other than the applied treatment, leading to uncertainty about any observed effects of the treatment. In the above scenario, to minimize the possibility that unintentional differences between incubators are the true cause of any observed differences, ideally at least 2 incubators would be used per temperature (either in parallel or in sequential replications of the experiment). As long as the tubes are randomized within incubators with respect to other variables, tests for interactions are valid. However, split-plot designs have less statistical power than fully-randomized designs.

Conclusions

In the fast-paced world of research, taking the time to mindfully plan an experiment (Box 1) may seem like a burden, but it is well worth the effort. Like any lab technique, experimental design becomes faster and less daunting with practice. Furthermore, the techniques of optimizing

Fig. 3 | Randomization of biological replicates across space, time, and batches can reduce experimental bias and reveal interactions between variables. A Top: An undetected temperature gradient within a lab causes a false positive result. The mutant strain grows more slowly than the wild-type on the cooler side of the lab. Bottom: After randomizing the flasks in space, temperature is no longer confounded with genotype and the mutation is revealed to have no effect on growth. **B** Top: A chronological confounding factor causes a false negative result. When cells are counted in all of the rich-media replicates first, the poor-media replicates systematically have more time to grow, masking the treatment effect. Other external variables that can change over time include the lab's temperature or humidity, the organism's age or circadian status, and researcher fatigue. Bottom: Randomizing the order of measurement eliminates the confounding factor, revealing the treatment effect. **C** Left: Batch effects exaggerate the similarity between the yellow and green groups and between the purple and blue groups. Right: Randomization of replicates from all four groups across batches leads to a more accurate measurement of the similarities and differences among the groups. Inclusion of positive and negative controls (black and white) can help to detect batch effects. **D** Randomization is necessary to test for interactions. Left: In hypothetical Experiments 1 and 2, one variable (genotype) is randomized but the other (ampicillin) is not. These observations, separated in time and/or space, cannot be used to conclude that ampicillin influences the effect of the mutation. Right: Both variables (genotype, ampicillin) are randomized and compared within a single experiment. A 2-way ANOVA confirms the interaction and the conclusion that ampicillin influences the mutation's effect on growth is valid. The two plots displaying the interaction are equivalent and interchangeable; the first highlights that the effect of the mutation is apparent only in the presence of ampicillin, while the second highlights that ampicillin inhibits growth only for the wild-type strain. **E**–**H** illustrate other patterns that can only be revealed in a properly randomized experiment. **E** A low-protein diet reduces respiration rate overall, but that effect is stronger for female than male animals. **F** Two plant genotypes show a rank change in relative fitness depending on the environment in which they are grown. **G** Two genes have epistatic effects on a phenotypic trait: a mutation in Gene 1 can either increase or decrease trait values depending on the allele at Gene 2. **H** This plot shows a lack of interaction between the pathogen strain and the host immunotype, as indicated by the (imaginary) parallel line segments that connect pairs of points. In contrast, the line segments that would connect pairs of points would not be parallel if an interaction were present (see **E**–**G**). In **H**, host immunotype A is more susceptible to both pathogen strains than host immunotype B, and pathogen strain 1 causes more severe disease than strain 2 regardless of host. Image credits: vector art courtesy of NIAID (orbital shaker) and Servier Medical Art (reservoir).

replication, signal:noise ratio, proper controls, and randomization are applicable to empirical research in any field. These skills will empower any biologist to conduct and report experiments with confidence. They may be learned through experience (especially the painful experience of realizing too late that an experiment is missing a critical component) and self-guided reading; however, it would be more efficient to instill them early on through formal coursework, thoughtful mentoring, and cultural change.

First, biology training programs—undergraduate, graduate, and beyond—should include at least one mandatory course in statistics, experimental design, or quantitative reasoning if they do not do so already. Courses that emphasize practical applications of statistical thinking, in addition to the underlying theory, are especially valuable to scientists-in-training who are eager to get in the lab and start generating data. Importantly, these courses should not focus solely on data analysis but rather demonstrate the relevance of statistical thought to the decisions that must be made before, during, *and* after an experiment. Cuddington et al.[49] highlight empirically-supported instruction methods that empower biologists to use quantitative skills in their research. Degree-granting programs, academic departments, and professional societies can all promote the improved quality and availability of quantitative training opportunities, including workshops and short courses that may be more accessible to established scientists with less time available for learning[50]. We note that courses focused on coding—another important skill for modern biology that is sometimes neglected in curricula[50]—are not substitutes for statistics and quantitative reasoning courses.

