# A Discussion on False Discovery Rate (FDR) with a Simulation Study and an Application on Genomics Data

Md Nahid Hassan, Md Faruk Hossain, Dwaipayan Mukhopadhyay, Anjan Mandal
Department of Mathematical Sciences, UNLV

April 25, 2020

**Abstract**

The false discovery rate (FDR) measures the proportion of false discoveries among a set of hypotheses tests(conducting multiple comparisons) called significant. This quantity is typically estimated based on $p$-values or test statistics. The common approach to the multiplicity problem calls for controlling the familywise error rate(FWER). This approach, though, has faults, and we point out a few. A different approach to problems of multiple significance testing is presented. It calls for controlling the expected proportion of falsely rejected hypothesis the false discovery rate. This error rate is equivalent to the FWER when all hypotheses are true but is smaller otherwise. Therefore, in problems where the control of the false discovery rate rather than that of the FWER is desired, there is potential for a gain in power. A simple sequential Bonferroni type procedure is proved to control the false discovery rate for independent test statistics. A simulation study and a real life genomics data shows that the gain in power is substantial. The informative variable in this study is the per-gene read depth. The framework this topics develop is quite general, and it should be useful in a broad range of scientific applications.

## 1 Introduction

With the advent of reliable statistical tools, multiple testing is commonplace in many scientific areas. For example, in genomics, RNA-seq technology is often used for testing thousands of genes for differential expression among more than one biological conditions. When pursuing multiple inferences, researchers tend to select the (statistically) significant ones for emphasis and support of conclusions. An unguarded use of single-inference procedures results in a greatly increased false positive rate. On this report, our work on the false discovery rate (FDR), and the paper Benjamini & Hochberg (1995), has its origins in two papers concerned with multiple testing of m hypotheses of which unknown $m_0$ are true. First was Schweder & Spjøtvoll (1982), who suggested plotting the ranked

$p$-values, assessing $m_0$ via an eye fitted line, and rejecting the other $m - m_0$ hypotheses. In Hochberg & Benjamini (1990) the authors developed their idea into an algorithm and incorporated the estimate $m_0$ into procedures such as Bonferroni, Holm or Hochberg.
Second was Sorić (1989), who argued forcefully against the use of uncontrolled single-hypothesis testing when many are tested, and used the expected number of false discoveries divided by the number of discoveries as a warning that 'a large part of statistical discoveries may be wrong'.
As the complexity and size of data increased through ages, additional information also increased along with that and in many emerging applications, additional information on the status of a null hypothesis or the power of a test have been used to help better estimate the FDR and Q value. An excellent example, in eQTL studies, gene–single nucleotide polymorphism (SNP) basepair distance informs the prior probability of association between a gene–SNP pair, with local associations generally more likely than distal associations Brem et al. (2002). The incorporation of additional information to estimate FDR in Bayesian setup had been discussed in Storey & Tibshirani (2003). Incorporation of additional information into calculating FDR had been discussed in the form of p value weighting procedure, Stratified FDR control (Sun and others, 2006), stratified local FDR thresholding (Ochoa and others, 2015), and covariate-adjusted conditional FDR estimation (Boca and Leek, 2018). There are pros and cons of using aforementioned procedures such as tratified FDR and local FDR rely on clearly defined strata, which may not always be available or make the best use of information. Covariate-adjusted conditional FDR estimation focuses on only one component of the FDR. $p$-value weighting has been a successful strategy. However, for a p-value weighting method, it remains challenging to derive weights that indeed result in improved power subject to a target FDR level, and how to obtain optimal weights under different optimality criteria is still an open problem. Furthermore, for a $p$-value weighting procedure that is also based on partitioning hypotheses into groups, its inferential results can be considerably affected by how the groups are formed.
Another major aspect of FDR in genomic studies have largely been dealing with "poor odds" problem. Often, initial findings are not replicated in subsequent studies. This lack of reproducibility reflects two difficulties. The first is the "poor odds" problem, i.e. the low probability that any statistically significant finding actually is true. The second is the multiplicity problem inherent in the large number of statistical inferences extracted from the study data. Whittemore (2007) defined Bayesian FDR and offered a simple comprehensive method to solve the multiplicity problem.
In this report we'll present in section 2 the definition and different theoretical aspects of the FDR which contains the classical definition of FDR, introduction and background of pFDR, Bayesian interpretation and some asymptotic properties of the same. In the Section 3 and in the Section 4 we will present a simulation study and a real data study involving genomic test on Leukemia dataset respectively.

