Peer-reviewed version available at *High-Throughput* **2018**, *7*, 23; doi:10.3390/ht7030023

Article

P-Value Histograms: Inference and Diagnostics

Patrick Breheny[1], Arnold Stromberg[2], and Joshua Lambert[2]*

[1] Department of Biostatistics, University of Iowa
[2] Department of Statistics, University of Kentucky
* Correspondence: joshua.lambert@uky.edu; Tel.: 1-859-257-6915

Abstract: It is increasingly common for experiments in biology and medicine to involve large numbers of hypothesis tests. A natural graphical method for visualizing these tests is to construct a histogram from the *p*-values of these tests. In this article, we examine the shapes, both normal and abnormal, that these histograms can take on, as well as present simple inferential procedures that help to interpret the shapes in terms of diagnosing potential problems with the experiment. We examine potential causes of these problems in detail, and discuss potential remedies. Throughout, examples of abnormal-looking *p*-value histograms are provided and based on case studies involving real biological experiments.

Keywords: p-value, histograms, inference, diagnostics

1. Introduction

Since the advent of high-throughput technology, it has become common for experiments in biology and medicine to involve large numbers of hypothesis tests. A natural graphical method for visualizing the body of these tests is to take the *p*-values from these tests and construct a histogram. If the null hypothesis is true for all features[1], these *p*-values follow a uniform distribution, which corresponds to a flat-looking histogram. Figure 1 illustrates an idealized version of this histogram in which 10,000 *p*-values have been drawn from a uniform random number generator.

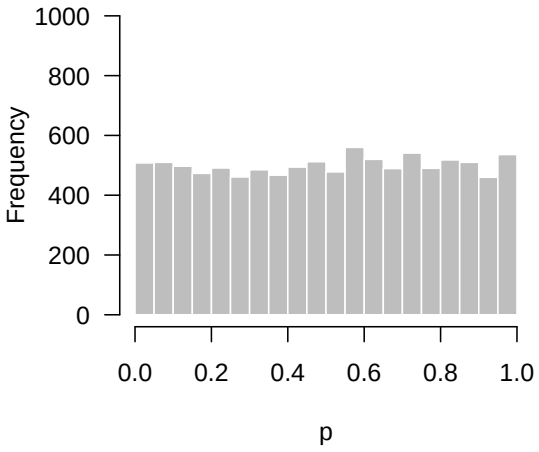


Figure 1. Simulated *p*-values from an idealized setting in which all null hypotheses are true.

Of course, one typically hopes that some of these null hypotheses are incorrect, and that there is an overabundance of low *p*-values. For example, Figure 2 illustrates the *p*-values of an experiment by Rogier *et al.* [1] where approximately 20,000 genes were compared using two sample t-tests. In the histogram, the *p*-values appear to be relatively uniform except for the clear overabundance of very low *p*-values.

There has been a tremendous amount of work in the past two decades, in particular involving false discovery rates [2], extending multiple comparison procedures to large-scale simultaneous inference questions such as these. Naturally, the vast majority of this work has focused on questions involving individual hypotheses. Our focus here, however, concerns what the *p*-value histogram says about the experiment as a whole. Some examples will help to illustrate what we mean by this.

[1] Throughout, we use the generic term "feature" to refer to the quantity being measured in a high-throughput experiment; in our examples the features are gene expression levels, but all of the ideas in the article are equally applicable to any high-throughput measurement such as metabolite or protein concentrations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2018 doi:10.20944/preprints201808.0091.v1

Peer-reviewed version available at High-Throughput 2018, 7, 23; doi:10.3390/ht7030023

Figure 3 displays another set of p-values from Rogier *et al.* [1] where two other groups were compared using two sample t-tests. In the experiment, not a single hypothesis could be rejected at the 10% false discovery rate level. And yet, as we can see from the figure, the p-values clearly do not seem to be uniformly distributed. There is an apparent overabundance of low p-values, suggesting the existence of genes in mice that genuinely responded to dextran sulfate sodium (DSS) in a three-way ANOVA. However, the experiment is not sufficiently powered to detect them after making corrections for multiple testing.