Second, principal investigators and other senior researchers should encourage their mentees to develop and use quantitative reasoning throughout all stages of the research lifecycle, not just during data analysis. For example, they can model this practice by working through the experimental design process (Box 1) with their lab members, especially beginning graduate students. Mentors can also identify gaps in students' quantitative training and steer them toward resources or coursework to fill those gaps[49]. By creating, modeling, and clearly communicating the expectation that their students will engage with statistics, mentors can provide a more well-rounded training experience to young biologists.

Finally, while educators and mentors have unique opportunities to teach experimental design, all biologists have the power to promote better research practices through their day-to-day work. This could take many forms, such as asking mentors, peers, and colleagues for feedback on experimental plans (Box 1) and offering such feedback to others; selecting helpful statistics-related papers (Table 1) for discussion during journal clubs or recommending them to peers; and independently seeking out quantitative training opportunities. In this way, any practicing biologist can contribute to a professional culture where rigorous experimental design is considered as fundamental to good research as proper lab technique.

References

1. Hu, J., Coombes, K. R., Morris, J. S. & Baggerly, K. A. The importance of experimental design in proteomic mass spectrometry experiments: Some cautionary tales. *Brief* **3**, 322–331 (2005)
2. Goeminne, L. J. E., Gevaert, K. & Clement, L. Experimental design and data-analysis in label-free quantitative LC/MS proteomics: a tutorial with MSqRob. *J. Proteom.* **171**, 23–36 (2018).
3. Ching, T., Huang, S. & Garmire, L. X. Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* **20**, 1684–1696 (2014).
4. Schurch, N. J. et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?. *RNA* **22**, 839–851 (2016).
5. Kirpich, A. et al. Variable selection in omics data: a practical evaluation of small sample sizes. *PLoS One* **13**, e0197910 (2018).
6. Zhou, H., He, K., Chen, J. & Zhang, X. LinDA: linear models for differential abundance analysis of microbiome compositional data. *Genome Biol.* **23**, 95 (2022).
7. Willis, A. D. & Clausen, D. S. Planning and describing a microbiome data analysis. *Nat. Microbiol.* **10**, 604–607 (2025).
8. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–2223 (2011).
9. Goeman, J. J. & Bühlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**, 980–987 (2007).
10. Colegrave, N. & Ruxton, G. D. Using biological insight and pragmatism when thinking about pseudoreplication. *Trends Ecol. Evol.* **33**, 28–35 (2018).
11. Kawecki, T. J. et al. Experimental evolution. *Trends Ecol. Evol.* **27**, 547–560 (2012).
12. Van den Bergh, B., Swings, T., Fauvart, M. & Michiels, J. Experimental design, population dynamics, and diversity in microbial experimental evolution. *Microbiol. Mol. Biol. Rev.* **82**, e00008–e00018 (2018).
13. Desai, M. M. Statistical questions in experimental evolution. *J. Stat. Mech. Theory Exp.* **2013**, P01003 (2013).
14. Neyman, J. & Pearson, E. S. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **231**, 289–337 (1933).