# 2 False Discovery Rate

With the increase in the size of data sets available the number of features in a dataset has also increased significantly. It is now often up to the statistician to find as many interesting features in a data set as possible rather than test a very specific hypothesis on

one item. For example, one is more frequently faced with the daunting task of estimating or performing hypothesis tests on thousands of parameters simultaneously. In this kind of situation, one is more interested in the total number of false positives compared to the total number of significant items, rather than making one or more Type I errors.
From Sorić (1989) V being the number of type I errors made, out of the R rejected, by defining

$$FDR - 1 = \frac{E(V)}{E(R)} \tag{1}$$

(Efron (2008) and earlier) we obtain a very appealing error rate, that rather than being merely a warning can serve as a worthy goal to control. Moreover, considering these quantities as a function of the level $\alpha$ at which the individual testing is done, a plausible estimator for the FDR is

$$Q(\alpha) = \frac{\alpha m_0}{R(\alpha)} \tag{2}$$

Indeed, the value depends on $m_0$, but we already had a way to estimate $m_0$ from corresponding setup of any statistical problem. We can obtain an estimate of the FDR by using a $step - up$ method on the sorted series of $p$-values (Benjamini & Hochberg (1995)).

## 2.1 Estimation of False Discovery Rate Using BH Procedure

| | Called significant | Not called significant | Total |
|---|---|---|---|
| True null hypothesis | $V$ | $U = m_0 - V$ | $m_0$ |
| True alternative hypothesis | $S$ | $T = m_1 - S$ | $m_1$ |
| Total | $R$ | $W = m - R$ | $m$ |

Table 1: Number of error committed when testing $m$ null hypothesis

Consider Table 1 giving the various outcomes that occur when m hypothesis tests are performed according to some significance rule, which can either be fixed or data-dependent. The FWER can formally be written as $Pr(V \geq 1)$. In a seminal paper, Benjamini & Hochberg (1995) introduce a new multiple hypothesis testing error measure called the false discovery rate (FDR), which they define as

$$\begin{aligned} FDR &= E\left[\frac{V}{R \vee 1}\right] \\ &= E\left[\frac{V}{R}|R > 0\right] Pr(R > 0) \end{aligned} \tag{3}$$

The only effect of the " R $\vee$ 1" in the denominator of the first expectation is that the ratio V /R is set to zero when R = 0. Benjamini & Hochberg (1995) prove that a particular $p$-value step-up method strongly controls the FDR when the true null hypotheses are

simple and independent, with an extension to "positive regression dependence" in Benjamini & Yekutieli (2001). This procedure was originally introduced by Simes (1986) to weakly control the FWER. When using this procedure, the realized V and R depend on the random outcome of a $p$-value-based algorithm. The Benjamini and Hochberg (1995) procedure works as follows. Suppose that the $p$-values resulting from the $m$ tests are ordered such that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. If we calculate

$$\hat{k} = \arg\max_{1 \leq k \leq m} \left\{ k : p_{(k)} \leq \alpha \cdot k/m \right\} \tag{4}$$

then rejecting the null hypotheses corresponding to $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ provides FDR $= m_0/m \cdot \alpha \leq \alpha$. If no $p$-value satisfies this inequality, then no hypothesis test is called significant. The FDR offers less stringent control over Type I errors than the FWER, and is therefore usually more powerful.

## 2.2 The Positive False Discovery Rate: An Improvement over FDR

Theoretically the above model may seem robust, however, in most cases there is positive probability that R = 0, so this definition is not well-defined. When $m_0 = m$, one would want the false discovery rate to be 1, and that one is not interested in cases where no test is significant. Shaffer (1995) also believed that the inclusion of $Pr(R > 0)$ in the definition of FDR is unsatisfying. These considerations lead Storey et al. (2003) to propose following as an alternative definition, called the *positive false discovery rate* (pFDR).

**Definition 1** *The positive false discovery rate is defined to be*

$$pFDR = E\left[\frac{V}{R} | R > 0\right] Pr(R > 0) \tag{5}$$

The Definition 1 given in Eq. 5 is called the pFDR because it is conditioned on the fact that at least one positive finding has occurred.

There are two clear approaches to false discovery rates that can be taken. The first is to fix the acceptable rate $\alpha$ beforehand and estimate a significance threshold to obtain this rate conservatively on average. The second is to fix the significance threshold and provide a conservative estimate of the rate over that threshold. When taking the first approach, one is forced to use the FDR since the pFDR cannot be controlled in this sense. The pFDR can be conservatively estimated in the second approach. On the other hand by considering false discovery rates for fixed significance regions, one can gain insight into the operating characteristics of the quantities, resulting in improved procedures.