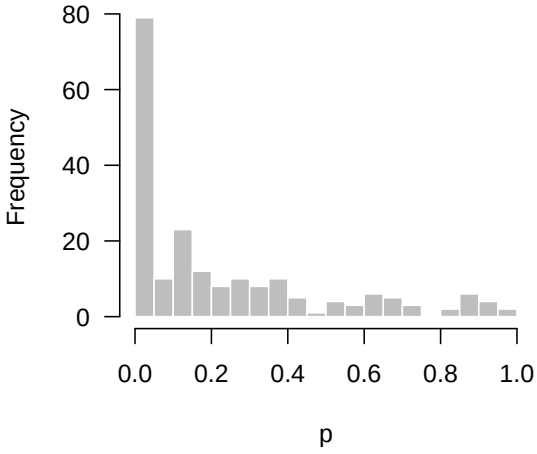


Figure 2. p-values from Rogier *et al.* [1]

Lastly, Figure 4 presents the p-values of an experiment by Fischl *et al.* [3] where paired t-tests were used to compare dependent samples. From the histogram, it appears as though something has gone wrong: there is an abundance not of low p-values but of p-values near 0.3. In summary, we have encountered four examples: no interesting features to detect (Figure 1), interesting features easily detected (Figure 2), interesting features present but unable to be detected (Figure 3), and finally, a problematic experiment (Figure 4). We discuss these cases in greater detail below and provide diagnostics for distinguishing between them.

2. Methods

2.1. Higher criticism

For the data presented in Figure 3, not a single null hypothesis could be rejected at a false discovery rate of 5%. And yet, it seems clear from looking at the histogram that *something* is going on and that more low p-values are present than one would expect by chance alone. This claim can be tested using quantiles of the binomial distribution. Let b denote the bin width of the histogram, m denote the number of hypotheses being tested, X denote the number of p-values in the bin closest to zero, and $F_\alpha(m, p)$ denote the α-level quantile of the cumulative distribution function (CDF) of the binomial distribution with size m and probability p. Then, under the global null hypothesis $H_{0j} : p_j \sim \text{Unif}(0, 1)$ for all j, the probability that X exceeds $F_{.95}(m, b)$ is only 5%. NOTE: Arguably, the .975 quantile could be used instead, as it would be consistent with the standard of always applying two-sided tests, although it would seem a one-sided test makes more sense here.

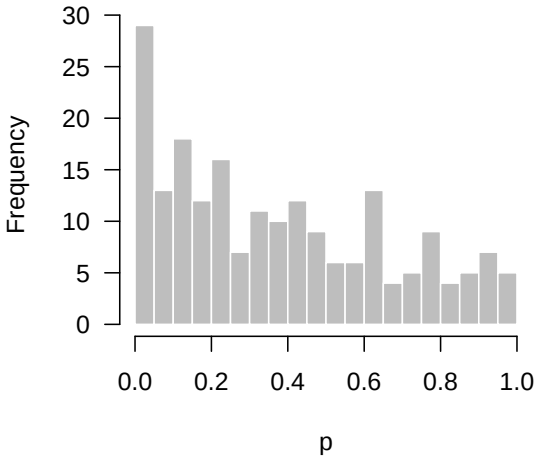
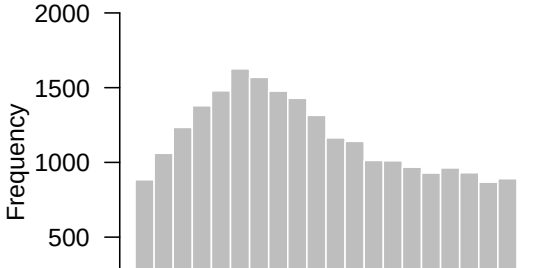


Figure 3. p-values from Rogier *et al.* [1]

Returning to our example from Rogier *et al.* [1] in Figure 3, $b = 0.05$ and $m = 201$, so $F_{.95}(m, b) = 15$. Figure 5 superimposes this threshold upon our earlier histogram. As the figure illustrates, the fact that 27 p-values fall below 0.05 provides significant evidence to reject the global null hypothesis, even though we cannot specifically reject any individual null hypothesis.

