POINTS OF SIGNIFICANCE

Nested designs

For studies with hierarchical noise sources, use a nested analysis of variance approach.

Many studies are affected by random-noise sources that naturally fall into a hierarchy, such as the biological variation among animals, tissues and cells, or technical variation such as measurement error. With a nested approach, the variation introduced at each hierarchy layer is assessed relative to the layer below it. We can use the relative noise contribution of each layer to optimally allocate experimental resources using nested analysis of variance (ANOVA), which generally addresses replication and blocking, previously discussed *ad hoc*[1,2].

Recall that factors are independent variables whose values we control and wish to study[3] and which have systematic effects on the response. Noise limits our ability to detect effects, but known noise sources (e.g., cell culture) can be mitigated if used as blocking factors[2]. We can model the contribution of each blocking factor to the overall variability, isolate it and increase power[2]. Statisticians distinguish between fixed factors, typically treatments, and random factors, such as blocks.

The impact of fixed and random factors in the presence of experimental error is shown in **Figure 1**. For a fixed factor (**Fig. 1a**), each of its levels (for example, a specific drug) has the same effect in all experiments and an unmodeled uncertainty due to experimental error. The levels of a fixed factor can be exactly duplicated (level A1 in **Fig. 1a** is identical for each experiment) and are of specific interest, usually the effect on the population mean.

In contrast, when we repeat an experiment, the levels of a random factor are sampled from a population of all possible levels of the factor (replicates) and are different across all the experiments, emphasized by unique level labels (B1–B9; **Fig. 1b**). Because the levels cannot be exactly duplicated, their effect is random and they are not of specific interest. Instead, we use the sample of levels to model the uncertainty added by the random factor (for example, all mice).

Fixed and random factors may be crossed or nested (**Fig. 2**). When crossed, all combinations of factors are used to study the main effects and interactions of two or more factors (**Fig. 2a**). In contrast, nested designs apply a hierarchy—some level combinations are not studied because the levels cannot be duplicated or reused (**Fig. 2b**). Random factors (for example, mouse and cell) are nested within the fixed factor (drug) to measure noise due to individual mice and cells and to generalize the effects of the fixed

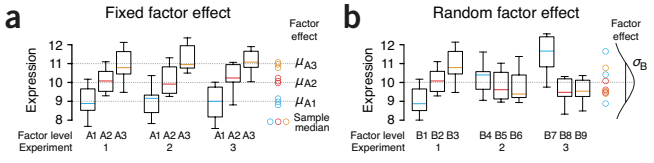

Figure 1 | Inferences about fixed factors are different than those about random factors, as shown by box-plots of $n = 10$ samples across three independent experiments. Circles indicate sample medians. Box-plot height reflects simulated measurement error ($\sigma_\varepsilon^2 = 0.5$). (**a**) Fixed factor levels are identical across experiments and have a systematic effect on the mean. (**b**) Random factor levels are samples from a population, have a random effect on the mean and contribute noise to the system ($\sigma_B^2 = 1$).

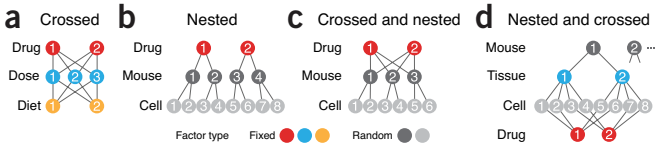

Figure 2 | Factors may be crossed or nested. (**a**) A crossed design examines every combination of levels for each fixed factor. (**b**) Nested design can progressively subreplicate a fixed factor with nested levels of a random factor that are unique to the level within which they are nested. (**c**) If a random factor can be reused for different levels of the treatment, it can be crossed with the treatment and modeled as a block. (**d**) A split plot design in which the fixed effects (tissue, drug) are crossed (each combination of tissue and drug are tested) but themselves nested within replicates.

factor on all mice and cells. If mice can be reused, we can cross them with the drug and use them as a random blocking factor[2] (**Fig. 2c**).

We will use the design in **Figure 2b** to illustrate the analysis of nested fixed and random factors using nested ANOVA, similar to the ANOVA discussed previously[2]. Now nesting is taken into account and the calculations have different interpretations because some of the factors are random. The fixed factor may have an effect on the mean, and the two random factors will add uncertainty. We will be able to estimate the amount of variance for each random factor and use it to better plan our replication strategy. We can maximize power (for example, within cost constraints) to detect a difference in means due to the top-level fixed factor or to detect variability due to random factors. The latter is biologically interesting when increased variance in cell response may be due to increased heterogeneity in the genotypes and implicated in drug resistance.

We will simulate the nested design in **Figure 2b** using three factors: A ($a = 2$ levels: control and treatment), B (mice, $b = 5$ levels, $\sigma_B^2 = 1$), C (cells, $c = 5$ levels, $\sigma_C^2 = 2$). Expression for each cell will be measured using three technical replicates ($\sigma_\varepsilon^2 = 0.5$, $n = 3$). The raw sample data of the simulation are shown in **Figure 3a**.