**Remark 1** *An example where confusion in the interpretation of FDR and pFDR is dangerous is the following. One can use the Benjamini & Hochberg (1995) procedure to yield on average that* $FDR \leq 0.1$. *But if* $Pr(R \geq 0) = 0.5$, *then we have actually only controlled* $pFDR \leq 0.2$, *a quantity twice as large.*

## 2.3 Bayesian Interpretation, Setup Under Mixture Models

By formulating the multiple testing problem in a simple Bayesian framework, we are able to construct procedures that control a quantity closely related to the FDR as we have previously defined. We assume that we have n hypotheses, which are null ($H_0$) with probability $\pi_0$ and non-null ($H_1$) with probability $1 - \pi_0$. The corresponding test statistics are $z \sim f_0$ under $H_0$ and $z \sim f_1$ under $H_1$. Thus formally the data has been generated by a heierarchal model : $H \sim Bernoulli(1 - \pi_0), z \sim f_H$;. marginally, the $z$ have a mixture distribution $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$.

Classical Benjamini-Hochberg theory only lets us discuss false discovery rates for tail sets of the open form. An advantage of Bayesian theory is that we can now compute and bound the FDR for generic measurable sets $A$. Efron (2008) likes to distinguish between the "local" and "global" FDR rates as following:

$$Global(typical)FDR : Fdr(z_c) = \psi((-\infty, z_c),$$
$$LocalFDR : fdr(z_c) = \psi(z_c),$$

where fdr(zc) will in general be well-defined provided all distributions have continuous densities. These two quantities can be very different.

Although the above framework assumes a simple versus simple testing situation with a fixed alternative value, as Storey & Tibshirani (2003) mentions, the relationship holds even for simple versus composite testing situation as long as one models the alternative parameter values as a random variable Then $H_1$ is a mixture of the alternative distributions. Thus, even though the problem of multiple testing belongs to the classical frequentist paradigm, the probabilities that one would like to estimate seems more natural to arise in a Bayesian framework. The FDR or pFDR is written in the form of a posterior probability and in case of simple versus composite testing one needs to resort to mixed effect models (Genovese & Wasserman (2002)) or a Bayesian model to incorporate the added variability of the alternative parameter into the analysis of FDR or pFDR. In the article Tang et al. (2007), authors have formulated a mixture model framework for the alternative $P$-value distribution under a simple versus composite testing situation and estimate $\pi_0$ and $F_1$ using a nonparametric Bayesian technique.

## 2.4 Connection to Classification Theory and Optimality under Sparsity

Although motivated by pure testing considerations, the Benjamini–Hochberg FDR controlling procedure Benjamini & Hochberg (1995) has shown remarkable properties as an estimation procedure Donoho et al. (2006). More specifically, it turns out to be adaptive to the amount of signal contained in the data, which has been referred to as "adaptation to unknown sparsity." An extension of the same property of FDR has been proven by Neuvial et al. (2012).
The first analysis of FDR thresholding with respect to the mis-classification risk, an

important theoretical breakthrough has been made by Bogdan et al. (2011). The major contribution of Bogdan et al. (2011) is to create an asymptotic framework in which several multiple testing procedures can be compared in a sparse Gaussian scale mixture model. In particular, they proved that FDR thresholding is asymptotically optimal (as the number $m$ of items goes to infinity) with respect to the mis-classification risk. Also, they proposed an optimal choice for the rate of $\alpha m_0$ as $m_0$ grows to infinity.

# 3 A Simulation Study on FDR Controlling Procedure

As we we were concerned with the following definition of FDR: $FDR = E(\frac{V}{R})$, where $\frac{V}{R} = 0$ when R=0, and $FDR + 1 = E(\frac{V}{R}|R > 0)$ (pFDR in Storey (2002)). We adopted the classical definition of FDR from Benjamini & Hochberg (1995) because controlling it assured weak control of the familywise error rate FWER=Pr(V1) when all hypotheses are true a property that we considered essential for use in medical research, and a property that the other definitions could not enjoy.

