

POINTS OF SIGNIFICANCE

Comparing samples—part II

When a large number of tests are performed, P values must be interpreted differently.

It is surprising when your best friend wins the lottery but not when a random person in New York City wins. When we are monitoring a large number of experimental results, whether it is expression of all the features in an 'omics experiment or the outcomes of all the experiments done in the lifetime of a project, we expect to see rare outcomes that occur by chance. The use of P values, which assign a measure of rarity to a single experimental outcome, is misleading when many experiments are considered. Consequently, these values need to be adjusted and reinterpreted. The methods that achieve this are called multiple-testing corrections. We discuss the basic principles of this analysis and illustrate several approaches.

Recall the interpretation of the P value obtained from a single two-sample t -test: the probability that the test would produce a statistic at least as extreme, assuming that the null hypothesis is true. Significance is assigned when $P \leq \alpha$, where α is the type I error rate set to control false positives. Applying conventional $\alpha = 0.05$, we expect a 5% chance of making a false positive inference. This is the per-comparison error rate (PCER).

When we now perform N tests, this relatively small PCER can result in a large number of false positive inferences, αN . For example, if $N = 10,000$, as is common in analyses that examine large gene sets, we expect 500 genes to be incorrectly associated with an effect for $\alpha = 0.05$. If the effect chance is 10% and test power is 80%, we'll conclude that 1,250 genes show an effect, and we will be wrong 450 out of 1,250 times. In other words, roughly 1 out of 3 'discoveries' will be false. For cases in which the effect chance is even lower, our list of significant genes will be over-run with false positives: for a 1% effect chance, 6 out of 7 (495 of 575) discoveries are false. The role of multiple-testing correction methods is to mitigate these issues—a large

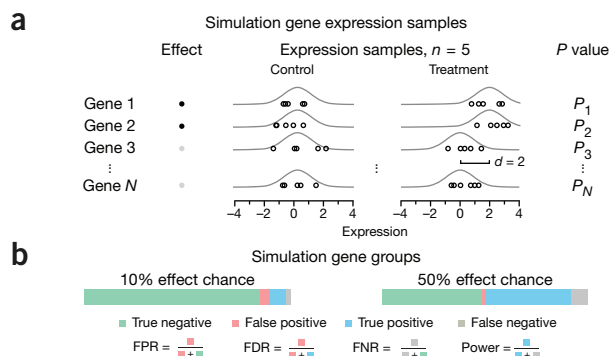


Figure 1 | The experimental design of our gene expression simulation. (a) A gene's expression was simulated by a control and treatment sample ($n = 5$ each) of normally distributed values ($\mu = 0$, $\sigma = 1$). For a fraction of genes, an effect size $d = 2$ (80% power) was simulated by setting $\mu = 2$. (b) Gene data sets were generated for 10% and 50% effect chances. P values were tested at $\alpha = 0.05$, and inferences were categorized as shown by the color scheme. For each data set and correction method, false positive rate (FPR), false detection rate (FDR) and power were calculated. FNR is the false negative rate.

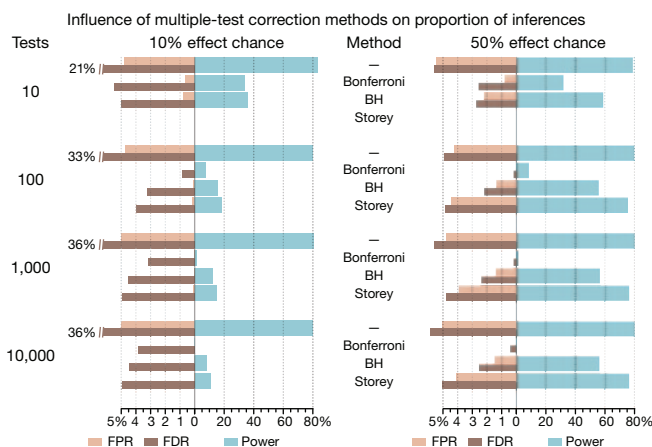


Figure 2 | Family-wise error rate (FWER) methods such as Bonferroni's negatively affect statistical power in comparisons across many tests. False discovery rate (FDR)-based methods such as Benjamini-Hochberg (BH) and Storey's are more sensitive. Bars show false positive rate (FPR), FDR and power for each combination of effect chance and N on the basis of inference counts using P values from the gene expression simulation (Fig. 1) adjusted with different methods (unadjusted (—), Bonferroni, BH and Storey). Storey's method did not provide consistent results for $N = 10$ because a larger number of tests is needed.

number of false positives and large fraction of false discoveries—while ideally keeping power high.