This is not a new idea in statistics, and dates back at least to John Tukey, who referred to this question as the "higher criticism" [4]. Tukey



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2018 doi:10.20944/preprints201808.0091.v1

Peer-reviewed version available at High-Throughput 2018, 7, 23; doi:10.3390/ht7030023

proposed the following test statistic, based on a normal approximation to the binomial:

$$\mathrm{HC}_{0.05} = \sqrt{m}\left\{\frac{x}{m} - 0.05\right\} - 0.05,$$

where x is the number of p-values that fall below 0.05. One may then reject the global null at a 5% significance level if HC $>$ 1.645. This leads to a very similar threshold as the above method for large numbers of tests (for example, with $m = 1,000$ tests, the binomial threshold is 62 and the Tukey threshold is 63). We prefer the more accurate binomial threshold for our purposes here, but note that Tukey's closed-form approach has advantages for theoretical study and has received renewed attention in the past decade in the field of high-dimensional data analysis [5–7].

So, what to make of situations like that in Figure 5? Obviously, the main point of these sorts of experiments is to assess the veracity of individual hypotheses, and in that sense an experiment giving rise to Figure 5 must be viewed as unsuccessful. However, the higher criticism here implies that there *is* something to find — this experiment failed to find it, but another experiment, perhaps carried out with an improved experimental design or additional observations, might be successful. This is in contrast to the conclusion one would reach after looking at the histogram in Figure 1, which suggests that there is little hope in conducting another experiment investigating the same biological question, as there is simply nothing to find.

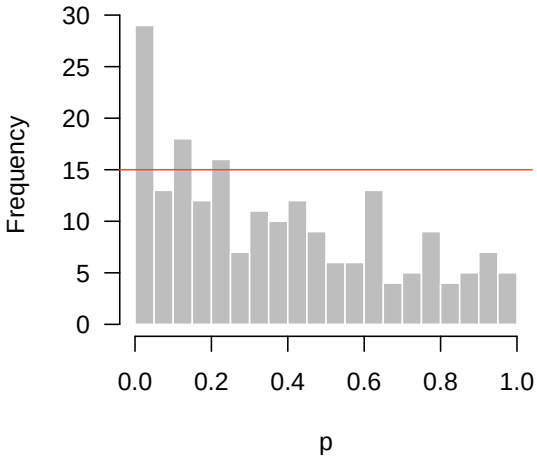


Figure 5. Rogier *et al.* [1]: Higher criticism

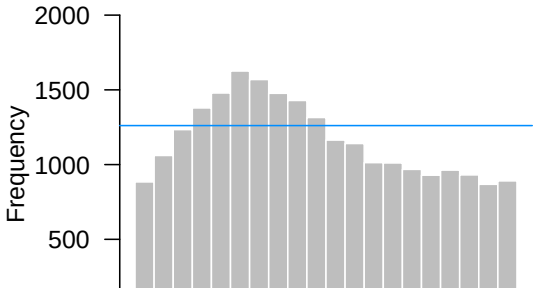
2.2. Quality control

The same basic idea can be used to test for departures from uniformity anywhere between 0 and 1, not necessarily only among low p-values. It is straightforward to extend the approach from Section 2.1 to this case using a Bonferroni correction. With a binwidth of 0.05, this amounts to checking 20 bins, and therefore using a corrected significance threshold of 0.05/20=0.0025, or equivalently, a frequency threshold of $F_{.9975}(m, b)$. For the data from the study by Fischl *et al.* [3] in Figure 4, $m = 23,332$ and $b = 0.05$, so the frequency threshold is 1261. In Figure 6, this threshold is superimposed on the original histogram.