15. Neyman, J. & Pearson, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference: part II. *Biometrika* **20A**, 263–294 (1928).
16. Lakens, D. & Caldwell, A. R. Simulation-based power analysis for factorial analysis of variance designs. *Adv. Methods Pract. Psychol. Sci.* **4**, 2515245920951503 (2021).
17. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* **41**, 1149–1160 (2009).
18. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G. Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191 (2007).
19. Champely, S. pwr: Basic Functions for Power Analysis. (2020).
20. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33**, 3486–3488 (2017).
21. By IMPACTT investigators. Beta-diversity distance matrices for microbiome sample size and power calculations — How to obtain good estimates. *Comput. Struct. Biotechnol. J.* **20**, 2259–2267 (2022).
22. Kohler, D. et al. MSstats version 4.0: statistical analyses of quantitative mass spectrometry-based proteomic experiments with chromatography-based quantification at scale. *J. Proteome Res.* **22**, 1466–1482 (2023).
23. Bi, R. & Liu, P. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinforma.* **17**, 146 (2016).
24. Lin, C.-W. et al. RNASeqDesign: a framework for ribonucleic acid sequencing genomewide power calculation and study design issues. *J. R. Stat. Soc. C Appl. Stat.* **68**, 683–704 (2019).
25. La Rosa, P. S. et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS One* **7**, e52078 (2012).
26. Kelly, B. J. et al. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. *Bioinformatics* **31**, 2461–2468 (2015).
27. Casals-Pascual, C. et al. Microbial diversity in clinical microbiome studies: sample size and statistical power considerations. *Gastroenterology* **158**, 1524–1528 (2020).
28. Ferdous, T. et al. The rise to power of the microbiome: power and sample size calculation for microbiome studies. *Mucosal Immunol.* **15**, 1060–1070 (2022).
29. Laukens, D., Brinkman, B. M., Raes, J., De Vos, M. & Vandenabeele, P. Heterogeneity of the gut microbiome in mice: guidelines for optimizing experimental design. *FEMS Microbiol. Rev.* **40**, 117–132 (2016).
30. Shansky, R. M. & Murphy, A. Z. Considering sex as a biological variable will require a global shift in science culture. *Nat. Neurosci.* **24**, 457–464 (2021).
31. Witjes, V. M., Boleij, A. & Halffman, W. Reducing versus embracing variation as strategies for reproducibility: the microbiome of laboratory mice. *Animals* **10**, 2415 (2020).
32. Bergelson, J., Kreitman, M., Petrov, D. A., Sanchez, A. & Tikhonov, M. Functional biology in its natural context: A search for emergent simplicity. *eLife* **10**, e67646 (2021).
33. Klironomos, J. N., Rillig, M. C. & Allen, M. F. Designing belowground field experiments with the help of semi-variance and power analyses. *Appl. Soil Ecol.* **12**, 227–238 (1999).
34. Berry, J. C. et al. Increased signal-to-noise ratios within experimental field trials by regressing spatially distributed soil properties as principal components. *eLife* **11**, e70056 (2022).
35. Falony, G. et al. Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
36. Kendziorski, C., Irizarry, R. A., Chen, K.-S., Haag, J. D. & Gould, M. N. On the utility of pooling biological samples in microarray experiments. *Proc. Natl Acad. Sci.* **102**, 4252–4257 (2005).
37. Todd, E. V., Black, M. A. & Gemmell, N. J. The power and promise of RNA-seq in ecology and evolution. *Mol. Ecol.* **25**, 1224–1241 (2016).
