*University of California, Berkeley*

U.C. Berkeley Division of Biostatistics Working Paper Series

*Year* 2005 *Paper* 186

# Application of a Multiple Testing Procedure Controlling the Proportion of False Positives to Protein and Bacterial Data

Merrill D. Birkner* Alan E. Hubbard†

Mark J. van der Laan‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, mbirkner@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, hubbard@stat.berkeley.edu

‡Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

# Application of a Multiple Testing Procedure Controlling the Proportion of False Positives to Protein and Bacterial Data

Merrill D. Birkner, Alan E. Hubbard, and Mark J. van der Laan

**Abstract**

Simultaneously testing multiple hypotheses is important in high-dimensional biological studies. In these situations, one is often interested in controlling the Type-I error rate, such as the proportion of false positives to total rejections (TPPFP) at a specific level, alpha. This article will present an application of the E-Bayes/Bootstrap TPPFP procedure, presented in van der Laan et al. (2005), which controls the tail probability of the proportion of false positives (TPPFP), on two biological datasets. The two data applications include firstly, the application to a mass-spectrometry dataset of two leukemia subtypes, AML and ALL. The protein data measurements include intensity and mass-to-charge (m/z) ratios of bone marrow samples, with two replicates per sample. We apply techniques to preprocess the data; i.e. correct for baseline shift of the data as well as appropriately smooth the intensity profiles over the m/z values. After preprocessing the data we show an application of a TPPFP multiple testing techniques (van der Laan et al. (2005)) to test the difference between two groups of patients (AML/ALL) with respect to their intensity values over various m/z ratios, thus indicative of testing proteins of different sizes. Secondly, we will show an illustration of the E-Bayes/Bootstrap TPPFP procedure on a bacterial data set. In this application we are interested in finding bacteria whose mean difference over time points is differentially expressed between two U.S. cities. With both of these data applications, we also show comparisons to the van der Laan et al. (2004b) tppfp augmentation method, and discover the E-Bayes/Bootstrap TPPFP method is less conservative, therefore rejecting more tests at a specific alpha level

# 1 Introduction

## 1.1 Motivation

Simultaneous hypothesis testing is present in various biological applications. Methods have been proposed to address situations of many simultaneous statistical tests (multiple testing). These methods control Type-I error rates in various manners. Original methods, such as the Bonferroni adjustment are extremely conservative, especially as the number of tests increases, which is the case in many genomic settings. Methods have been developed, using either the marginal or joint distribution of the test statistics, to control various Type-I error rates at a specific $\alpha$ level. Multiple testing revolves around developing a procedure which controls the Type-I error rate close to the nominal $\alpha$ level, therefore correctly rejecting the alternative hypotheses.

Multiple testing procedures are based on a variety of Type-I error rates. Some of the popular Type-I error rates include the family wise error rate (FWER), which controls the probability of rejecting more than one false positive; generalized family wise error rate (gFWER), which controls the probability of rejecting more than a user defined number, $k$, false positives; tail probability of the proportion of false positives (TPPFP), which controls the proportion of false positives to total rejections at a user defined value $q$, $q \in (0, 1)$; False Discovery Rate (FDR), or controlling the mean of the proportion of false positives to total rejections. FWER is an extremely conservative method (e.g. Bonferroni), and often too conservative for most biological applications, therefore leading scientists to be interested in methods which will allow some false positives, but at a given number or proportion. A method controlling the TPPFP is attractive especially since it deals with the proportion of false positives to total rejections, instead of an absolute number of false rejections. It will allow some false positives as long as the probability of the proportion of false positives to total rejections is small. Also, as compared to the FDR methods, TPPFP controls the actual proportion of false positives to total rejections, whereas the FDR controls that proportion on average, therefore making a method controlling the TPPFP favorable in some settings, particularly since the expected number of false positives can be highly variable (e.g. when the test statistics are highly dependent).