As another example of an experiment whose p-value histogram displays a strange departure from uniformity, Figure 7 presents the p-values of an unpublished NanoString gene expression experiment conducted in 2012 by Dr. Luke Bradley at the University of Kentucky. These p-values were extracted from a two-way interaction effect in a three-way ANOVA model for each gene.

This procedure and the bound we have described are useful as a test of quality control. Here, it establishes that the excess of p-values around 0.3 in Figure 6 and the excess of p-values around 1 in Figure 7 are not due merely to random chance, but that some systematic deviation from the theoretical null distributions of the test statistics has occurred.

In contrast, Figure 8 presents results from an experiment by Matthews and Bridges [8], in which steers were assigned randomly to graze either in a



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2018 doi:10.20944/preprints201808.0091.v1

Peer-reviewed version available at High-Throughput 2018, 7, 23; doi:10.3390/ht7030023

pasture that contained high levels of ergot alkaloids ($n = 10$) or one that did not ($n = 9$). The p-values come from a two-sample t-test of gene expression levels in the liver of the two groups of steers, as measured by NanoString. Although there is something of an abundance of p-values near 0.6, this excess is well within the bounds of random variation.

2.3. Causes of anomalous p-value histograms

In this section, we explore some of the potential causes of the anomalous p-value histograms we have shown above. A related discussion is given by Brad Efron in Section 5 of Efron [9] and Chapter 6.4 of Efron [10]; we hope to add to Efron's remarks by providing specific instances of these violations to illustrate the connection between the cause and the resulting shape of the p-value histogram. In Sections 2.3.1 and 2.3.2, we simulate $m = 10,000$ features belonging to two groups and use a two-sample t-test to test the null hypothesis that the means of the two groups are the same.

2.3.1. Low power

Here, we simulate $n = 4$ observations in each of two groups from the standard normal distribution. For 80% of the features, there is no difference in the means. For the remaining 20%, the difference in means was drawn from a Uniform(-2,2) distribution. The p-value histogram and accompanying higher criticism threshold are shown in the left panel of Figure 9.

With $n = 4$, there is insufficient evidence to reject any of the individual null hypotheses, even at a liberal FDR cutoff of 30%. Nevertheless, the higher criticism threshold clearly indicates that some of the features are non-null. The middle panel of Figure 9 shows a decomposition of the p-value histogram, revealing the contributions from the null and non-null features. As one might imagine from the shape of the histogram, the rise on the left side results from the fact that most of the non-null features have low p-values.

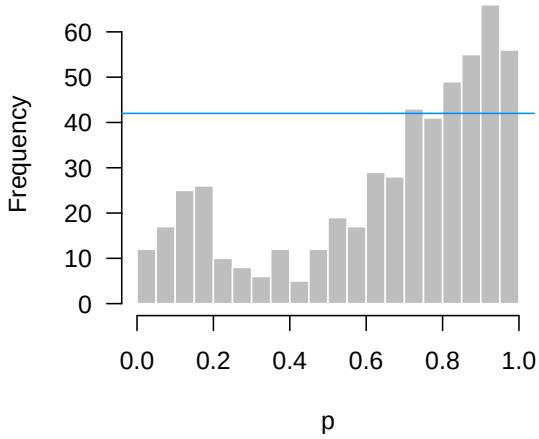
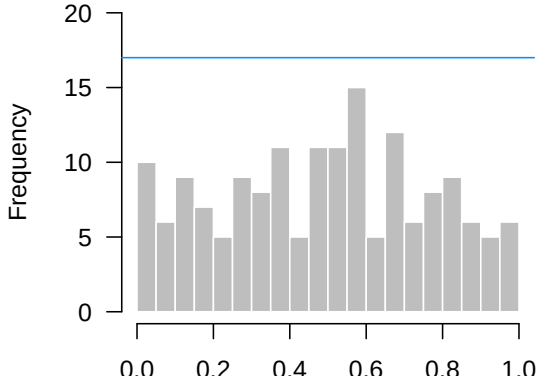