There are many adjustment methods; we will discuss common ones that adjust the P value. To illustrate their effect, we performed a simulation of a typical 'omics expression experiment in which N genes are tested for an effect between control and treatment (Fig. 1a). Some genes were simulated to have differential expression with an effect size $d = 2$, which corresponded to a test power of 80% at $\alpha = 0.05$. The P value for the difference in expression between control and treatment samples was computed with a two-sample t -test. We created data sets with $N = 10, 100, 1,000$ and $10,000$ genes and an effect chance (percentage of genes having a nonzero effect) of 10% and 50% (Fig. 1b). We performed the simulation 100 times for each combination of N and effect chance to reduce the variability in the results to better illustrate trends, which are shown in Figure 2.

Figure 1b defines useful measures of the performance of the multiple-comparison experiment. Depending on the correction method, one or more of these measures are prioritized. The false positive rate (FPR) is the chance of inferring an effect when no effect is present. Without P value adjustment, we expect FPR to be close to α . The false discovery rate (FDR) is the fraction of positive inferences that are false. Technically, this term is reserved for the expected value of this fraction over all samples—for any given sample, the term false discovery percentage (FDP) is used, but either can be used if there is no ambiguity. Analogously to the FDR, the false nondiscovery rate (FNR) measures the error rate in terms of false negatives. Together the FDR and FNR are the multiple-test equivalents of type I and type II error levels. Finally, power is the fraction of real effects that are detected¹. The performance of popular correction methods is illustrated using FPR, FDR and power in Figure 2.

The simplest correction method is Bonferroni's, which adjusts the P values by multiplying them by the number of tests, $P' = PN$, up to a maximum value of $P' = 1$. As a result, a P value may lose its significance in the context of multiple tests. For example, for $N = 10,000$ tests, an observed $P = 0.00001$ is adjusted $P' = 0.1$. The effect of this

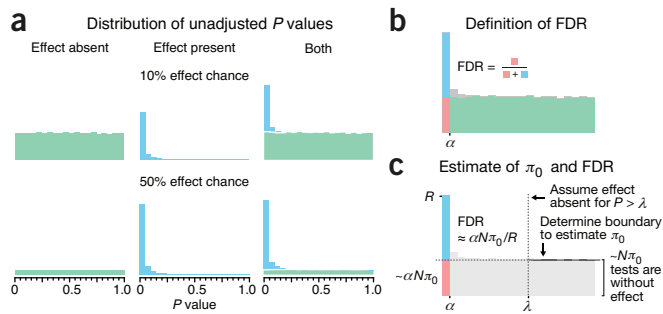


Figure 3 | The shape of the distribution of unadjusted P values can be used to infer the fraction of hypotheses that are null and the false discovery rate (FDR). (a) P values from null are expected to be distributed uniformly, whereas those for which the null is false will have more small values. Shown are distributions from the simulation for $N = 1,000$. (b) Inference types using color scheme of **Figure 1b** on the P value histogram. The FDR is the fraction of $P < \alpha$ that correspond to false positives. (c) Storey's method first estimates the fraction of comparisons for which the null is true, π_0 , by counting the number of P values larger than a cutoff λ (such as 0.5) relative to $(1 - \lambda)N$ (such as $N/2$), the count expected when the distribution is uniform. If R discoveries are observed, about $\alpha N \pi_0$ are expected to be false positives, and FDR can be estimated by $\alpha N \pi_0 / R$.

correction is to control the probability of committing even one type I error across all tests. The chance of this is called the family-wise error rate (FWER), and Bonferroni's correction ensures that $\text{FWER} < \alpha$.

FWER methods such as Bonferroni's are extremely conservative and greatly reduce the test's power in order to control the number of false positives, particularly as the number of tests increases (**Fig. 2**). For $N = 10$ comparisons, our simulation shows a reduction in power for Bonferroni from 80% to ~33% for both 10% and 50% effect chance. These values drop to ~8% for $N = 100$, and by the time we are testing a large data set with $N = 10,000$, our power is ~0.2%. In other words, for a 10% effect chance, out of the 1,000 genes that have an effect, we expect to find only 2! Unless the cost of a false positive greatly outweighs the cost of a false negative, applying Bonferroni correction makes for an inefficient experiment. There are other FWER methods (such as Holm's and Hochberg's) that are designed to increase power by applying a less stringent adjustment to the P values. The benefits of these variants are realized when the number of comparisons is small (for example, < 20) and the effect rate is high, but neither method will rescue the power of the test for a large number of comparisons.