This article presents two data applications of the E-Bayes/Bootstrap TPPFP approach, outlines in detail in van der Laan et al. (2005). This approach controls the TPPFP at a user defined level $q$, with probability

$1-\alpha$. van der Laan et al. (2005) outlines this procedure and provides finite and asymptotic rational of the proposed procedure, as well as simulations showing the method is more powerful and less conservative in the finite setting, relative to competing TPPFP procedures. Since this method is less conservative, we are apt to properly reject more null hypotheses at a nominal $\alpha$ level as compared to other more conservative methods. In this article, this technique will be applied to two separate datasets, which are described in detail in section 3.

The first application is to an AML/ALL leukemia dataset, in which we are interested in finding proteins which are present (with greater intensity) in one leukemia subtype as compared to the other (AML versus ALL). Mass-spectrometry profiles are often analyzed to determine the differences in the protein profiles (intensity) between the samples. The spectrums display the mass to charge ratio (m/z) versus intensity for each sample and the peaks of the spectrums are compared; the m/z values correspond to different proteins (e.g. depending on their size).

Preprocessing of mass-spectrometry data is necessary in order to correct for phenomenons such as baseline shift and other sources of experimental error. After preprocessing the data, we are interested in applying a multiple testing method to this data in order to determine which m/z ratios are differentially expressed with respect to intensity, between AML and ALL samples. In addition, we want to employ a technique that gives accurate (and not overly conservative) control if these ratios are highly dependent. Thus, we applied the E-Bayes/Bootstrap TPPFP technique which controls the probability that the proportion of false positives, among the rejections, exceeds a user supplied $q$ (e.g. $q = 0.1$), at an $\alpha$ level.

The second application was to bacteria microarray data, which is used to catalog the relative abundance of thousands of types of bacteria in various U.S. cities. Comparing these geographic-specific arrays will therefore allow researches to distinguish between natural occurring bacteria and anomalies occurring in the various cities. We are interested in comparing the mean expression difference over various time points between Austin, TX versus San Antonio, TX. A multiple testing procedure is used to determine which bacterial agents are differentially expressed between the two cites.

# 2 Methods

## 2.1 Multiple Testing Methodology

The E-Bayes/Bootstrap TPPFP method aims to control the proportion of false positives to total rejections at a user defined level $q$, with probability $1 - \alpha$. As discussed in van der Laan et al. (2005), the recently developed, resampling based E-Bayes/Bootstrap TPPFP approach has proven to be less conservative and thus more powerful, as compared to other methods such as the augmentation approach outlined in van der Laan et al. (2004b), and the Lehmann and Romano (2003) tppfp techniques. The procedure involves 1) specifying a conditional distribution for a guessed set of true nulls, given the data, which asymptotically is degenerate at the true set of nulls, and 2) specifying a generally valid null distribution for the vector of test-statistics proposed in Pollard and van der Laan (2003), and generalized in subsequent articles Dudoit et al. (2004), van der Laan et al. (2004a), and van der Laan et al. (2004b). The finite and asymptotic results are outlined in the van der Laan et al. (2005) as well as relevant simulations, which illustrate comparisons of the power and error rate of this procedure in various situations. We will briefly outline the procedure before applying it to the actual datasets, but refer the reader to van der Laan et al. (2005) for a more detailed description of the procedure.

Let $X_1, ...X_n$ be i.i.d. observations and $X \sim P$. We will define $H_{0j}, j = 1, ..., m$ as the $m$ null hypotheses about $P$, $H_{0j} : P \in M_j$. We will define $T_n = (T_n(1), \ldots, T_n(m))$ as the test-statistics corresponding to null hypotheses $H_1, ..., H_m$ for each m/z value or bacterial species, in the respective datasets, with $m$ corresponding to the number of tests performed. This vector of test statistics has an unknown distribution $Q_n$ Given a user supplied $q$ and $\alpha \in (0, 1)$, the procedure selects a common cut-off $c_n$ such that,

$$Pr\left(\frac{\sum_{j=1}^{m} I(T_n(j) > c_n, j \in \mathcal{S}_0)}{\sum_{j=1}^{m} I(T_n(j) > c_n)} > q\right) \leq \alpha,$$

where $j \in \mathcal{S}_0$ indicates a null hypothesis, and $T_n(j) > c_n$ indicates a rejection of $H_{0j}$.