Figure 7. p-values from Bradley NanoString experiment

However, this is not true for *all* of the non-null features. With insufficient power, many of the non-null features turn out to have moderate, or even large p-values and can be found throughout all bins of the histogram. Obtaining these results is likely to be disappointing, since no significant features could be detected, but the p-value histogram and higher criticism indicate reasons for optimism. Although the initial experiment was unable to distinguish null and non-null features, there are indeed interesting features to be discovered, and a second, more adequately powered experiment may be successful at finding them.

To illustrate this, we simulated data under the same settings as above, but with a sample size of $n = 10$ in each group. In marked contrast to the previous results, we can now safely reject 504 null hypotheses at the 5% FDR level. These results are displayed on



the right panel of Figure 9, and show much clearer separation between null and non-null features.

2.3.2. Incorrect distributional assumptions

In Figure 10, we simulate $n = 3$ observations in each of two groups from the exponential distribution with rate 1, then apply a two-sample t-test for each feature. Thus, in this example, all 10,000 features satisfy the null hypothesis. The derivation of p-values from the t-test assumes normally distributed data; here, that assumption is highly inaccurate, the exponential distribution being both highly skewed and having considerably thicker tails than the normal distribution.

Problems with distributional assumptions can be alleviated by choosing more robust, nonparametric methods. For example, replacing the t-test in the above example with a Wilcoxon rank sum test produces an appropriate, uniform-looking histogram. In addition, distributional problems are alleviated as n increases due to the central limit theorem. Increasing n to 30 in each group for this setting also yields a flat, uniform-looking histogram essentially indistinguishable from Figure 1.

2.3.3. Correlation among features

Perhaps the most common cause of an abnormal-looking histogram, however, is correlation among features. With respect to p-value histograms, correlation among the features being tested does not necessarily alter the shape of the histogram: marginally, each p-value still follows a uniform distribution under the null. However, it does mean that there is a greater chance of seeing an irregular deviation from uniformity in the p-value histogram. For example, imagine a bundle of highly correlated features. Due to the correlation, these features will have similar p-values. Where the bundle lies is uniformly distributed, but wherever it lands, a "bump" will appear in the histogram.

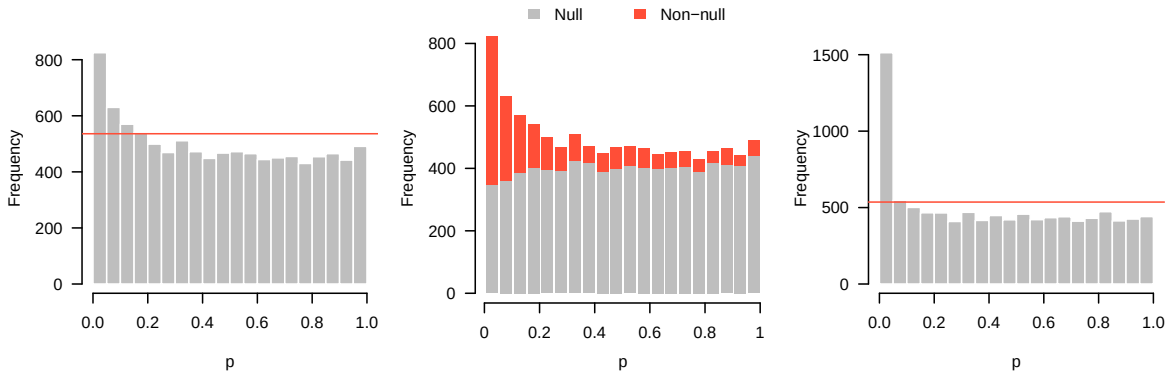


Figure 9. Left: Simulated data with low power. Middle: Same data as in left panel, showing contributions from null and non-null genes. Right: Data simulated under same conditions as left panel, but with adequate power.