In most situations, we are willing to accept a certain number of false positives, measured by FPR, as long as the ratio of false positives to true positives is low, measured by FDR. Methods that control FDR—such as Benjamini-Hochberg (BH), which scales P values in inverse proportion to their rank when ordered—provide better power characteristics than FWER methods. Our simulation shows that their power does not decrease as quickly as Bonferroni's with N for a small effect chance (for example, 10%) and actually increases with N when the effect chance is high (**Fig. 2**). At $N = 1,000$, whereas Bonferroni correction has a power of $< 2\%$, BH maintains 12% and 56% power at 10% and 50% effect rate while keeping FDR at 4.4% and 2.2%, respectively. Now, instead of identifying two genes at $N = 10,000$ and effect rate 10% with Bonferroni, we find 88 and are wrong only four times.

The final method shown in **Figure 2** is Storey's, which introduces two useful measures: π_0 and the q value. This approach is based on the observation that if the requirements of the t -test are met, the distribution of its P values for comparisons for which the null is true is expected

to be uniform (by definition of the P value). In contrast, comparisons corresponding to an effect will have more P values close to 0 (**Fig. 3a**). In a real-world experiment we do not know which comparisons truly correspond to an effect, so all we see is the aggregate distribution, shown as the third histogram in **Figure 3a**. If the effect rate is low, most of our P values will come from cases in which the null is true, and the peak near 0 will be less pronounced than for a high effect chance. The peak will also be attenuated when the power of the test is low.

When we perform the comparison $P \leq \alpha$ on unadjusted P values, any values from the null will result in a false positive (**Fig. 3b**). This results in a very large FDR: for the unadjusted test, $\text{FDR} = 36\%$ for $N = 1,000$ and 10% effect chance. Storey's method adjusts P values with a rank scheme similar to that of BH but incorporates the estimate of the fraction of tests for which the null is true, π_0 . Conceptually, this fraction corresponds to part of the distribution below the optimal boundary that splits it into uniform (P under true null) and skewed components (P under false null) (**Fig. 3b**). Two common estimates of π_0 are twice the average of all P values (Pound and Cheng's method) and $2/N$ times the number of P values greater than 0.5 (Storey's method). The latter is a specific case of a generalized estimate in which a different cutoff, λ , is chosen (**Fig. 3c**). Although π_0 is used in Storey's method in adjusting P values, it can be estimated and used independently. Storey's method performs very well, as long as there are enough comparisons to robustly estimate π_0 . For all simulation scenarios, power is better than BH, and FDR is more tightly controlled at 5%. Use the interactive graphs in **Supplementary Table 1** to run the simulation and explore adjusted P -value distributions.

The consequences of misinterpreting the P value are repeatedly raised^{2,3}. The appropriate measure to report in multiple-testing scenarios is the q value, which is the FDR equivalent of the P value. Adjusted P values obtained from methods such as BH and Storey's are actually q values. A test's q value is the minimum FDR at which the test would be declared significant. This FDR value is a collective measure calculated across all tests with $\text{FDR} \leq q$. For example, if we consider a comparison with $q = 0.01$ significant, then we accept an FDR of at most 0.01 among the set of comparisons with $q \leq 0.01$. This FDR should not be confused with the probability that any given test is a false positive, which is given by the local FDR. The q value has a more direct meaning to laboratory activities than the P value because it relates the proportion of errors in the quantity of interest—the number of discoveries.

The choice of correction method depends on your tolerance for false positives and the number of comparisons. FDR methods are more sensitive, especially when there are many comparisons, whereas FWER methods sacrifice sensitivity to control false positives. When the assumptions of the t -test are not met, the distribution of P values may be unusual and these methods lose their applicability—we recommend always performing a quick visual check of the distribution of P values from your experiment before applying any of these methods.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2900).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Martin Krzywinski & Naomi Altman

1. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 1139–1140 (2013).
2. Nuzzo, R. *Nature* **506**, 150–152 (2014).
3. Anonymous. Trouble at the lab. *Economist* 26–30 (19 October 2013).

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.