## 3.1 The Procedure

Suppose we wish to test $m$ hypotheses, $H_1 \ldots, H_m$, and we have a $p$-value $p^i$ for each hypothesis $H_i$. That is,

$$P_{H_i}(p^i \leq \alpha) = \alpha.$$

According to Benjamini and Hochberg (1995) procedure these $p$-values are independent. This procedure controls the false discovery rate (FDR), which is the expected proportion of the rejected tests that should not have been rejected:

$$\mathbf{E}_{\{H_i:i\in S\}}\left[\frac{\sum_{i\in S} 1\{H_i \quad \text{rejected}\}}{\max[1, \sum_{i=1}^{m} 1\{H_i \quad \text{rejected}\}]}\right].$$

Given a desired FDR $q$, the BH procedure finds a data-adaptive threshold level $\hat{p}(q)$ and rejects all $H_i$ for which $\hat{p}_i \leq \hat{p}(q)$. The threshold level is given by comparing the sorted $p$-values $\hat{p}_{(1)} \leq \cdots \leq \hat{p}_{(n)}$ to a line of slope $q/n$ and identifying the largest $p$-value that is below this line. That is, $\hat{p}(q) = \hat{p}_{(}\hat{i})$ where

$$\hat{i} = \max\{i : \hat{p}_i \leq qi/m\}.$$

In the simulation in Section 3.3, we verify that the BH procedure works.

## 3.2 Distribution of $p$-value

Suppose we were given high-throughput gene expression data that was measured for several individuals in two populations. We are asked to report which genes have different average expression levels in the two populations. If instead of thousands of genes, we

were handed data from just one gene, we could, for example, use a t-test or some other test.

We consider $p$-values are random variables. To see this, consider the example in which we define a $p$-value from a t-test with a large enough sample size to use the CLT approximation. Now we define

$$p = 2\{1 - \Phi(Z)\},$$

where $Z$ is a random variable and $\Phi$ is a deterministic function, $p$ is also a random variable. We create a Monte Carlo simulation using the mice data from Churchill et al. (2012) to imitate a situation 10,000 times showing how the values of $p$ change.

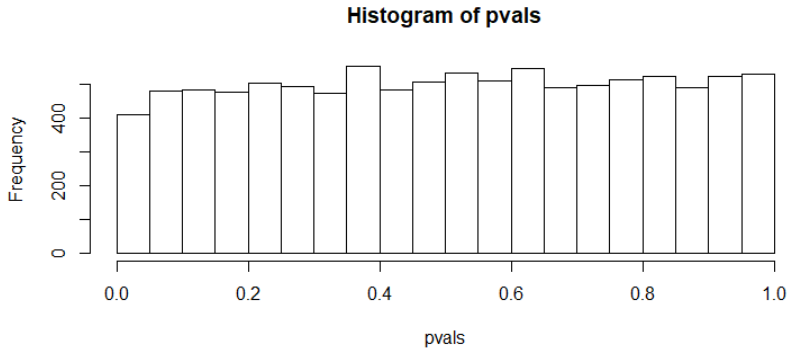


Figure 1: Simulated $p$-values from t-test

The histogram given in Fig. 1 shows that the distribution of the $p$-value is uniformly distributed.

## 3.3 Simulation Study

Throughout this section we will be using the type I error and type II error terminology defined in Table 1 in Section 2. We, also, use the possibilities using the notation from Benjamini and Hochberg (1995) in the given Table 1 in Section 2. We will also refer to them as false positives and false negatives respectively.

To describe the entries in the Table 1 in Section 2, we use, as an example, a data-set representing measurements from 10,000 genes, which means that the total number of tests that we are conducting is: $m = 10,000$. The number of genes for which the null hypothesis is true, which in most cases represent the "non-interesting" genes, is $m_0$, while the number of genes for which the null hypothesis is false is $m_1$. For this we can also say that the alternative hypothesis is true. In general, we are interested in detecting as many as the cases for which the alternative hypothesis is true (true positives), without incorrectly detecting cases for which the null hypothesis is true (false positives). For most high-throughput experiments, we assume that $m_0$ is much greater than $m_1$. For example, we test 10,000 expecting 100 genes or less to be interesting. This would imply that $m_1 \leq 100$ and $m_0 \geq 19,900$.

We use a Monte Carlo simulation using the mice data from Churchill et al. (2012) to