### 2.1.1 E-Bayes/Bootstrap TPPFP Approach

Our method for choosing $c$ involves controlling the tail probability of a random variable $\tilde{r}_n(c)$ defined as:

$$\tilde{r}_n(c) = \frac{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_{0n})}{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_{0n}) + \sum_j I(T_n(j) > c, j \notin \mathcal{S}_{0n})}.$$

$\tilde{r}_n(c)$ represents a guessed proportion of false positives among rejections, defined by drawing a random set $\mathcal{S}_{0n}$ (a guessed set of true null hypotheses $\mathcal{S}_0$) and a draw $\tilde{T}_n$ from a null distribution for the test-statistic vector. We want $\tilde{r}_n(c)$ to dominate in distribution the true proportion of false positives: $\frac{\sum I(T_n(j)>c, j \in S_0)}{\sum I(T_n(j)>c)}$. Clearly the random variable $\tilde{r}_n(c)$ is defined by the proposed definition of $\tilde{T}_n(j)$ and $\mathcal{S}_{0n}$.

**Derivation of $\tilde{T}_n(j)$:**

In order to estimate $\tilde{T}_n$, we bootstrap the data $(X_1^{\#}, \ldots, X_n^{\#})$ $B^*$ times (e.g. $B^* = 10,000$). Each iteration, we recalculate the $m$ test-statistics. This $m \times B^*$ matrix, $\tilde{T}_n^*$, represents a draw from the test-statistic vector under the empirical distribution $P_n$. We then calculate the row-specific means and center the $\tilde{T}_n^*$ matrix at its null value. Each column of this matrix specifies a draw of $\tilde{T}_n = (\tilde{T}_n(j) : j = 1, \ldots, m)$.

**Derivation of $B_n(j) = I(j \in S_{0n})$:**

We will define the distribution of our guessed set of nulls $\mathcal{S}_{0n}$, and describe how this random set is drawn. This random set is defined by drawing a null or alternative status for each of the test statistics. The working model for defining the distribution of the guessed set $\tilde{\mathcal{S}}_{0n}$ will assume $T_n(j) \sim p_0 f_0 + (1 - p_0) f_1$, a mixture of a null density $f_0$ and alternative density $f_1$. Let $B(j)$ represent the underlying Bernoulli random variable, such that $f_0 \sim (T_n(j)|B(j) = 0)$, is the density of $T_n(j)$ if $H_0(j)$ is true, and $f_1 \sim (T_n(j)|B(j) = 1)$ is the density of $T_n(j)$ if $H_0(j)$ is false.

Under this working model, the posterior probability defined as the probability that $T_n(j)$ came from a true $H_{0j}$, given its observed value $T_n(j)$, can now be calculated:

$$P(B(j) = 0|T_n(j)) = p_0 \frac{f_0(T_n(j))}{f(T_n(j))}$$

We will use this posterior probability as the Bernoulli probability on $H_{0j}$ being true, given the test statistic, where we have to specify or estimate $p_0, f_0$ and $f$. Since $f_0$ plays the roll of the density of test-statistics under the null hypothesis, in some situations $f_0$ is simply known: e.g., $f_0 \sim N(0,1)$. However, in cases where the marginal distribution of $T_n(j)$ is not known if $H_{0j}$ is true, one can use a kernel density (**density()** in R with a given kernel and bandwidth) on the mean centered elements in the matrix representing $B$ draws of $\tilde{T}_n$. The elements from this matrix are pooled into a vector of length $m * B^*$ in the kernel density function. In order to estimate the density $f$, we can again apply a kernel smoother on the bootstrapped test statistics, before they are mean centered. Again, the elements of the matrix are pooled into a vector of length $m * B^*$ in the kernel density function.