The higher criticism and quality control bounds in Sections 2.1 and 2.2 are based on the assumption that the features being tested are mutually independent of each other. The primary practical consequence of correlation among features is that that the QC bound given in Section 2.2 is too low, leading one to conclude that an error has occurred when the irregular shape may simply be explained by correlation among the features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2018 doi:10.20944/preprints201808.0091.v1

Peer-reviewed version available at High-Throughput 2018, 7, 23; doi:10.3390/ht7030023

Fortunately, given an adequate sample size, it is possible to assess the impact of correlation among features using permutation approaches. The idea underlying the permutation approach is simple. Let $\mathbf{X}$ denote the $n \times m$ matrix of feature values (here, gene expression data), with each row of $\mathbf{X}$ denoting an experimental unit consisting of m features. By permuting the rows of $\mathbf{X}$, we accomplish two things. First, we eliminate any association between $\mathbf{X}$ and any other variables or group memberships that we are testing for. Second, by permuting entire rows of $\mathbf{X}$ intact, we preserve any correlation among the rows that is present in the data. Thus, by carrying out the original test on random permutations of $\mathbf{X}$, we obtain p-values from the null distribution but without assuming independence among features.

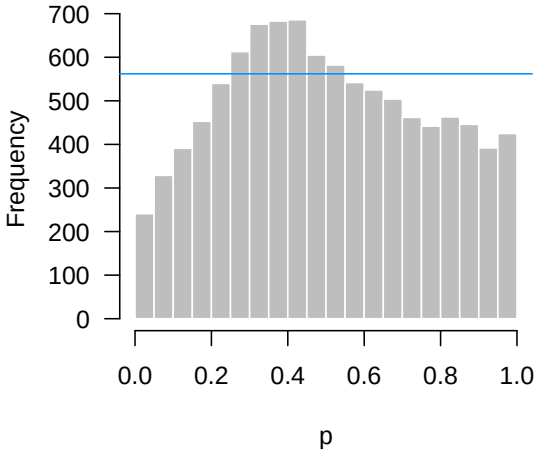


Figure 10. A t-test was applied, even though the data come from a highly non-normal (exponential) distribution.

We repeated the test for the two-way interaction in the Bradley data seen in Figure 7 for 1,000 random permutations of the expression data. For each permutation, we made a p-value histogram and recorded the count in the most highly populated bin. Figure 11 plots the histogram of the original p-values with two lines superimposed. One is the original quality control metric from Section 2.2 which assumes independence among the hypothesis tests, the other is the 95th percentile of the maximum counts from the permutation histograms.

The difference between the lines is striking. In this experiment, the correlation between genes is quite high (root-mean-square correlation among the 536 genes selected for the NanoString experiment was 0.75). As a result, the spike of p-values near 0.9 observed in the data could easily have arisen simply from the correlation among genes. In fact, given the correlation among features, the abnormal-looking histogram of Figure 7 is not particularly abnormal at all, a point clearly communicated by the large gap between the p-value histogram and the "Permutation" line in Figure 11.

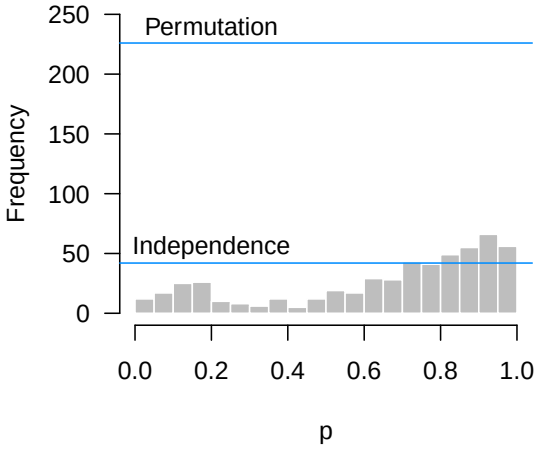


Figure 11. Bradley experiment: Permutation vs independence approaches.