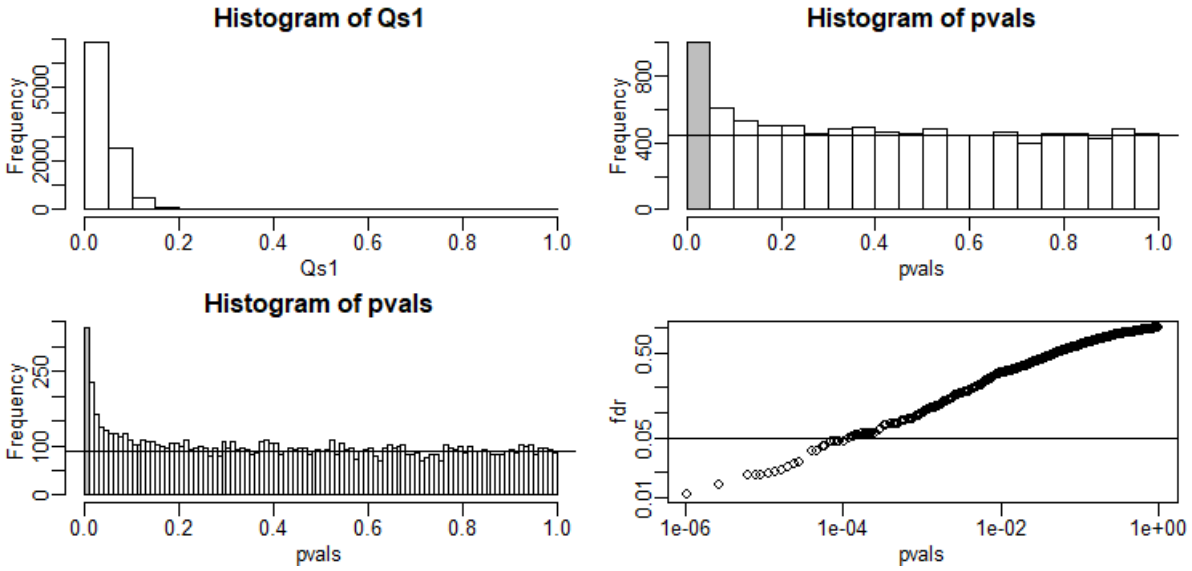


Figure 2: Simulated $p$-values from t-test at different cut-off

imitate a situation in which we perform tests for 10,000 different fad diets, none of them having an effect on weight. This implies that the null hypothesis is true for diets and thus $m = m_0 = 10,000$ and $m_1 = 0$. Let's run the tests with a sample size of $N = 12$ and compute $R$. Our procedure will declare any $p$-value smaller than $\alpha = 0.05$ as significant.

By the definition of FDR mentioned in Eq. (1), we can write

$$Q = \frac{V}{R}$$

with $Q = 0$ when $R = 0$ and $V = 0$. Note that $R = 0$ (nothing called significant) implies $V = 0$ (no false positives). So $Q$ is a random variable that can take values between 0 and 1 and we can define a rate by considering the average of $Q$. To better understand this concept here, we compute Q for the procedure: call everything $p$-value $< 0.05$ significant.

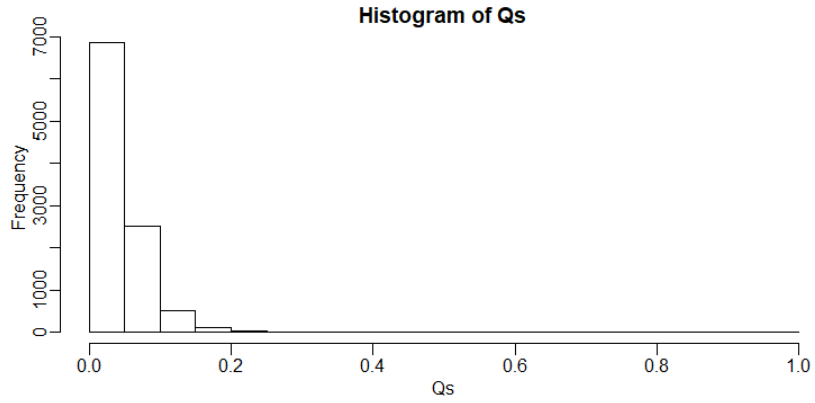


Figure 3: Simulated FDR for Benjamini-Hochberg method

We use Monte Carlo simulation that generates 10,000 experiments 1,000 times, each time saving the observed $Q$. The histogram of these values is given in Fig. 2 (top left).

Top right graph of Fiq. 2 shows that The first bar (grey) on the left represents cases with $p$-values smaller than 0.05. From the horizontal line we can infer that about $1/2$ are

false positives. This is in agreement with an FDR of 0.50. If we look at the bar for 0.01 (bottom left graph of Fig. 2), we can see a lower FDR, as expected, but would call less features significant.

As we consider a lower and lower $p$-value cut-off, the number of features detected decreases (loss of sensitivity), but our FDR also decreases (gain of specificity). So we need to decide on this cut-off. One approach is to set a desired FDR level $\alpha$, and then develop procedures that control the error rate: FDR $\leq \alpha$, which is known as Benjamini-Hochberg proocedure. Benjamini and Hochberg (1995) showed mathematically that this procedure has FDR lower than 5%.