Finally, $p_0$ represents the proportion of nulls $\mid \mathcal{S}_0 \mid /m$ and typically the user might use a conservative $p_0^*$ for this true proportion of nulls. The most conservative prior, $p_0^* = 1$, will be used throughout this paper. Now, given $T_n$, we can define the random set

$$\mathcal{S}_{0n} = \{j : C(j) = 1\}, C(j) \sim Bernoulli\left(\min\left(1, p_0^* \frac{f_0(T_n(j))}{f(T_n(j))}\right)\right).$$

Given the data $X_1, \ldots, X_n$ (i.e., $P_n$), $\mathcal{S}_{0n}$ and $\tilde{T}_n$ are drawn independently.

We will now draw $(\mathcal{S}_{0n}, (\tilde{T}_n(j))$ $B^*$ times, and each time calculate the corresponding realization of $\tilde{r}_n(c)$, where $T_n$ is fixed at the true original test statistics. This provides us with a sample of $B^*$ realizations of $(\tilde{r}_n^b(c) : c \geq 0)$, $b = 1, \ldots, B^*$, conditional on the data $P_n$ (and thus, conditional on $T_n$ as well).

The cut-off $c$ is set so that the tail probability, at a user supplied level $q$, of the random variable, $\tilde{r}_n(c)$, equals $\alpha$. To do so, we will then choose $c$ such that average over $B^*$ draws of both $\tilde{T}_n(j)$ and $\mathcal{S}_{0n}(j)$ equals $\alpha$.

Specifically, we set

$$c_n = \inf\left\{c : \frac{1}{B^*}\sum_{b=1}^{B^*} I(\tilde{r}_n^b(c) > q) \leq \alpha\right\}.$$

### 2.1.2 Augmentation Technique

An augmentation TPPFP procedure was also applied the multiple testing procedure outlined in Pollard and van der Laan (2003). This augmentation

corresponds to merely adding the $[\frac{q}{1-q}r_0]$ most significant rejections to the rejection set of the FWER method, where $r_0$ is the set of initial rejections form the FWER procedure. As the FWER procedure, we use the single-step maxT based on the resampling-based null distribution $\tilde{T}_n$ described above. Further detail of this method can be found in Pollard and van der Laan (2003).

## 2.2 Adjusted $p$-values

Both the E-Bayes/Bootstrap TPPFP and Augmentation techniques provide adjusted $p$-values as a summary measure for each test. Adjusted $p$-values provide a measure of the probability of making a Type-I error taking into account that one made multiple tests. The $j^{th}$ adjusted $p$-value can be interpreted as the nominal alpha level one would use to just reject the $j^{th}$ specific test-statistic. Displaying these adjusted $p$-values provide a summary measure of the tests and therefore make them easier to compare.

# 3 Data Applications

In the following section, the leukemia and bacteria datasets will be presented, as well as outline the leukemia preprocessing steps. We will then present the application of the E-Bayes/Bootstrap TPPFP approach, as well as the van der Laan et al. (2004b) augmentation technique. Firstly we will describe the datasets, followed by the results of the multiple testing application.

## 3.1 Leukemia: AML/ALL Data

### 3.1.1 Motivation

Leukemia is a form of cancer that originates in the cells in the bone marrow. Leukemia occurs as a result of an excess of abnormal white blood cells in the blood, which are known as leukocytes. The healthy version of these cells are used by the body to defend the body against infectious agents such as viruses and bacteria. But in the case of leukemia, these damaged cells become poor at fighting infection and the abnormal cells multiply excessively and do not die off as they should.