Nested ANOVA calculations begin with the sum of squared deviations (SS) to partition the variance among the factors, exactly as in regular ANOVA. For example, the first blue arrow in **Figure 3a** represents the difference between the averages of all points from mouse B4 (X_{14}..) and all points from the control (X_{1}...). Factor C has the largest deviations (**Fig. 3b**) because it was modeled to be the largest source of noise ($\sigma_C^2 = 2$). The distinction between regular and nested ANOVA is how the mean squares (MS) enter into the calculation of the *F*-ratio for each factor. The *F*-ratio is a ratio of MS values, and the denominator corresponds to the MS of the next nested factor (for example, MS_B/MS_C) and not MS_E (see **Supplementary Table 1** for nested ANOVA formulas and calculated values; see **Supplementary Table 2** for expected values of MS). The *F*-test uses the ratio of between-group sample variance (estimate of population variance from sample means) and within-group variance (estimate of population variance from sample variances) to test whether group means differ (for fixed factors). In the case of random factors, the interpretation is whether the factor contributes noise in addition to the noise due to the factor nested within it (for example, is there more mouse-to-mouse variability than would be expected from cell-to-cell variability?).

At the bottom of the nested hierarchy ($n = 3$ technical replicates per cell), we find $MS_E = 0.55$, which is an estimate of $\sigma_\varepsilon^2 = 0.5$ in our simulation. We find statistically significant (at $\alpha = 0.05$) contributions to noise from both mice (factor B) and cells (factor C) with estimated variance contributions of 0.84 and 2.1, respectively, which matches

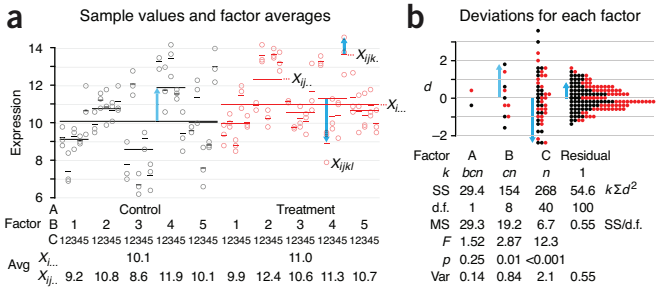

Figure 3 | Data and analysis for a simulated three-factor nested experiment. (**a**) Simulated expression levels, X_{ijkl}, measured for $a = 2$ levels of factor A (control and treatment, i), $b = 5$ of factor B (mice, j), $c = 5$ of factor C (cells, k) and $n = 3$ technical replicates (l). Averages across factor levels are shown as horizontal lines and denoted by dots in subscript for the factor's index. Blue arrows illustrate deviations used for calculation of sum of squares (SS). Data are simulated with $\mu_c = 10$ for control and $\mu_t = 11$ for treatment and $\sigma_B^2 = 1$, $\sigma_C^2 = 2$, $\sigma_\varepsilon^2 = 0.5$ for noise at mouse, cell and technical replicate levels, respectively. Values below the figure show factor levels and averages at levels of A ($X_{i...}$) and B ($X_{ij..}$). Labels for the levels of B and C are reused but represent distinct individual mice and cells. (**b**) Histogram of deviations (d) for each factor. Three deviations illustrated in **a** are identified by the same blue arrows. Nested ANOVA calculations show number of times (k) each deviation (d) contributes to SS, degrees of freedom (d.f.), mean squares (MS), F-ratio, P value and the estimated variance contribution of each factor.

our inputs $\sigma_B^2 = 1$ and $\sigma_C^2 = 2$. Because the top-layer factor is fixed and not considered a source of noise, its variance component is not a useful quantity—of interest is its effect on the mean. Unfortunately, we were unable to detect a difference in means for A ($P = 0.25$) because of poor power due to our allocation of replicates. It is useful to relate the F-test for factor A to a two-sample t-test to understand the statistical quantities involved and calculate power.

The F-test for the top-layer factor A ($F = \mathrm{MS}_A/\mathrm{MS}_B$) tests the difference between the variances of treatment and mouse means. Any treatment effect on the mean will show up as additional variance, which we stand a chance to detect. Because we have only two levels of factor A, the F-test, which has degrees of freedom (d.f.) of $a - 1 = 1$ and $a(b - 1) = 8$, is equivalent to the two-sample t-test for samples of size b, $2(b - 1)$ d.f. and with $t = \sqrt{F}$. This t-test is applied to the control and treatment samples formed using $b = 5$ averages X_{ij}.. (**Fig. 3a**) whose expected variance is $\mathrm{E}[\mathrm{Var}(X_{ij}..)] = \sigma_B^2 + \sigma_C^2/c + \sigma_\varepsilon^2/(cn) = 1.43$ (ref. 1). This quantity is estimated by $\mathrm{MS}_B/(cn) = 1.28$, which is exactly the average variance of the two sample variances 1.73 and 0.83 (**Supplementary Table 3**). These samples yield the control and treatment means of 10.1 and 11.0 (X_{i}...; **Fig. 3a**) and a t-statistic of $0.9/\sqrt{(2\mathrm{MS}_B/(bcn))} = 1.24$, which yields the same P value of 0.25 as from the F-test.