In summary, Monte-Carlo simulation confirmed that the FDR is in fact lower than .05 (simulated value of FDR is 0.03625253, Fiq. 5). Although we will end up with more false positives, FDR gives us much more power. This makes it particularly appropriate for discovery phase experiments where we may accept FDR levels much higher than 0.05.

# 4 An Application of FDR Methodology to Leukemia Dataset

## 4.1 Data Description

Leukemia dataset (learning set) contains gene expression levels (3051 genes and 38 patient samples) from Golub et al. (1999). This dataset has been pre-processed: capping into floor of 100 and ceiling of 16000; filtering by exclusion of genes with max/min<=5 or max-min<=500, where max and min refer respectively to the maximum and minimum intensities for a particular gene across mRNA samples; 2-base logarithmic transformation. These samples include 11 acute myeloid leukemia (AML) and 27 acute lymphoblastic leukemia (ALL) which can be further subtyped into 19 B-cell ALL and 8 T-cell ALL. There are also gene identifiers and tumor class labels (0 for ALL, 1 for AML). In the gene expression data we have-

- **golub.cl:** Numeric vector indicating the tumor class, 27 acute lymphoblastic leukemia (ALL) cases (code 0) and 11 acute myeloid leukemia (AML) cases (code 1).
- **golub.gnames:** a matrix containing the names of the 3051 genes for the expression matrix golub. The three columns correspond to the gene index, ID, and Name, respectively

## 4.2 Statistical Analysis and Results

### 4.2.1 Computing Test Statistics

We compute two-sample t-statistics that compares the gene expressions for each gene in the ALL and AML cases. This can be done with the **mt.teststat** function. The default

test is the two-sample Welch t-test.

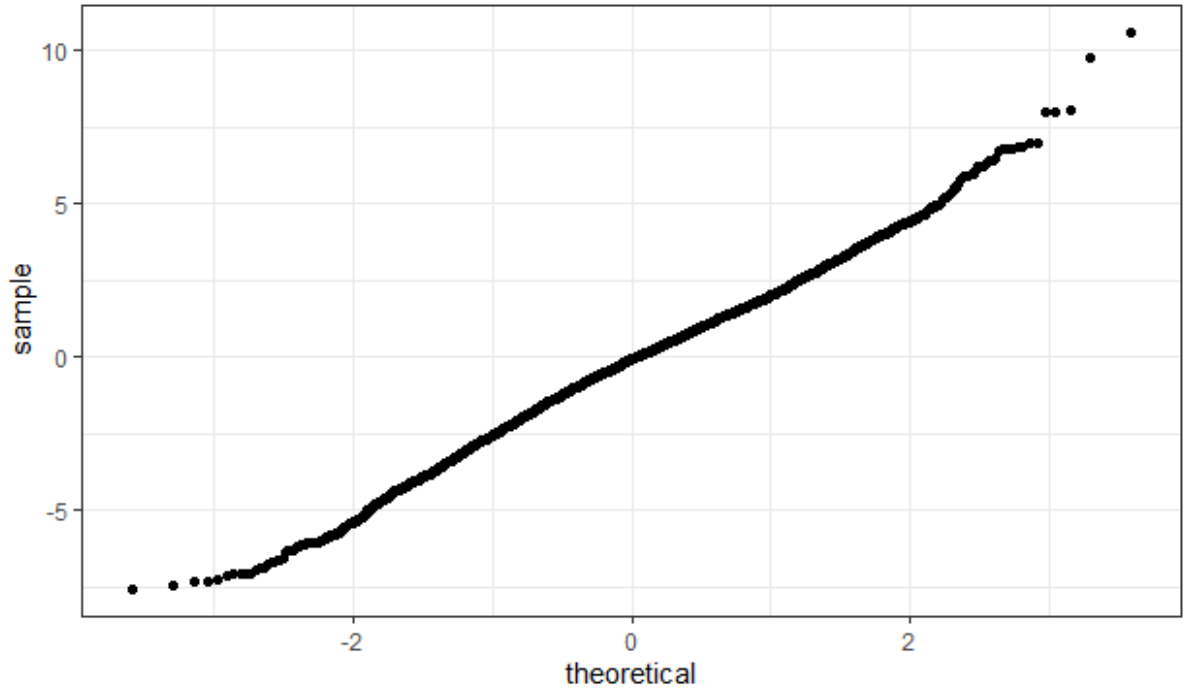


Figure 4: Gene expressions for each gene in the ALL and AML cases

### 4.2.2 Adjusting $p$-values

The **mt.rawp2adjp** function computes adjusted $p$-values for simple multiple testing procedures from a vector of raw (unadjusted) $p$-values. The procedures include the Benjamini & Hochberg (1995) and Benjamini & Yekutieli (2001) procedures for (strong) control of the false discovery rate (FDR). First we will compute raw nominal two-sided $p$-values. For this data, we'll assume that it's safe to use a standard normal distribution for the 3,051 test statistics.

