Supplementary Text: Capturing Heterogeneity in Gene Expression Studies by "Surrogate Variable Analysis"

Jeffrey T. Leek[1] and John D. Storey[12*]

Address: [1]Department of Biostatistics and [2]Department of Genome Sciences, University of Washington, Seattle, WA 98195 USA

[*]Address for correspondence: `jstorey@u.washington.edu`

Nested KS-Tests: A Procedure to Test Whether a Procedure is Valid

The false discovery rate (FDR) has been discussed extensively and it has been pointed out that the distribution of the null p-values must be "correct" or conservative for FDR estimation or any other standard statistical significance measure to behave properly. What is meant for distribution of the null p-values to be correct is that they are Uniformly distributed in the interval (0,1). The null p-values are have a conservative distribution or they are pushed towards 1 relative to the Uniform(0,1). P-values are constructed to have the Uniform distribution property under the null hypothesis, and if this cannot be done exactly the conservative version is calculated [1]. In a simulation study where the right answer is known, there is no off-the-shelf approach to test whether the null p-values have a proper distribution.

In this study, we use a Kolmogorov-Smirnov (KS) test on the set of null p-values for deviation from the Uniform. However, we want to test whether this is true over many repeated simulations to avoid "getting lucky" on one particular simulated data set. If the set of null p-values are Uniform, then the p-value resulting from the KS test should also follow the Uniform distribution. Therefore, by examining the KS test p-values over all simulations, we can again apply a KS test to verify that these are Uniformly distributed. Here we have employed this nested KS test to compare the relative behavior of each multiple testing procedure discussed. If the quantiles of the KS test p-values follow the diagonal line in a quantile-quantile plot against the quantiles of the Uniform distribution, then this is very strong evidence that the p-values resulting from the procedure are "correct."

On Capturing EH at the P-value or Test-statistic Level

As described in the main text, it appears that methods that adjust for multiple testing dependence by modifying p-values or test-statistics [2–6] are not capable of addressing EH at a sufficient level of generality. A heuristic argument was given in the main text; here, we present a more theoretical argument. Consider the scenario where m p-values are calculated $p_1, p_2, \ldots, p_m$, corresponding to m hypothesis tests. Figure 5 of the main text provides a graphical motivation for the counter-example. The histogram present in this figure is actually composed of p-values all corresponding to true null hypotheses. These p-values should therefore be Uniformly distributed. However, the unmodeled factors manifested as EH have biased their distribution towards zero. At the same time, it is straightforward to simulate a mixture of m_0 correctly distributed Uniform null p-values and $m - m_0$ correctly distributed alternative p-values pushed towards zero, which as a set are indistinguishable in distribution from those in Figure 5.

The reason for this is that it is mathematically impossible to identify a signature of the umodeled factors causing the EH from a one-dimensional summary of each test's data, whether it be a p-value or a test-statistic. The actual set of data itself has to be examined, as is done in the SVA approach. That is, EH can only be in general at the level of the original observed data. To simplify the mathematical argument (although the argument easily extends to general assumptions), assume that the p-values either follow a null probability density function g_0 (which would be Uniform(0,1) for p-values when no EH is present) or an alternative probability density function g_1. Therefore, a randomly selected p-value follows the mixture distribution $\overline{g} = \pi_0 g_0 + (1 - \pi_0) g_1$, where π_0 is the proportion of null hypotheses that are true.

It is impossible to de-convolute this mixture in general. Without specifying π_0 or g_0, there are an infinite number of configurations of π_0, g_0, and g_1 that yield our observed $\overline{g}$. This can be seen by noting that $g_1 = (\overline{g} - \pi_0 g_0)/(1 - \pi_0)$. As we vary g_0 and π_0, we see an infinite number of possible g_1. If we assume that g_0 is known, then one can use the entire set of p-values to estimate $\overline{g}$. It then becomes possible to obtain conservative estimates of π_0 and g_1 based on knowledge of g_0 and the estimate of $\overline{g}$ [7]. However, if there is EH, then it is not possible to assume anything *a priori* about g_0. The p-values or test-statistics therefore cannot be utilized to deconvolute the mixture $\overline{g} = \pi_0 g_0 + (1 - \pi_0) g_1$. In this case, there is not sufficient information to distinguish the null and alternative distributions because the parameters of the mixture are unidentifiable. Therefore, one cannot adjust for EH based only on p-values or test-statistics.

The method in Efron (2004, 2007) [2,3] is an example of a test-statistic only based approach to correct for multiple testing dependence. Efron (2004, 2007) recognizes that when there is large-scale dependence among tests, this may distort the assumed "theoretical null" g_0, and therefore proposes the use of an empirical null distribution $\tilde{g}_0$ to correct for this distortion. The goal is to identify the

correct $\tilde{g}_0$ by only considering the observed statistics and the theoretical null. Efron (2004, 2007) first transforms the observed statistics so that the theoretical null g_0 is a $N(0,1)$ distribution. The empirical null is calculated by scaling and shifting g_0 into the estimated $\tilde{g}_0$ so that in an interval around zero, $\tilde{g}_0 \approx \overline{g}$. This empirical null can then be utilized in place of the theoretical null for calculating the significance of each statistic. Due to the argument given above, in general it will not be possible to obtain a correct null distribution from methods such as this that are based only on test-statistics or p-values. We implemented the method of Efron (2004, 2007) [2, 3] and used that proposed empirical null distribution to calculate p-values for the same thousand simulated studies we described in the main text and compared the p-values corresponding to true nulls to the Uniform(0,1) distribution. The results, shown in Supplementary Figure 11 indicate that the null distribution is not accurately estimated using the empirical null, as we would expect from the theoretical result described above.

The fact that the method of Efron (2004, 2007) does not estimate the true underlying null distribution in this example can also be understood from Figure 5. Working with transformed observed statistics, so that the initial theoretical null is $N(0,1)$, is equivalent to calculating p-values from the observed statistics so that their theoretical null is Uniform(0,1). The transformation he makes on the Normal distribution scale to obtain $\tilde{g}_0$ is equivalent to transforming the set of observed p-values (calculated from the theoretical null) so that this set is approximately Uniformly distributed in some interval $[\lambda, 1]$. It can be seen from Figure 5 that for most intervals of the form $[\lambda, 1]$ where $\lambda \geq 0.4$, the p-values are already indistinguishable from the Uniform distribution before any transformation has taken place. Therefore, the method of Efron (2004, 2007) does not properly capture and correct for EH in this example.

References

1. Dabney AR, Storey JD (2006) A reanalysis of a published affymetrix genechip control dataset. Genome Biology 7:401.

2. Efron B (2004) Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association 99:96–104.

3. Efron B (2007) Correlation and large-scale simultaneous significance testing. Journal of the American Statistical Association 102:93–103.

4. Cai GQ, Sarkar SK (2006) Modified simes' critical values under positive dependence. Journal of Statistical Planning and Inference 136:4129–4146.

5. Benjamini Y, Yekultieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165 – 1188.

6. Pawitan Y, Calza S, Ploner A (2006) Estimation of false discovery proportion under general dependence. Bioinformatics 22:3025–3031.

7. Storey JD (2002) A direct approach to false discovery rates. JRSSB 64:479–498.

8. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, et al. (2001) Gene-expression profiles in hereditary breast cancer. New Engl J Med 344:539–548.