We can now calculate the t-test power for our scenario. For a difference in means of $d = 1$, the power using samples of size $b = 5$ is 0.21, using the expected variance 1.43. In practice, we might run a trial experiment to determine this value using $\mathrm{MS}_B/(cn)$. Clearly, our initial choice of b, c and n was an inadequate design—we should aim for a power of at least 0.8. If variance is kept at 1.43 ($c = 5$, $n = 5$), this power can be achieved for a sample size $b = 24$. With 24 mice, the expected variance of the average across mice would be $\mathrm{E}[\mathrm{Var}(X_{i}...)] = 1.43/24$. Dividing this into the total variance due to replication ($\sigma_B^2 + \sigma_C^2 + \sigma_\varepsilon^2 = 3.5$), we can calculate the effective sample size, 57 (ref. 1). As we've previously seen, this can be achieved with the fewest number of measurements if we have $b = 57$ mice and $c = n = 1$. If we assume the cost of mice, cells and technical replicates to be 100, 10 and 1, respectively, these designs would cost 3,960 ($b = 24$, $c = 5$, $n = 3$) and 6,327 ($b = 57$, $c = 1$, $n = 1$). Let's see if we can use fewer mice and increase replication to obtain the same power at a lower cost.

The nested analysis provides a general framework for these cost and power calculations. The optimum number of replicates at each level can be calculated on the basis of the cost of replication and the variance at the level of the factor. We want to minimize $\mathrm{Var}(X_{i}...) = \sigma_B^2/b + \sigma_C^2/(bc) + \sigma_\varepsilon^2/(bcn)$ within the cost constraint $K = bC_B + bcC_D + bcnC_D$ (C_X is cost per replicate at factor X) with the goal of finding values of b, c and n that provide the largest decrease in the variance per unit cost. The optimum number of technical replicates is $n^2 = C_C/C_D \times \sigma_\varepsilon^2/\sigma_C^2$. In other words, subreplicates are preferred to replicates when they are cheaper and their factor is a source of greater noise. With the costs as given above ($C_C/C_N = 10$) we find $n^2 = 10 \times 0.5/2 = 2.5$ and $n = 2$. We can apply the same equation for the number of cells, $c^2 = C_B/C_C \times \sigma_C^2/\sigma_B^2$, where C_B is the cost of a mouse. Using the same tenfold cost ratio, $c^2 = 10 \times 2/1 = 20$ and $c = 5$. For $c = 5$ and $n = 2$, $\mathrm{Var}(X_{ij}..)$ is 1.45, and we would reach a power of 0.8 if we had $b = 24$ mice. This experiment is slightly cheaper than the one with $n = 3$ (3,840 vs. 3,960).

Two components affect power in detecting differences in means. Subreplication at the cell and technical layer helps increase power by decreasing the variance of mouse averages, $\mathrm{Var}(X_{ij}..)$, used for t-test samples. The number of mice also increases power because it decreases the standard error of $X_{ij}..$ (the precision of mouse averages) because sample size is increased. To obtain the largest power to detect a treatment effect with the fewest number of measurements, it is always best to pick as many mice as possible: effective sample size is largest and variance of sample averages is lowest.

The number of replicates also affects our ability to detect the noise contribution from each random factor. If detecting and estimating variability in mice and cells is of interest, we should aim to increase the power of the associated F-tests (**Supplementary Table 1**). For example, under the alternative hypothesis of a nonzero contribution of cells to noise (σ_C^2), the F-statistic will be distributed as a multiple of the null hypothesis F-statistic, $F_{u,v} \times (n\sigma_C^2 + \sigma_\varepsilon^2)/\sigma_\varepsilon^2$. The multiplication factor is the ratio of expected MS values (**Supplementary Table 2**). For our simulation values, the multiple is 13 and the d.f. are $u = 40$ and $v = 100$. The critical F-value is 1.52, and our power is the P value for 1.52/13, which is essentially 1 (this is why the P value for factor C in **Fig. 2b** is very low). For level B we have $u = 8$, $v = 40$, a multiple of 3.3 (21.5/6.5) and a power of 0.72. The power of our design to detect noise within mice and cells was much higher than that for detecting an effect of the treatment on the means.

Nested designs are useful for understanding sources of variability in the hierarchy of the subsamples and can reduce the cost of the experiment when costs vary across the hierarchy. Statistical conclusions can be made only about the layers actually replicated—technical replication cannot replace biological replication for biological inference.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.3137).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski, Naomi Altman & Paul Blainey

1. Blainey, P., Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 879–880 (2014).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 699–700 (2014).
3. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 597–598 (2014).

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. Paul Blainey is an Assistant Professor of Biological Engineering at MIT and Core Member of the Broad Institute.