The resulting $p$-values are displayed based on the original data order. Only the first 10 entries are shown in Table 2 .

We can also adjust the $p$-values with permutations using the **mt.maxT()** and **mt.minP()** functions.

From the vignette: we compute permutation adjusted $p$-values for the **maxT** and **minP** step-down multiple testing procedure described in Westfall et al. (1993). These

| Raw $p$-value | Benjamini and Hochberg | Benjamini and Yekutieli |
|---|---|---|
| 0.07854436 | 0.1819581 | 1 |
| 0.36289759 | 0.5354583 | 1 |
| 0.92191171 | 0.9590019 | 1 |
| 0.73463771 | 0.8385259 | 1 |
| 0.17063542 | 0.3187987 | 1 |
| 0.20585260 | 0.3617836 | 1 |
| 0.47019947 | 0.6378740 | 1 |
| 0.59364760 | 0.7374670 | 1 |
| 0.98666904 | 0.9931796 | 1 |
| 0.89267891 | 0.9428699 | 1 |

Table 2: Adjusted $p$-value table in original order

procedure provide strong control of the FWER and also incorporate the joint dependence structure between the test statistics. There are thus in general less conservative than the standard Bonferroni procedure. The permutation algorithm for the **maxT** and **minP** procedures is described in Ge et al. (2003). The $p$-values are sorted in decreasing order of the absolute values of the test statistics: 10.577748, 9.775847, 8.032939, 7.983260, 7.965528, −7.548348.
The functions **mt.sample.teststat** and **mt.sample.rawp** can be used to investigate the permutation distribution of test statistics and raw $p$-values. As for example the following histogram gives a vector of 10,000 permutation test statistics for gene 1.

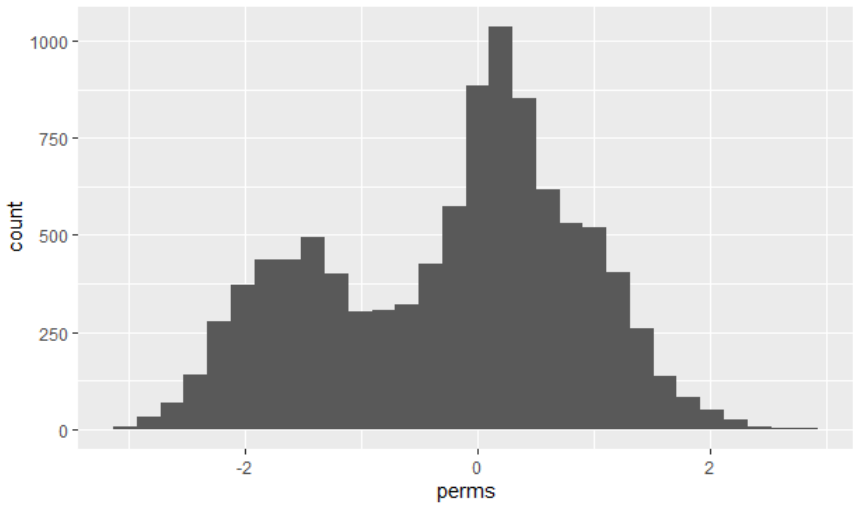


Figure 5: Histogram of 10,000 permutation test statistics for gene 1.

### 4.2.3 Q Value

Note that the FDR is a property of a list of features, not each specific feature. The Q value relates FDR to an individual feature. To define the Q value we order features we tested by $p$-value then compute the FDR for a list with the most significant, the two most significant, the three most significant, etc. The FDR of the list with the, say, m most

significant tests is defined as the Q value of the $m^{th}$ most significant feature. In other words, the Q value of a feature, is the FDR of the biggest list that includes that gene. In Table 3, we display first 10 entries of the calculation of Q value of respective genes:

| Serial Number | Q Value |
|---|---|
| 1 | 0.08314218 |
| 2 | 0.26920003 |
| 3 | 0.48014054 |
| 4 | 0.40918425 |
| 5 | 0.12669612 |
| 6 | 0.15024811 |
| 7 | 0.33520063 |
| 8 | 0.37546226 |
| 9 | 0.49683551 |
| 10 | 0.47249846 |

Table 3: Q value for first 10 genes

In our Leukemia dataset, we calculated that number of genes which achieve an FDR $< 0.05$ is 902. In other words, out of all the genes in the dataset, 902 genes have almost 5% false positives.