There are more than a dozen varieties of leukemia, and AML and ALL are two of the main varieties. Acute myelogenous leukemia (AML) develops when

there is a defect in the immature cells in the bone marrow. The exact cause of acute leukemias is unknown, but some environmental factors are linked with AML, including exposure to radiation and organic solvents, such as benzene (Greaves, 1997; Bunin, 2004). AML occurs in all ages but are more often prevalent in older adults (Bunin, 2004). Acute lymphocytic leukemia (ALL) symptoms result from the body not producing enough healthy blood cells. An ALL patient's bone marrow makes too many immature white blood cells. Normal blast cells turn into a type of white blood cell called granulocytes, but the leukemia blast cells do not. At the same time, the marrow cannot grow enough normal red blood cells, white blood cells and platelets (Greaves, 2002). Again, a few environmental factors are linked with ALL (Greaves, 1997; Pui et al., 2001).

Researchers are interested in determining the differences between these two types of leukemia varieties, in order to facilitate the treatment of this disease. Proteins have been found to be linked to cancers and especially to leukemia. Researchers can now develop drugs to target the specific protein, thus disabling it's function. We are therefore interested in finding potential biomarkers (proteins) which are of different intensities between the two types of leukemia.

The data which is used in this section is mass-spectrometry data, which consists of mass/charge and intensity values. Mass spectrometry has been a popular tool in the field of proteomics. This field is centered around the identification of proteins in the body and determining their role in processes, such as the transmission of disease. Mass spectrometry is used to identify and quantify proteins from biological samples. The process labels the mass and charge of potential proteins, and their relative abundance in a sample.

### 3.1.2 Data Structure

The data structure consists of two replicates each for 7 samples of AML and 13 samples of ALL. Each sample contained approximately 100 different m/z values and respective intensity values. We are interested in obtaining an intensity value for a specific number of unique m/z values, averaged over the replicates. After preprocessing the data, a test is predetermined for each unique m/z value across the two leukemia samples intensity measurements.

### 3.1.3 Preprocessing

Before the data can be analyzed, preprocessing steps including correcting for the baseline shift, smoothing the mass-intensity profiles, and choosing unique mass values at which to measure the intensity level, are applied. A multiple testing procedure is then applied to the data to determine which mass levels have differentially expressed intensity levels between AML and ALL patients. Note: Examples of the preprocessing steps (in one sample, two replicates) is illustrated in the plots at the end of this paper.

#### Baseline Correction

As previously mentioned, there is often a shift in raw mass-intensity data. Quantile regression was applied to predict the intensity quantile (0.02 quantile) by m/z value for each of the samples. The intensity is then adjusted by subtracting the observed peak from this predicted quantile.

#### Smoothed Intensity

To "smooth" over the error in the m/z estimate per sample, a smoother is used, such as **ksmooth()** in R (using a box kernel), with bounded support (i.e. we expect most m/z ratios should have 0 intensity). The kernel bandwidth is chosen, per biologic sample, by using a simple cross-validation technique on the replicates.

For each bandwidth (1-10 m/z) the smoothing algorithm is trained on one replicate of a biological sample (subject) and used to predict the intensities of its matched replicate. We then reverse the roles of the two replicates and train the smoothing algorithm on the second replicate and test it on the first replicate. The mean squared error is recorded each time the algorithm is trained on the second replicate for each bandwidth. This is then repeated over all samples/replicates. The average MSE is calculated for each bandwidth and the bandwidth with the smallest MSE is chosen, which was 9 in this example. Finally, the original data is reduced to a set of unique m/z ratios (that are non-zero in at least one biological sample).

Finally, after smoothing, the replicate profiles are averaged to get one protein expression/biologic replicate. This processing stream results in a data matrix with 204 unique protein intensities (the rows) for each of the 21

biologic samples (the columns).

### 3.1.4 Application to AML ALL data

The difference in the mean intensities of the AML versus the ALL samples at each of the 204 m/z ratios is tested. The test-statistics will be defined as: $T_n(j) = \sqrt{n}\frac{(\mu^{AML}(j)-\mu^{ALL}(j))}{\sigma_{AML/ALL}(j)}, j = 1, ..., 204$, where $\sigma^2_{AML/ALL}$ is the pooled variance of the two samples. The null hypothesis is that $(\mu_{AML}-\mu_{ALL}) = 0$ and the alternative hypothesis is that $(\mu_{AML}-\mu_{ALL}) \neq 0$. The E-Bayes/Bootstrap TPPFP procedure is used to determine those m/z ratios which have significantly different mean intensities between AML and ALL, while controlling the proportion of false positives to total rejections at a level $q = 0.1$, with probability 0.95 ($\alpha = 0.05$).

