POINTS OF SIGNIFICANCE

Replication

Quality is often more important than quantity.

Science relies heavily on replicate measurements. Additional replicates generally yield more accurate and reliable summary statistics in experimental work. But the straightforward question, 'how many and what kind of replicates should I run?' belies a deep set of distinctions and tradeoffs that affect statistical testing. We illustrate different types of replication in multilevel ('nested') experimental designs and clarify basic concepts of efficient allocation of replicates.

Replicates can be used to assess and isolate sources of variation in measurements and limit the effect of spurious variation on hypothesis testing and parameter estimation. Biological replicates are parallel measurements of biologically distinct samples that capture random biological variation, which may itself be a subject of study or a noise source. Technical replicates are repeated measurements of the same sample that represent independent measures of the random noise associated with protocols or equipment. For biologically distinct conditions, averaging technical replicates can limit the impact of measurement error, but taking additional biological replicates is often preferable for improving the efficiency of statistical testing.

Nested study designs can be quite complex and include many levels of biological and technical replication (**Table 1**). The distinction between biological and technical replicates depends on which sources of variation are being studied or, alternatively, viewed as noise sources.

An illustrative example is genome sequencing, where base calls (a statistical estimate of the most likely base at a given sequence position) are made from multiple DNA reads of the same genetic locus. These reads are technical replicates that sample the uncertainty in the sequencer readout but will never reveal errors present in the library itself. Errors in library construction can be mitigated by constructing technical replicate libraries from the same sample. If additional resources are available, one could potentially return to the source tissue and collect multiple samples to repeat the entire sequencing workflow. Such replicates would be technical if the samples were considered to be from the same aliquot or biological if considered to be from different aliquots of biologically distinct material[1]. Owing to historically high costs per assay, the field of genome sequencing has not demanded such replication. As the need for accuracy increases and the cost of sequencing falls, this is likely to change.

Table 1 | Replicate hierarchy in a hypothetical mouse single-cell gene expression RNA sequencing experiment

	Replicate type	Replicate category[a]
Animal study subjects	Colonies	B
	Strains	B
	Cohoused groups	B
	Gender	B
	Individuals	B
Sample preparation	Organs from sacrificed animals	B
	Methods for dissociating cells from tissue	T
	Dissociation runs from given tissue sample	T
	Individual cells	B
	RNA-seq library construction	T
Sequencing	Runs from the library of a given cell	T
	Reads from different transcript molecules	V[b]
	Reads with unique molecular identifier (UMI) from a given transcript molecule	T

[a]Replicates are categorized as biological (B), technical (T) or of variable type (V). [b]Sequence reads serve diverse purposes depending on the application and how reads are used in analysis.

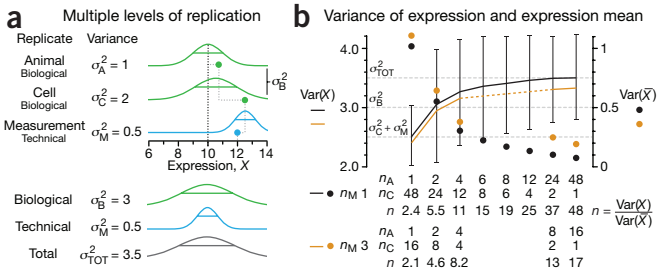

Figure 1 | Replicates do not contribute equally and independently to the measured variability, which can often underestimate the total variability in the system. (**a**) Three levels of replication (two biological, one technical) with animal, cell and measurement replicates normally distributed with a mean across animals of 10 and ratio of variances 1:2:0.5. Solid green (biological) and blue (technical) dots show how a measurement of the expression (X = 12) samples from all three sources of variation. Distribution s.d. is shown as horizontal lines. (**b**) Expression variance, Var(X), and variance of expression mean, Var($\bar{X}$), computed across 10,000 simulations of $n_A n_C n_M$ = 48 measurements for unique combinations of the number of animals (n_A = 1 to 48), cells per animal (n_C = 1 to 48) and technical replicate measurements per cell (n_M = 1 and 3). The ratio of Var(X) and Var($\bar{X}$) is the effective sample size, n, which corresponds to the equivalent number of statistically independent measurements. Horizontal dashed lines correspond to biological and total variation. Error bars on Var(X) show s.d. from the 10,000 simulated samples (n_M = 1).