38. Colovas, J., Bintarti, A. F., Mechan Llontop, M. E., Grady, K. L. & Shade, A. Do-it-yourself mock community standard for multi-step assessment of microbiome protocols. *Curr. Protoc.* **2**, e533 (2022).
39. Chen, K. et al. The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. *Mol. Cell. Biol.* **36**, 662–667 (2016).
40. Hutchinson, M. I., Bell, T. A. S., Gallegos-Graves, L. V., Dunbar, J. & Albright, M. Merging fungal and bacterial community profiles via an internal control. *Microb. Ecol.* **82**, 484–497 (2021).
41. Fierer, N. et al. Guidelines for preventing and reporting contamination in low-biomass microbiome studies. *Nat. Microbiol.* **10**, 1570–1580 (2025).
42. Kim, D. et al. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* **5**, 52 (2017).
43. Leek, J. T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
44. Goh, W. W. B., Wang, W. & Wong, L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* **35**, 498–507 (2017).
45. Han, W. & Li, L. Evaluating and minimizing batch effects in metabolomics. *Mass Spectrom. Rev.* **41**, 421–442 (2022).
46. Manga, P. et al. Replicates, read numbers, and other important experimental design considerations for microbial RNA-seq identified using bacillus thuringiensis datasets. *Front. Microbiol.* **7**, 794 (2016).
47. Zhou, L., Chi-Hau Sue, A. & Bin Goh, W. W. Examining the practical limits of batch effect-correction algorithms: when should you care about batch effects?. *J. Genet. Genom.* **46**, 433–443 (2019).
48. Gelman, A. & Stern, H. The difference between "significant" and "not significant" is not itself statistically significant. *Am. Stat.* **60**, 328–331 (2006).
49. Cuddington, K. et al. Challenges and opportunities to build quantitative self-confidence in biologists. *BioScience* **73**, 364–375 (2023).
50. Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A. & Schneider, M. V. A global perspective on evolving bioinformatics and data science training needs. *Brief. Bioinform.* **20**, 398–404 (2019).
51. Paulson, D. S. *Biostatistics and Microbiology: A Survival Manual* (Springer Science & Business Media, 2008).
52. Whitlock, M. C. & Schluter, D. *The Analysis of Biological Data* (Macmillan Higher Education, 2019).
53. Ruxton, G. D. & Colegrave, N. *Experimental Design for the Life Sciences*. (Oxford University Press, Oxford, New York, 2017).
54. Dean, A., Voss, D. & Draguljić, D. *Design and Analysis of Experiments*. (Springer International Publishing, 2017). https://doi.org/10.1007/978-3-319-52250-0.
55. Hinkelmann, K. & Kempthorne, O. *Principles of Experimental Design* (John Wiley & Sons, Ltd, 2007).
56. Casler, M. D. Fundamentals of experimental design: guidelines for designing successful experiments. *Agron. J.* **107**, 692–705 (2015).
57. Oberg, A. L. & Vitek, O. Statistical design of quantitative mass spectrometry-based proteomic experiments. *J. Proteome Res.* **8**, 2144–2156 (2009).
58. Rasch, D., Pilz, J., Verdooren, L. R. & Gebhardt, A. *Optimal Experimental Design with R* (Chapman and Hall/CRC, 2011). https://doi.org/10.1201/b10934.
59. Chen, Z.-M. et al. Greater enhancement of Bacillus subtilis spore yields in submerged cultures by optimization of medium composition through statistical experimental designs. *Appl. Microbiol. Biotechnol.* **85**, 1353–1360 (2010).
60. Groemping, U. & Morgan-Wall, T. CRAN task view: design of experiments (DoE) & analysis of experimental data. https://CRAN.R-project.org/view=ExperimentalDesign (2025).