Correlation among features also affects the higher criticism threshold of Section 2.1, although not as much as for quality control thresholds. The same permutation approach can be applied to obtain correlation-adjusted higher criticism thresholds, although in this case we would examine the 95th percentile of the counts for the first bin rather than the maximum count. For the Rogier *et al.* [1] data of Figure 5, the higher criticism bound assuming independence was 15, while the higher criticism bound obtained from the permutation approach was 19.4. This is far less dramatic than the difference in Figure 11 because while correlation leads to bumps in the p-value histogram, those bumps are not systematically located in the lowest bin.

Unfortunately, there are limitations to the permutation approach. One is that it can be computer-intensive if p is large or if the tests themselves are time-consuming to perform. The other issue is that permutation approaches cannot be applied to very small samples. For example, we cannot use a permutation approach to investigate the Fischl *et al.* [3] data from Figure 6, which involves a one-sample t-test with only 3 pairs of subjects. Although the idea can be extended to paired data (by randomly assigning signs to the differences rather than permuting rows), in this case there are only

four distinct random assignments that can be made, and hence four different null histograms to serve as a reference for comparison, which is not sufficient for estimating a 95th percentile.

This is a fundamental limitation with applying permutation approaches to small samples, although the number of available permutations rapidly increases with sample size. For example, in a two-sample study with $n = 3$ in each group, only 10 distinct permutations are available; however, with $n = 10$ in each group, the number of permutations increases to 92,378.

For both of these reasons (small sample sizes and computational burden), it is desirable to develop an analytic method for estimating higher criticism and quality control thresholds that account for correlation among features. Such a development is beyond the scope of this manuscript, but we re-examine this issue in the discussion.

2.4. Remedies

When faced with an abnormal-looking p-value histogram, what action should a researcher take? In this section, we describe possible remedies.

One potential remedy is to increase the sample size by collecting more data. This is most clearly indicated in situations like Figure 3, where there is a clear indication that non-null features are present, but unable to be reliably distinguished from noise. The higher criticism threshold is potentially a very useful tool to guide this decision in terms of whether the additional cost of collecting more data is likely to bear fruit or not.

Alternatively, abnormal-looking p-value histograms may serve as an indication that the assumptions being made in the statistical analysis are not being met (see Section 2.3.2) and that one should consider an alternative approach – for example, a Wilcoxon rank sum test instead of a two-sample $t-$test. It is worth noting that higher sample sizes are beneficial here as well. Not only do larger sample sizes increase the robustness of many statistical tests, they also allow one to fit less restrictive statistical models.

Lastly, we note that abnormal p-value histograms may also indicate that the experimental design should be revised. Although to some extent correlation among features is an unavoidable biological fact, it is also the case that careful experimental designs (randomization, blocking, balance, etc.) reduce this correlation and the potential for confounding factors to induce correlation in an experiment.

An element of design particularly relevant to expression and other sorts of "-omic" data is the issue of normalization. Proper normalization procedures substantially reduce correlations in this sort of data [11]. However, while normalization procedures are well-developed for long-standing technologies such a microarray data [12], this is often not the case for more recent technologies such as NanoString and RNA-Seq.

3. Discussion

In this article, we have taken a closer look at p-value histograms with respect to two questions of vital practical importance:

- Higher criticism: Is there a significant excess of low p-values? In other words, is there any evidence of a systematic biological response in the experiment?
- Quality control: Has something gone wrong in this experiment?

We present straightforward analytic diagnostics to address these questions, as well as a permutation-based approach capable of accounting for correlation among features. As Figure 11 demonstrates, correlation among features is an important issue as it has the potential to dramatically affect p-value histograms.