### 3.1.5 Results

There are 20 m/z values out of the 204 with an unadjusted $p$-value less than $\alpha = 0.05$. With the tppfp augmentation method no m/z are rejected at an $\alpha = 0.05$ and only one is rejected at an $\alpha = 0.1$ level. The E-Bayes/Bootstrap TPPFP rejects 3 m/z ratios at an $\alpha = 0.05$ and also three are rejected at an $\alpha = 0.1$ level. Interestingly, the proprietary Biomarker Wizard software (Ciphergen Biosystems, Fremont, CA) also found these masses to be significant, based on another algorithm. [These were found through the software's autodetection; therefore anything with a signal to noise ratio greater than 2, the peak had to be present in at least 25 percent of the samples, and the mass window of 0.8 percent mass]. These results illustrate the importance of the E-Bayes/Bootstrap TPPFP method, especially in the cases of few significant associations in the data.

The mass to charge ratios have yet to be identified as unique proteins. However, researchers plan to follow this analysis and ID the most significant mass to charge ratios by SDSPAGE separation and LD MS/MS peptide identification procedures.

## 3.2 Airborne Bacterial Data

### 3.2.1 Motivation

Bacteria are naturally occurring in air and researchers have been interested in various mechanisms to monitor and evaluate the type and concentration

Table 1: Adjusted $p$-values: Top 10 m/z Ratios:

| m/z | E-Bayes/Bootstrap TPPFP ($q = 0.1$) | Augmentation ($q = 0.1$) |
|---|---|---|
| 4968.104 | 0.039 | 0.051 |
| 3333.169 | 0.043 | 0.0595 |
| 4941.165 | 0.0491 | 0.1515 |
| 3201.327 | 0.215 | 0.352 |
| 8457.161 | 0.3197 | 0.437 |
| 3281.276 | 0.3404 | 0.4535 |
| 3908.681 | 0.3586 | 0.460 |
| 2908.314 | 0.3605 | 0.4615 |
| 10527.394 | 0.3897 | 0.467 |
| 10509.961 | 0.3999 | 0.467 |

in the air samples. This is also of interest given the concern of terrorism by biological agents. The Department of Homeland Security, in conjunction with the Lawrence Berkeley Laboratory, Division for Environmental Biotechnology initiated this project. The goal of this project is to catalog thousands of different types of bacteria from cities throughout the United States. The concentration of the bacteria is monitored over several weeks. The goal of this study is to catalog the various natural airborne pathogens. Therefore it can be used as a baseline to compare future levels and it could also help identify disease causing bacteria.

The process of determining the bacteria consists of using a special Affymetrix glass chip. This microarray process quantifies and classifies environmental DNA from a range of prokaryotic and eukaryotic origins. The array has been designed based on 62,358 probes which are matched to both prokaryotic and eukaryotic ribosomal RNA genes (DeSantis et al., 2005). The bacterial DNA is separated, then it is fluorescently labelled and placed on the slide. The more matches to a specific bacteria, the higher the chance that the bacteria is in the air sample. Detailed information regarding the array process and technique can be found in Wilson et al. (May 2002) and DeSantis et al. (2005). After processing, the resulting data for each biological replicate is the expression for each of the 420 bacterial species or samples.