How does one determine the types, levels and number of replicates to include in a study, and the extent to which they contribute information about important sources of variation? We illustrate the approach to answering these questions with a single-cell sequencing scenario in which we measure the expression of a specific gene in liver cells in mice. We simulated three levels of replication: animals, cells and measurements (**Fig. 1a**). Each level has a different variance, with animals ($\sigma_A^2 = 1$) and cells ($\sigma_C^2 = 2$) contributing to a total biological variance of $\sigma_B^2 = 3$. When technical variance from the assay ($\sigma_M^2 = 0.5$) is included, these distributions compound the uncertainty in the measurement for a total variance of $\sigma_{TOT}^2 = 3.5$. We next simulated 48 measurements, allocated variously between biological replicates (the number of animals, n_A and number of cells sampled per animal, n_C) and technical replicates (number of measurements taken per cell, n_M) for a total number of measurements $n_A n_C n_M = 48$. Although we will always make 48 measurements, the effective sample size, n, will vary from about 2 to 48, depending on how the measurements are allocated. Let us look at how this comes about.

Our ability to make accurate inferences will depend on our estimate of the variance in the system, Var(X). Different choices of n_A, n_C and n_M impact this value differently. If we sample n_C = 48 cells from a single animal (n_A = 1) and measure each n_M = 1 times, our estimate of the total variance σ_{TOT}^2 will be Var(X) = 2.5 (**Fig. 1b**). This reflects cell and measurement variances ($\sigma_C^2 + \sigma_M^2$) but not animal variation; with only one animal sampled we have no way of knowing what the animal variance is. Thus Var(X) certainly underestimates σ_{TOT}^2, but we would not know by

how much. Moreover, the uncertainty in Var(X) (error bar at $n_A = 1$; **Fig. 1b**) is the error in $\sigma_C^2 + \sigma_M^2$ and not σ_{TOT}^2. At another extreme, if all our measurements are technical replicates ($n_A = n_C = 1$, $n_M = 48$) we would find Var(X) = 0.5 (not represented in **Fig. 1**). This is only the technical variance; if we misinterpreted this as biological variation and used it for biological inference, we would have an excess of false positives. Be on the lookout: unusually small error bars on biological measurements may merely reflect measurement error, not biological variation. To obtain the best estimate of σ_{TOT}^2 we should sample $n_C = 1$ cells from $n_A = 48$ animals because each of the 48 measurements will independently sample each of the distributions in **Figure 1a**.

Our choice of the number of replicates also influences Var($\bar{X}$), the precision in the expression mean. The optimal way to minimize this value is to collect data from as many animals as possible ($n_A = 48$, $n_C = n_M = 1$), regardless of the ratios of variances in the system. This comes from the fact that n_A contributes to decreasing each contribution to Var($\bar{X}$), which is given by $\sigma_A^2/n_A + \sigma_C^2/n_A n_C + \sigma_M^2/n_A n_C n_M$. Although technical replicates allow us to determine σ_M^2, unless this is a quantity of interest, we should omit technical replicates and maximize n_A. Of course, good blocking practice suggests that samples from the different animals and cells should be mixed across the sequencing runs to minimize the effect of any systematic run-to-run variability (not present in simulated data here).

The value in additional measurements can be estimated by the prospective improvement in effective sample size. We have seen before that the variance in the mean of a random variable is related to its variance by Var(X) = nVar($\bar{X}$). The ratio of Var(X) to Var($\bar{X}$) can therefore be used as a measure of the equivalent number of independent samples. From **Figure 1b**, we can see that n = 48 only for n_A = 48 and drops to n = 25 for n_A, n_C = 12, 4 and is as low as about 2 for n_A = 1. In other words, even though we may be collecting additional measurements they do not all contribute equally to an increase in the precision of the mean. This is because additional cell and technical replicates do not correspond to statistically independent values: technical replicates are derived from the same cell and the cell replicates from the same animal. If it is necessary to summarize expression variability at the level of the animals, then cells from a given animal are pseudoreplicates—statistically correlated in a way that is unique to that animal and not representative of the population under study. Not all replicates yield statistically independent measures, and treating them as if they do can erroneously lower the apparent uncertainty of a result.