61. Antony, J. *Design of Experiments for Engineers and Scientists* (Elsevier, 2023).
62. Qin, L.-X. et al. Blocking and randomization to improve molecular biomarker discovery. *Clin. Cancer Res.* **20**, 3371–3378 (2014).
63. Gilman, J., Walls, L., Bandiera, L. & Menolascina, F. Statistical design of experiments for synthetic biology. *ACS Synth. Biol.* **10**, 1–18 (2021).
64. Casler, M. D. Blocking principles for biological experiments. in *Applied Statistics in Agricultural, Biological, and Environmental Sciences* 53–72 (John Wiley & Sons, Ltd, 2018). https://doi.org/10.2134/appliedstatistics.2015.0074.c3
65. Borges, A. et al. Can spatial modeling substitute for experimental design in agricultural experiments?. *Crop Sci.* **59**, 44–53 (2019).
66. Lazic, S. E., Clarke-Williams, C. J. & Munafò, M. R. What exactly is 'N' in cell culture and animal experiments?. *PLoS Biol.* **16**, e2005282 (2018).
67. Verma, J. P. *Repeated Measures Design for Empirical Researcher* (John Wiley & Sons, 2015).
68. Guang, A. Power analyses for microbiome studies with micropower. *Comput. Biol. Core* https://medium.com/brown-compbiocore/power-analyses-for-microbiome-studies-with-micropower-8ff28b36dfe3 (2020).
69. Schmid, K. T. et al. scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. *Nat. Commun.* **12**, 6625 (2021).
70. Anderson, M. J., de Valpine, P., Punnett, A. & Miller, A. E. A pathway for multivariate analysis of ecological communities using copulas. *Ecol. Evol.* **9**, 3276–3294 (2019).
71. Sze, M. A. & Schloss, P. D. Looking for a signal in the noise: revisiting obesity and the microbiome. *mBio* **7**, https://doi.org/10.1128/mbio.01018-16 (2016).
72. Koren, O. et al. Host remodeling of the gut microbiome and metabolic changes during pregnancy. *Cell* **150**, 470–480 (2012).
73. Walter, J., Armet, A. M., Finlay, B. B. & Shanahan, F. Establishing or exaggerating causality for the gut microbiome: lessons from human microbiota-associated rodents. *Cell* **180**, 221–232 (2020).
74. Zhao, F. et al. A short-term feeding of dietary casein increases abundance of lactococcus lactis and upregulates gene expression involving obesity prevention in cecum of young rats compared with dietary chicken protein. *Front. Microbiol.* **10**, 2411 (2019).
75. Bartlett, A., Blakeley-Ruiz, J. A., Richie, T., Theriot, C. M. & Kleiner, M. Large quantities of bacterial DNA and protein in common dietary protein source used in microbiome studies. *Proteomics* e202400149, https://doi.org/10.1002/pmic.202400149 (2025).
76. Dollive, S. et al. Fungi of the murine gut: episodic variation and proliferation during antibiotic treatment. *PLoS One* **8**, e71806 (2013).
77. Johnston-Monje, D., Lundberg, D. S., Lazarovits, G., Reis, V. M. & Raizada, M. N. Bacterial populations in juvenile maize rhizospheres originate from both seed and soil. *Plant Soil* **405**, 337–355 (2016).
78. Salter, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
79. Olomu, I. N. et al. Elimination of "kitome" and "splashome" contamination results in lack of detection of a unique placental microbiome. *BMC Microbiol.* **20**, 157 (2020).
80. The importance of no evidence *Nat. Hum. Behav.* **3**, 197–197 (2019).
81. Nuzzo, R. Scientific method: Statistical errors. *Nature* **506**, 150–152 (2014).
82. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychol. Bull.* **86**, 638–641 (1979).
83. Sterling, T. D., Rosenbaum, W. L. & Weinkam, J. J. Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. *Am. Stat.* **49**, 108–112 (1995).
84. Nissen, S. B., Magidson, T., Gross, K. & Bergstrom, C. T. Publication bias and the canonization of false facts. *eLife* **5**, e21451 (2016).
85. Murad, M. H., Chu, H., Lin, L. & Wang, Z. The effect of publication bias magnitude and direction on the certainty in evidence. *BMJ Evid.-Based Med.* **23**, 84–86 (2018).
86. Kers, J. G. & Saccenti, E. The power of microbiome studies: some considerations on which alpha and beta metrics to use and how to report results. *Front. Microbiol.* **12**, 796025 (2022).
87. Chalmers, I. & Glasziou, P. Avoidable waste in the production and reporting of research evidence. *Lancet* **374**, 86–89 (2009).
88. Purgar, M., Klanjscek, T. & Culina, A. Quantifying research waste in ecology. *Nat. Ecol. Evol.* **6**, 1390–1397 (2022).
89. Soderberg, C. K. et al. Initial evidence of research quality of registered reports compared with the standard publishing model. *Nat. Hum. Behav.* **5**, 990–997 (2021).

Acknowledgements

Many thanks to Caetano Antunes, Josie Chandler, Natalie Ford, Nichole Ginnan, Zach Harris, and Joel Swift for helpful feedback on early drafts of the manuscript. This work was supported by awards from the National Science Foundation (IOS-2016351 to MRW, IOS-2033621 to MRW and MK), the USDA National Institute for Food and Agriculture (2022-67013-36672 to MK and MRW), and the National Institute of General Medical Sciences of the National Institutes of Health (R35GM138362 to MK).

Author contributions

M.R.W. wrote the initial draft; M.K. contributed substantially to advanced drafts.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-62616-x.

Correspondence and requests for materials should be addressed to Maggie R. Wagner.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.