### 3.2.2 Applying Multiple Testing

The dataset analyzed here consists of 17 arrays containing samples collected at different times in San Antonio and Austin, Texas. Thus, the final data matrix is 17 columns, where each entry is the log-base 2 relative expression of a bacterial species (the row) for Austin versus San Antonio for one time point (the column). The test statistics, which we are interested in testing is the if the mean of the difference is equal to 0. Therefore, $T_n(j) = \frac{\sqrt{n}(\mu_{diff}-0)}{\sigma_{diff}}, j = 1, ..., 420$. This is computed for each of the 420 unique bacterium and both the augmentation technique as well as the E-Bayes/Bootstrap TPPFP technique are applied at a $q = 0.1$ and controlling at an $\alpha = 0.05$.

### 3.2.3 Results

Table 2 illustrates the adjusted $p$-values produced by the augmentation technique. The table illustrates that more are rejected with the E-Bayes/Bootstrap TPPFP procedure as compared to the Augmentation techniques at both an $\alpha = 0.05$ and $\alpha = 0.1$. Both of these procedures used $q = 0.1$.

# 4 Discussion

This article presented two separate types of data structures to which the E-Bayes/Bootstrap TPPFP technique, presented in van der Laan et al. (2005), was applied. As previously mentioned, the TPPFP is an appropriate Type-I error rate to control in many biological applications controlling the TPPFP. This error rate is less conservative than the family-wise error rate. The first dataset was comparing two types of leukemia in regards to their differential protein intensity levels. An initial preprocessing technique was applied to the data before the multiple testing procedures. The application of the E-Bayes/Bootstrap TPPFP approach resulted in rejecting more m/z values as compared to the augmentation approach. The bacterial application also elucidated several bacteria that were differentially expressed between the two cities, again with the E-Bayes/Bootstrap TPPFP approach providing more rejections as compared to the augmentation approach. We suggest that both the examples and simulations in the paper as well as the data applications prove that the E-Bayes/Bootstrap TPPFP approach is a more powerful technique to control the proportion of false positives to total rejections at a given level $q$, as compared to various other methods controlling the TPPFP.

Table 2: Adjusted $p$-values of Top Bacteria

| Organism | E-Bayes/Bootstrap TPPFP | Augmentation |
|---|---|---|
| A.ferrooxidans subgroup CtaxTah | 0.00095 | 0.002 |
| Lactobacillus fermentum | 0.00102 | 0.00225 |
| Calyptogena symbionts Calyptogena magnifica | 0.00121 | 0.0025 |
| Catellatospora citrea | 0.001858 | 0.0035 |
| Vr.pantothenticus subgroup compost | 0.00380 | 0.00525 |
| Bacillus alcalophilus | 0.00382 | 0.00525 |
| Dfm.ruminis subgroup | 0.0161 | 0.023 |
| Mlm.methanica subgroup gamma SA51 | 0.019 | 0.029 |
| Pseudonocardia thermophila | 0.0252 | 0.03725 |
| B.cereussubgroup Gram-positive D-Su1-25 | 0.0243 | 0.03725 |
| Bacillus endophyticus | 0.0254 | 0.045 |
| Thermophilic streptomyces Streptomyces | 0.02773 | 0.0735 |
| Klebsiella pneumoniae c3 | 0.03159 | 0.0785 |
| Clostridium beijerinckii | 0.05896 | 0.087 |
| Clostridium collagenovorans | 0.06846 | 0.144 |
| B.cohnii subgroup str. HTA437. | 0.0685 | 0.1575 |
| Microcoleus sociatus | 0.0695 | 0.179 |
| Taxeobacter ocellatus | 0.06998 | 0.18825 |
| Environmental clone iii1-8 group soil clone | 0.0701 | 0.18975 |
| Environmental clone opb45 group soil clone S079 | 0.0701 | 0.18975 |
| Pae.validus subgroup SCBP-S17 | 0.0767 | 0.18975 |
| Myb.tuberculosis subgroup Mycobacterium | 0.0777 | 0.22575 |
| Achromatium assemblage Agricultural soil clone | 0.0806 | 0.229 |

# References


Greta R. Bunin. Nongenetic Causes of Childhood Cancers: Evidence from International Variation, Time Trends, and Risk Factor Studies. *Toxicology and Applied Pharmacology*, 199, 2004.