The number of replicates has a practical effect on inference errors in analysis of differences of means or variances. We illustrate this by enumerating inference errors in 10,000 simulated drug-treatment experiments in which we vary the number of animals and cells (**Fig. 2**). We assume a 10% effect chance for two scenarios: a twofold increase in variance, σ_C^2, or a 10% increase in mean, μ_A, using the same values for other variances and 48 total measurements as in **Figure 1**. Applying the t-test, we show false discovery rate (FDR) and power for detecting these differences (**Fig. 2**). If we want to detect a difference in variation across cells, it is best to choose $n_A \approx n_C$ in our range. On the other hand, when we are interested in changes in mean expression across mice, it is better to sample as many mice as possible. In either case, increasing the number of measurements from 48 to 144 by taking three technical replicates (n_M = 3) improves inference only slightly.

Biological replicates are preferable to technical replicates for inference about the mean and variance of a biological population.

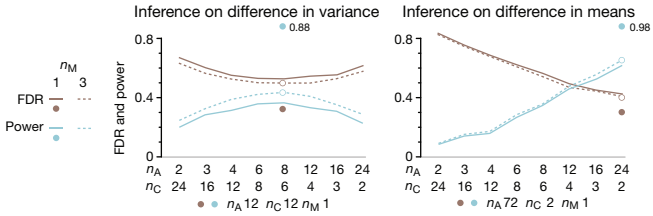

Figure 2 | The number of replicates affects FDR and power of inferences on the difference in variances and means. Shown are power and FDR profiles of a test of difference in cell variances (left) and animal means (right) for 48 (n_M = 1) or 144 (n_M = 3) measurements using different combinations of n_A and n_C. Vertical arrows indicate change in FDR and power when technical replicates are replaced by biological replicates, as shown by n_A,n_C,n_M, for the same number of measurements (144). Values generated from 10,000 simulations of a 10% chance of a treatment effect that increases cell variance $2\sigma_C^2$ or animal mean 1.1 × μ_A. Samples were tested with two-sample t-test (sample size n_A) at two-tailed α = 0.05.

(**Fig. 2**). For example, changing n_A,n_C,n_M from 8,6,3 (where power is highest) to 12,12,1 doubles the power (0.43 to 0.88) in detecting a twofold change in variance. In the case of detecting a 10% difference in means, changing n_A,n_C,n_M from 24,2,3 to 72,2,1 increases power by about 50% from 0.66 to 0.98. Practically, the cost difference between biological and technical replicates should be considered; this will affect the cost-benefit tradeoff of collecting additional replicates of one type versus the other. For example, if the cost units of animals to cells to measurements is 10:1:0.1 (biological replicates are likely more expensive than technical ones) then an experiment with n_A,n_C,n_M of 12,12,1 is about twice as expensive as that with 8,6,3 (278 versus 142 cost units). However, power in detecting a change in variance is doubled as well, so the cost increase is commensurate with increase in efficiency. In the case of detecting differences in means, 72,2,1 is about three times as expensive as 24,2,3 (878 versus 302 cost units) but increases power only by 50%, making this a lower-value proposition.

Typically, biological variability is substantially greater than technical variability, so it is to our advantage to commit resources to sampling biologically relevant variables unless measures of technical variability are themselves of interest, in which case increasing the number of measurements per cell, n_M, is valuable.

Good experimental design practice includes planning for replication. First, identify the questions the experiment aims to answer. Next, determine the proportion of variability induced by each step to distribute the capacity for replication of the experiment across steps. Be aware of the potential for pseudoreplication and aim to design statistically independent replicates.

As our capacity for higher-throughput assays increases, we should not be misled into thinking that more is always better. Clear thinking about experimental questions and sources of variability is still crucial to produce efficient study designs and valid statistical analyses.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Paul Blainey, Martin Krzywinski & Naomi Altman

1. Robasky, K., Lewis, N.E. & Church, G.M. *Nat. Rev. Genet.* **15**, 56–62 (2014).

Paul Blainey is an Assistant Professor of Biological Engineering at MIT and Core Member of the Broad Institute. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University.