How does multiple testing correction work?

William S Noble

When prioritizing hits from a high-throughput experiment, it is important to correct for random events that falsely appear significant. How is this done and what methods should be used?

Imagine that you have just invested a substantial amount of time and money in a shotgun proteomics experiment designed to identify proteins involved in a particular biological process. The experiment successfully identifies most of the proteins that you already know to be involved in the process and implicates a few more. Each of these novel candidates will need to be verified with a follow-up assay. How do you decide how many candidates to pursue?

The answer lies in the tradeoff between the cost associated with a false positive versus the benefit of identifying a novel participant in the biological process that you are studying. False positives tend to be particularly problematic in genomic or proteomic studies where many candidates must be statistically tested.

Such studies may include identifying genes that are differentially expressed on the basis of microarray or RNA-Seq experiments, scanning a genome for occurrences of candidate transcription factor binding sites, searching a protein database for homologs of a query protein or evaluating the results of a genome-wide association study. In a nutshell, the property that makes these experiments so attractive—their massive scale—also creates many opportunities for spurious discoveries, which must be guarded against.

In assessing the cost-benefit tradeoff, it is helpful to associate with each discovery a statistical confidence measure. These measures may be stated in terms of *P*-values, false discovery rates or *q*-values. The goal of this article is to provide an intuitive understanding of these confidence measures, a sense for how they are computed and some guidelines for how to select an appropriate measure for a given experiment.

William S. Noble is at the Department of Genome Sciences, Department of Computer Science and Engineering, University of Washington, Seattle, Washington, USA.
e-mail: william-noble@u.washington.edu

As a motivating example, suppose that you are studying CTCF, a highly conserved zinc-finger DNA-binding protein that exhibits diverse regulatory functions and that may play a major role in the global organization of the chromatin architecture of the human genome[1]. To better understand this protein, you want to identify candidate CTCF binding sites in human chromosome 21. Using a previously published model of the CTCF binding motif (**Fig. 1a**)[2], each 20 nucleotide (nt) sub-sequence of chromosome 21 can be scored for its similarity to the CTCF motif. Considering both DNA strands, there are 68 million such subsequences. **Figure 1b** lists the top 20 scores from such a search.

Interpreting scores with the null hypothesis and the *P*-value

How biologically meaningful are these scores? One way to answer this question is to assess the probability that a particular score would occur by chance. This probability can be estimated by defining a 'null hypothesis' that represents, essentially, the scenario that we are not interested in (that is, the random occurrence of 20 nucleotides that match the CTCF binding site).

The first step in defining the null hypothesis might be to shuffle the bases of chromosome 21. After this shuffling procedure, high-scoring occurrences of the CTCF motif will only appear because of random chance. Then, the shuffled chromosome can be rescanned with the same CTCF matrix. Performing this procedure results in the distribution of scores shown in **Figure 1c**.

Although it is not visible in **Figure 1c**, out of the 68 million 20-nt sequences in the shuffled chromosome, only one had a score ≥26.30. In statistics, we say that the probability of observing this score under the null hypothesis is 1/68 million, or 1.5×10^{-8} . This probability—the probability that a score at least as large as the observed score would occur in data drawn according to the null hypothesis—is called the *P*-value.

Likewise, the *P*-value of a candidate CTCF binding site with a score of 17.0 is equal to the percentage of scores in the null distribution that are ≥17.0. Among the 68 million null scores shown in **Figure 1c**, 35 are ≥17.0, leading to a *P*-value of 5.5×10^{-7} (35/68 million). The *P*-value associated with score x corresponds to the area under the null distribution to the right of x (**Fig. 1d**).

Shuffling the human genome and rescanning with the CTCF motif is an example of an 'empirical null model'. Such an approach can be inefficient because a large number of scores must be computed. In some cases, however, it is possible to analytically calculate the form of the null distribution and calculate corresponding *P*-values (that is, by defining the null distribution with mathematical formulae rather than by estimating it from measured data).

In the case of scanning for CTCF motif occurrences, an analytic null distribution (gray line in **Fig. 1d**) can be calculated using a dynamic programming algorithm, assuming that the sequence being scanned is generated randomly with a specified frequency of each of the four nucleotides[3]. This distribution allows us to compute, for example, that the *P*-value associated with the top score in **Figure 1b** is 2.3×10^{-10} (compared to 1.5×10^{-8} under the empirical null model). This *P*-value is more accurate and much cheaper to compute than the *P*-value estimated from the empirical null model.