### 4.2.4 Controlling FDR

FWER is appropriate one wants to guard against any false positives. However, in many cases (particularly in genomics) we can live with a certain number of false positives. In these cases, the more relevant quantity to control is the FDR.The FDR control has generated a lot of interest due to its more balanced trade-off between error rate control and power than the traditional Family-wise Error Rate control Procedures controlling FDR include Benjamini & Hochberg (1995), Benjamini & Yekutieli (2001) and Benjamini & Hochberg (2000). Here are the steps for Benjamini and Hochberg FDR:
1. Sorting nominal $p$-values from small to big: $p_1 \leq p_2 \leq \ldots \leq p_m$
2. Finding a highest rank of j with $p_j < (j/m) * \delta$ , where $\delta$ is the controlled FDR level.
3. Declaring the tests of rank 1, 2, ..., j as significant, and their adjusted $p$-values as $p_j * m/j$.

Here “Control of the false discovery rate" refers to the expected proportion of false positives (rather than a tail probability) and “Control of FDR at level alpha" means $E(V_n/R_n) \leq \alpha = .05$. We run **MTP** on the first gene in the golub dataset. Table 4 represents the summary of $p$-value.

# 5 Concluding Remarks

The approach to multiple significance testing in this paper is philosophically different from the classical approach of controlling FWER. The classical approach requires the

| | Mean |
|---|---|
| Adjusted $p$-value | 0.06 |
| Unadjusted $p$-value | 0.03 |
| Statistic | 1.7592 |
| Estimate | 0.4923 |

Table 4: Summary of $p$-value

control of the FWER in the strong sense, a conservative type I error rate control against any configuaration of the hypotheses tested. The new approach calls for the control of the FDR instead, and thereby also the control of the FWER in the weak sense. In many applications this is the desirable control against errors originated from multiplicity.
Within the framework suggested, other procedures may be developed, including procedures which utilize the structure of specific problems such as pairwise comparisons in analysis of variance. A different direction, which we pursued here, is to mimic an adaptive method which incorporates the ideas of Benjamini & Hochberg (1995). In this report we have focused on presenting the approaches that calls for controlling the FDR along with simulation and a real genomics dataset and we have demonstrated that it can be developed into a simple and powerful procedure. Thus the cost paid for the control of "true-rejection" does not not need to be large. The FDR procedure thus contributes considerably to the area of hypothesis testing where high rate of rejection of null hypotheses might pose a problem such as but not limited to the area of multiple-comparison problems, classification theory.

# References


Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, *25*(1), 60–83.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 1165–1188.

Bogdan, M., Chakrabarti, A., Frommlet, F., & Ghosh, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, 1551–1579.

Brem, R. B., Yvert, G., Clinton, R., & Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*, *296*(5568), 752–755.

Churchill, G. A., Gatti, D. M., Munger, S. C., & Svenson, K. L. (2012). The diversity outbred mouse population. *Mammalian Genome*, *23*(9-10), 713–718.

Donoho, D., Jin, J., et al. (2006). Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *The Annals of Statistics*, *34*(6), 2980–3018.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 1–22.

Ge, Y., Dudoit, S., & Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, *12*(1), 1–77.

Genovese, C., & Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(3), 499–517.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... others (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531–537.

Hochberg, Y., & Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine*, *9*(7), 811–818.

Neuvial, P., Roquain, E., et al. (2012). On false discovery rate thresholding for classification under sparsity. *The Annals of Statistics*, *40*(5), 2572–2600.

Schweder, T., & Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, *69*(3), 493–502.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*(1), 561–584.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(3), 751–754.

Sorić, B. (1989). Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, *84*(406), 608–610.

Storey, J. D., et al. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, *31*(6), 2013–2035.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*(16), 9440–9445.

Tang, Y., Ghosal, S., & Roy, A. (2007). Nonparametric Bayesian estimation of positive false discovery rates. *Biometrics*, *63*(4), 1126–1134.

Westfall, P. H., Young, S. S., & Wright, S. P. (1993). On adjusting P-values for multiplicity. *Biometrics*, *49*(3), 941–945.

Whittemore, A. S. (2007). A Bayesian false discovery rate for multiple testing. *Journal of Applied Statistics*, *34*(1), 1–9.