Our derivation of higher criticism bounds in Section 2.1 and quality control bounds in Section 2.2 assumes that the p-values are "proper" in the sense that $\Pr(p < \alpha) = \alpha$ (i.e., the p-values are uniformly distributed) under the null hypothesis . Many common tests, especially those involving discrete outcomes, are *valid* in that $\Pr(p < \alpha) \leq \alpha$ under the null, but not proper. For these conservative tests,

the higher criticism derivation still holds, although like the tests themselves, the threshold will be conservative. However, for the quality control bound, this issue causes a problem, as a bump in the histogram could be the result of the conservative nature of the test and not an actual problem with the experiment. The quality control bounds derived in Section 2.2 are not likely to be useful for such tests, although the permutation approach may still be used.

An additional factor that can distort p-value histograms, but which is not discussed in Section 2.3, is the effect of correlation among sampling units, possibly brought on by unmeasured confounding variables. The effect of correlation among samples (as opposed to correlation among features) is to broaden the null distribution. If this correlation is not accounted for, it will lead to an inflation of test statistics and a failure to preserve the proper size of the test, rejecting the null hypothesis too often. This is obviously an important issue, although p-value histograms are of little help in diagnosing this issue, since when this issue is present, the histogram appears similar to "ideal" results, with a clear excess of small p-values.

Finally, as noted in Section 2.3.3, it is desirable to develop an analytic method capable of computing higher criticism and quality control thresholds without the need for a permutation approach. Such a method, however, would need to both estimate and account for all pairwise correlations among the features. This is potentially a very large number, especially for genome-wide expression studies. These statistical challenges are not necessarily insurmountable, but they do fall beyond the intended scope of this article; it is a problem we are currently working on.

Despite these limitations, it is our hope that the tools and examples presented in this article will be useful to researchers engaged in testing of high-throughput biological data, particularly since the notion of "troubleshooting" such experiments is largely absent from the scientific literature as problematic and underpowered studies often go unpublished.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

Appendix

The histograms can easily be reproduced in R (www.r-project.org) with the following code, which assumes that a vector p of p-values has already been calculated:

```
b <- 0.05
hist(p, breaks=seq(0, 1, b), col="gray", border="white")

# Higher criticism:
abline(h=qbinom(.95, length(p), b), col="red")

# Quality control:
abline(h=qbinom(1-b*.05, length(p), b), col="blue")
```

References

1. Rogier, E.W.; Frantz, A.L.; Bruno, M.E.C.; Wedlund, L.; Cohen, D.A.; Stromberg, A.J.; Kaetzel, C.S. Secretory antibodies in breast milk promote long-term intestinal homeostasis by regulating the gut microbiota and host gene expression. *Proceedings of the National Academy of Sciences* **2014**, *111*, 3074–3079. doi:10.1073/pnas.1315792111.
2. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B* **1995**, *57*, 289–300.
3. Fischl, A.M.; Heron, P.M.; Stromberg, A.J.; Mcclintock, T.S. Activity-Dependent Genes in Mouse Olfactory Sensory Neurons. *Chemical Senses* **2014**, *39*, 439–449. doi:10.1093/chemse/bju015.
4. Tukey, J. The Philosophy of Multiple Comparisons. *Statistical Science* **1991**, *6*, 100–116.

5. Donoho, D.; Jin, J. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **2004**, *32*, 962–994.
6. Donoho, D.; Jin, J. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences* **2008**, *105*, 14790.
7. Hall, P.; Jin, J. Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics* **2010**, *38*, 1686–1732.
8. Matthews, J.C.; Bridges, P.J. NutriPhysioGenomics applications to identify adaptations of cattle to consumption of ergot alkaloids and inorganic versus organic forms of selenium: altered nutritional, physiological and health states? *Animal Production Science* **2014**, *54*, 1594. doi:10.1071/an14274.
9. Efron, B. Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **2008**, *23*, 1–22.
10. Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*; Cambridge University Press, 2010.
11. Qiu, X.; Brooks, A.; Klebanov, L.; Yakovlev, A. The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics* **2005**, *6*, 120. doi:10.1186/1471-2105-6-120.
12. Irizarry, R.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y.; Antonellis, K.; Scherf, U.; Speed, T. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **2003**, *4*, 249.