In practice, determining whether an observed score is statistically significant requires comparing the corresponding statistical confidence measure (the *P*-value) to

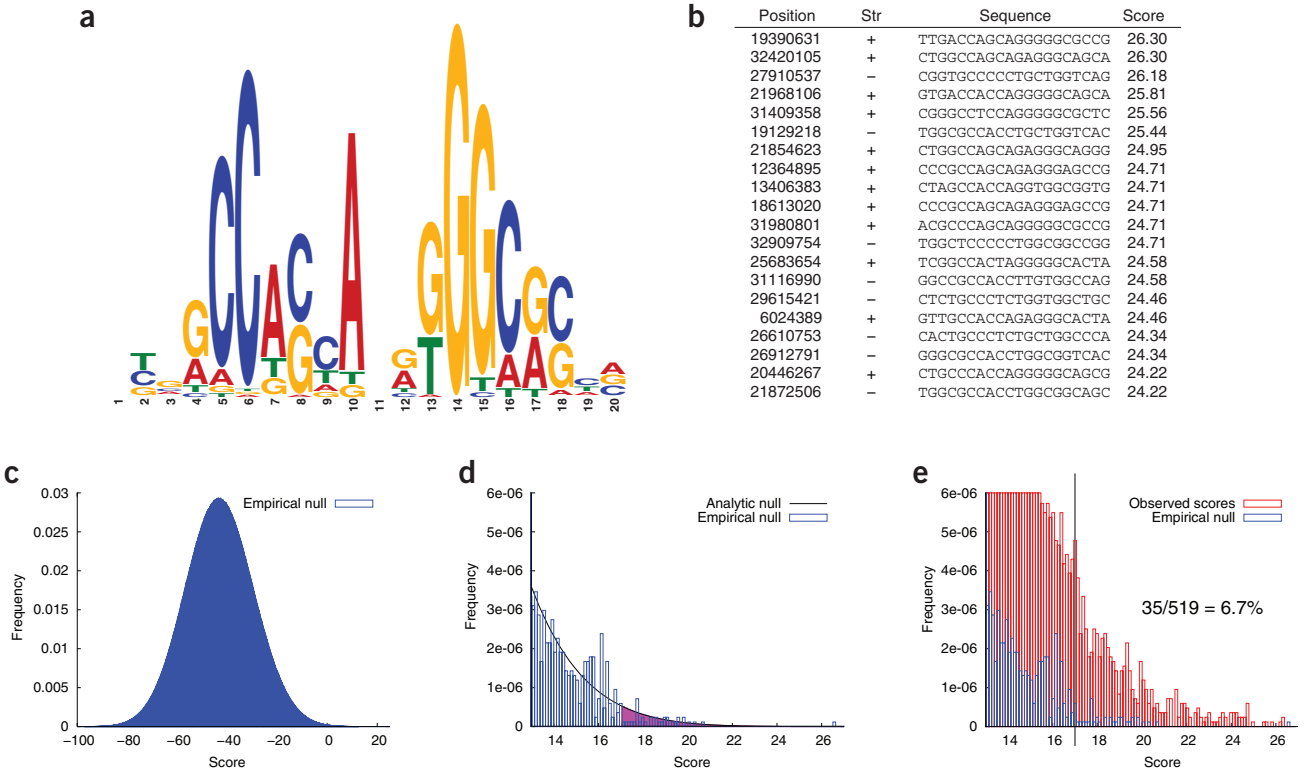

Figure 1 Associating confidence measures with CTCF binding motifs scanned along human chromosome 21. (**a**) The binding preference of CTCF[2] represented as a sequence logo[9], in which the height of each letter is proportional to the information content at that position. (**b**) The 20 top-scoring occurrences of the CTCF binding site in human chromosome 21. Coordinates of the starting position of each occurrence are given with respect to human genome assembly NCBI 36.1. (**c**) A histogram of scores produced by scanning a shuffled version of human chromosome 21 with the CTCF motif. (**d**) This panel zooms in on the right tail of the distribution shown in **c**. The blue histogram is the empirical null distribution of scores observed from scanning a shuffled chromosome. The gray line is the analytic distribution. The *P*-value associated with an observed score of 17.0 is equal to the area under the curve to the right of 17.0 (shaded pink). (**e**) The false discovery rate is estimated from the empirical null distribution for a score threshold of 17.0. There are 35 null scores >17.0 and 519 observed scores >17.0, leading to an estimate of 6.7%. This procedure assumes that the number of observed scores equals the number of null scores.

a confidence threshold α. For historical reasons, many studies use thresholds of $\alpha = 0.01$ or $\alpha = 0.05$, though there is nothing magical about these values. The choice of the significance threshold depends on the costs associated with false positives and false negatives, and these costs may differ from one experiment to the next.