Todd Z. DeSantis, Carol E. Stone, Sonya R. Murray, Jordan P. Moberg, and Gary L. Andersen. Rapid Quantification and Taxonomic Classification of Environmental Dna from both Prokaryotic and Eukaryotic Origins Using a Microarray. *FEMS Microbiology Letters*, 2005.

Sandrine Dudoit, Mark J. van der Laan, and Katherine S. Pollard. Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL `http://www.bepress.com/sagmb/vol3/iss1/art13`. Article 13.

M. F. Greaves. Aetiology of Acute Leukemia. *The Lancet*, 394, 1997.

Mel Greaves. Childhood Leukemia. *BMJ*, 324, 2002.

Haleem J. Issaq, TImothy D. Veenstra, Thomas P. Conrads, and Donna Felschow. The Seldi-tof ms Approach to Proteaomics: Protein Profiling and Biomarker Identification. *Biochemical and Biophysical Research COmmunications*, 292(3), 2002.

E.L. Lehmann and J.P Romano. Generalizations of the Family-wise Error Rate. Technical report, Department of Statistics, Stanford University, 2003.

Richard J. Q. McNally and Tim O. B. Eden. An Infectious Aetiology for Childhood Acute Leukemia: A Review of the Evidence. *BMJ*, 127, 2004.

Katherine S. Pollard and Mark J. van der Laan. Resampling-based Multiple Testing: Asymptotic Control of Type I error and Applications to Gene Expression Data. Technical Report 121, Division of Biostatistics, University of California, Berkeley, June 2003. URL `http://www.bepress.com/ucbbiostat/paper121`.

Ching-Hon Pui, Dario Campana, and William E. Evans. Childhood Acute Lymphblastic Leukemia- Current Status and Future Perspectives. *The Lancet Oncology*, 2, 2001.

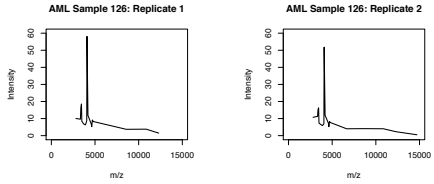
Mark J. van der Laan, Sandrine Dudoit, and Katherine S. Pollard. Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004a. URL `http://www.bepress.com/sagmb/vol3/iss1/art15`. Article 15.

Mark J. van der Laan, Sandrine Dudoit, and Katherine S. Pollard. Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives. Technical Report 1, 2004b. URL `http://www.bepress.com/sagmb/vol3/iss1/art15`. Article 15.

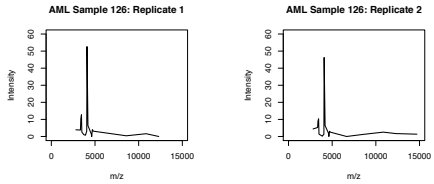
Mark J. van der Laan, Merrill D. Birkner, and Alan E. Hubbard. Resampling Based Multiple Testing Procedure Controlling Tail Probability of the Proportion of False Positives. Technical report, Division of Biostatistics, University of California, Berkeley, March 2005. URL `http://www.bepress.com/ucbbiostat/paper`.

Kenneth H. Wilson, Wendy J. Wilson, Jennifer L. Radosevich, Todd Z. DeSantis, Vijay S. Viswanathan, Thomas A. Kuczmarsk, and Gary L. Andersen. High-Density Microarray of Small-Subunit Ribosomal Dna Probes. *Applied and Environmental Microbiology*, 68(5), May 2002.

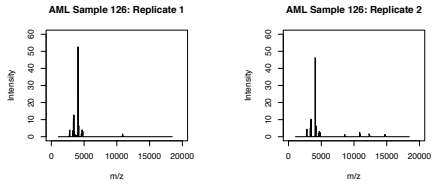
## Raw Data




## Baseline Corrected




## Smoothed Intensity




## Average over Replicates: Smoothed Intensity