Why *P*-values are problematic in a high-throughput experiment

Unfortunately, in the context of an experiment that produces many scores, such as scanning a chromosome for CTCF binding sites, reporting a *P*-value is inappropriate. This is because the *P*-value is only statistically valid when a single score is computed. For instance, if a single 20-nt sequence had been tested as a match to the CTCF binding site, rather than scanning all of chromosome 21, the *P*-value could be used directly as a statistical confidence measure.

In contrast, in the example above, 68 million 20-nt sequences were tested. In the case of a score of 17.0, even though it is associated with a seemingly small *P*-value of 5.5×10^{-7} (the chance of obtaining such a *P*-value from null data is less than one in a million), scores of 17.0 or larger were in fact observed in a scan of the shuffled genome, owing to the large number of tests performed. We therefore need a 'multiple testing correction' procedure to adjust our statistical confidence measures based on the number of tests performed.

Correcting for multiple hypothesis tests

Perhaps the simplest and most widely used method of multiple testing correction is the Bonferroni adjustment. If a significance threshold of α is used, but n separate tests are performed, then the Bonferroni adjustment deems a score significant only if the corresponding *P*-value is $\leq \alpha/n$. In the CTCF example, we considered 68 million distinct 20-mers as candidate CTCF sites, so achieving statistical significance at $\alpha = 0.01$ according to the Bonferroni criterion would require a *P*-value $< 0.01/(68 \times 10^6) = 1.5 \times 10^{-10}$. Because the smallest observed *P*-value in **Figure 1b** is 2.3×10^{-10}, no scores are deemed significant after correction.

The Bonferroni adjustment, when applied using a threshold of α to a collection of n scores, controls the 'family-wise error rate'. That is, the adjustment ensures that for a given score threshold, one or more larger scores would be expected to be observed in the null distribution with a probability of α. Practically speaking, this means that, given a set of CTCF sites with a Bonferroni adjusted significance threshold of $\alpha = 0.01$, we can be 99% sure that none of the scores would be observed by chance when drawn according to the null hypothesis.

In many multiple testing settings, minimizing the family-wise error rate is too strict. Rather than saying that we want to be 99% sure that none of the observed scores is drawn according to the null, it is frequently sufficient to identify a set of scores for which a specified percentage of scores are drawn according to the null. This is the basis of multiple testing correction using false discovery rate (FDR) estimation.

The simplest form of FDR estimation is illustrated in **Figure 1e**, again using an empirical null distribution for the CTCF scan. For a specified score threshold $t = 17.0$, we count the number s_{obs} of observed scores $\geq t$ and the number s_{null} of null scores $\geq t$. Assuming that the total number of observed scores and null scores are equal, then the estimated FDR is simply s_{null}/s_{obs}. In the case of our CTCF scan, the FDR associated with a score of 17.0 is 35/519 = 6.7%.

Note that, in **Figure 1e**, FDR estimates were computed directly from the score. It is also possible to compute FDRs from *P*-values using the Benjamini-Hochberg procedure, which relies on the *P*-values being uniformly distributed under the null hypothesis[4]. For example, if the *P*-values are uniformly distributed, then the *P*-value 5% of the way down the sorted list should be ~0.05. Accordingly, the procedure consists of sorting the *P*-values in ascending order, and then dividing each observed *P*-value by its percentile rank to get an estimated FDR. In this way, small *P*-values that appear far down the sorted list will result in small FDR estimates, and vice versa.

In general, when an analytical null model is available, you should use it to compute *P*-values and then use the Benjamini-Hochberg procedure because the resulting estimated FDRs will be more accurate. However, if you only have an empirical null model, then there is no need to estimate *P*-values in an intermediate step; instead you may directly compare your score distribution to the empirical null, as in **Figure 1e**.

These simple FDR estimation methods are sufficient for many studies, and the resulting estimates are provably conservative with respect to a specified null hypothesis; that is, if the simple method estimates that the FDR associated with a collection of scores is 5%, then on average the true FDR is ≤5%. However, a variety of more sophisticated methods have been developed for achieving more accurate FDR estimates (reviewed in ref. 5). Most of these methods focus on estimating a parameter π_0, which represents the percentage of the observed scores that are drawn according to the null distribution. Depending on the data, applying such methods may make a big difference or almost no difference at all. For the CTCF scan, one such method[6] assigns slightly lower estimated FDRs to each observed score, but the number of sites identified at a 5% FDR threshold remains unchanged relative to the simpler method.

Complementary to the FDR, Storey[6] proposed defining the *q*-value as an analog of the *P*-value that incorporates FDR-based multiple testing correction. The *q*-value is motivated, in part, by a somewhat unfortunate mathematical property of the FDR: when considering a ranked list of scores, it is possible for the FDR associated with the first m scores to be higher than the FDR associated with the first $m + 1$ scores. For example, the FDR associated with the first 84 candidate CTCF sites in our ranked list is 0.0119, but the FDR associated with the first 85 sites is 0.0111. Unfortunately, this property (called nonmonotonicity, meaning that the FDR does not consistently get bigger) can make the resulting FDR estimates difficult to interpret. Consequently, Storey proposed defining the *q*-value as the minimum FDR attained at or above a given score. If we use a score threshold of T, then the *q*-value associated with T is the expected proportion of false positives among all of the scores above the threshold. This definition yields a well-behaved measure that is a function of the underlying score. We saw, above, that the Bonferroni adjustment yielded no significant matches at $\alpha = 0.05$. If we use FDR analysis instead, then we are able to identify a collection of 519 sites at a *q*-value threshold of 0.05.

In general, for a fixed significance threshold and fixed null hypothesis, performing multiple testing correction by means of FDR estimation will always yield at least as many significant scores as using the Bonferroni adjustment. In most cases, FDR analysis will yield many more significant scores, as in our CTCF analysis. The question naturally arises, then, whether a Bonferroni adjustment is ever appropriate.

Costs and benefits help determine the best correction method

Like choosing a significance threshold, choosing which multiple testing correction method to use depends upon the costs associated with false positives and false negatives. In particular, FDR analysis is appropriate if follow-up analyses will depend upon groups of scores. For example, if you plan to perform a collection of follow-up experiments and are willing to tolerate having a fixed percentage of those experiments fail, then FDR analysis may be appropriate. Alternatively, if follow-up will focus on a single example, then the Bonferroni adjustment is more appropriate.

It is worth noting that the statistics literature describes a related probability score, known as the 'local FDR'[7]. Unlike the FDR, which is calculated with respect to a collection of scores, the local FDR is calculated with respect to a single score. The local FDR is the probability that a particular test gives rise to a false positive. In many situations, especially if we are interested in following up on a single gene or protein, this score may be precisely what is desired. However, in general, the local FDR is quite difficult to estimate accurately.

Furthermore, all methods for calculating *P*-values or for performing multiple testing correction assume a valid statistical model—either analytic or empirical—that captures dependencies in the data. For example, scanning a chromosome with the CTCF motif leads to dependencies among overlapping 20-nt sequences. Also, the simple null model produced by shuffling assumes that nucleotides are independent. If these assumptions are not met, we risk introducing inaccuracies in our statistical confidence measures.

In summary, in any experimental setting in which multiple tests are performed, *P*-values must be adjusted appropriately. The Bonferroni adjustment controls the probability of making one false positive call. In contrast, false discovery rate estimation, as summarized in a *q*-value, controls the error rate among a set of tests. In general, multiple testing correction can be much more complex than is implied by the simple methods described here. In particular, it is often possible to design strategies that minimize the number of tests performed for a particular hypothesis or set of hypotheses. For more in-depth treatment of multiple testing issues, see reference 8.

ACKNOWLEDGMENTS

National Institutes of Health award P41 RR0011823.

1. Phillips, J.E. & Corces, V.G. *Cell* **137**, 1194–1211 (2009).
2. Kim, T.H. *et al. Cell* **128**, 1231–1245 (2007).
3. Staden, R. *Methods Mol. Biol.* **25**, 93–102 (1994).
4. Benjamini, Y. & Hochberg, Y. *J. R. Stat. Soc., B* **57**, 289–300 (1995).
5. Kerr, K.F. *Bioinformatics* **25**, 2035–2041 (2009).
6. Storey, J.D. *J. R. Stat. Soc. Ser. A Stat. Soc.* **64**, 479–498 (2002).
7. Efron, B., Tibshirani, R., Storey, J. & Tusher, V. *J. Am. Stat. Assoc.* **96**, 1151–1161 (2001).
8. Dudoit, S. & van der Laan, M.J. *Multiple Testing Procedures with Applications To Genomics* (Springer, New York, 2008).
9. Schneider, T.D. & Stephens, R.M. *Nucleic Acids Res.* **18**, 6097–6100 (1990).