*University of California, Berkeley*
U.C. Berkeley Division of Biostatistics Working Paper Series

*Year* 2005 *Paper* 200

# Issues of Processing and Multiple Testing of SELDI-TOF MS Proteomic Data

Merrill D. Birkner[*] Alan E. Hubbard[†] Mark J. van der Laan[‡]
Christine F. Skibola[**] Christine M. Hegedus[††] Martyn T. Smith[‡‡]

[*] 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

[**] Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, chrisfs@berkeley.edu

[††] Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, chegedus@berkeley.edu

[‡‡] Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, martynts@berkeley.edu

# Issues of Processing and Multiple Testing of SELDI-TOF MS Proteomic Data

Merrill D. Birkner, Alan E. Hubbard, Mark J. van der Laan, Christine F. Skibola, Christine M. Hegedus, and Martyn T. Smith

## Abstract

A new data filtering method for SELDI-TOF MS proteomic spectra data is described. We examined technical repeats (2 per subject) of intensity versus m/z (mass/charge) of bone marrow cell lysate for two groups of childhood leukemia patients: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). As others have noted, the type of data processing as well as experimental variability can have a disproportionate impact on the list of "interesting" proteins (see Baggerly et al. (2004)). We propose a list of processing and multiple testing techniques to correct for 1) background drift; 2) filtering using smooth regression and cross-validated bandwidth selection; 3) peak finding; and 4) methods to correct for multiple testing (van der Laan et al. (2005)). The result is a list of proteins (indexed by m/z) where average expression is significantly different among disease (or treatment, etc.) groups. The procedures are intended to provide a sensible and statistically driven algorithm, which we argue provides a list of proteins that have a significant difference in expression. Given no sources of unmeasured bias (such as confounding of experimental conditions with disease status), proteins found to be statistically significant using this technique have a low probability of being false positives.

# 1 Introduction

This study is based on the analysis of array-based proteomic data obtained by surface enhanced laser desorption ionization mass spectrometry (SELDI-TOF MS) of childhood leukemia samples. Two sets of samples of bone marrow cell lysate from children with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) were used, as originally described in Hegedus et al. (2005). Leukemia is a group of cancers characterized by the uncontrolled proliferation of blood precursor cells of the myeloid or lymphoid lineage. ALL and AML are the most common leukemias among children representing approximately 31 percent of cancers in children under 14 years of age (Smith et al., 2005). These leukemias are further classified by immunophenotypic and cytogenetic characteristics. Cases with high hyperdiploidy (greater than 50 chromosomes) and those harboring t(12;21) constitute a majority of childhood ALL (Greaves, 2002). With the exception of a few known risk factors such as benzene and radiation exposure, little is known about the causes of leukemia. Researchers are interested in determining differences in protein expression between leukemic subtypes in order to distinguish between subtypes and investigate possible mechanisms of leukemogenesis. Previous microarray and proteome studies have successfully identified such markers (Golub et al., 1999; Valk et al., 2004; Ohmine et al., 2001; Kohlmann et al., 2004; Yeoh et al., 2002; Ross et al., 2003, 2004; Cui et al., 2004, 2005; Hegedus et al., 2005; Issaq et al., 2002). However, few studies have used SELDI-TOF MS for proteomic analysis of bone marrow from childhood leukemia cases. Here we have used raw data from SELDI-TOF MS analysis of bone marrow described in Hegedus et al. (2005). The bone marrow cell lysate from ALL and AML cases was analyzed to generate data consisting of mass-to-charge ratios (m/z) representing individual proteins and their corresponding intensities, which represent the relative abundance.

This proteomic data is not a straightforward (exact) measurement of underlying protein abundances and is victim to sources of experimental variability (for instance, see Baggerly et al. (2004)), which are a nuisance to finding which proteins are related to the question of interest. As vendors have done (e.g., Ciphergen Biosystems), we provide a series of processing steps that are meant to minimize the sources of nuisance variation. We rely on having technical replicate measures of the samples on a child, which provide a convenient motivation for choosing optimal processing parameters using statistical criteria. Although not discussed in this paper, optimal designs should insure that the data is not confounded by experimental variation, which is most efficiently done by design (for instance, making sure that either experimental conditions are homogenous for all samples or at least samples are evenly distributed across experimental conditions with respect to the factors of interest). Rarely (if ever) can a set of processing techniques overcome unmeasured confounding.

We discuss two classes of data processing/filtering problems that are typical of genomic/proteomic data: pre-processing and selection of proteins of interest by multiple testing. In Section 2 we first discuss our processing algorithm and give some arguments why it should be relatively robust (provide reproducible results) and also suggests augmentations that will make it more flexible. We then follow with a discussion of multiple testing in general and a newly introduced method that provides accurate and yet not overly conservative control for experimentwise (Type I) error rates. We conclude the paper with the analysis of the childhood leukemia data and a short discussion.

# 2 Data Pre-Processing

In this section, we first give the specific structure of the leukemia, SELDI-TOF MS spectral proteomic data for our childhood leukemia subjects. We then discuss how this structure can be utilized for optimally smoothing the intensity vs. m/z data to derive summary intensity measurements for each child for a common set of m/z values.

## 2.1 Data Structure

The dataset consists of two replicates each of AML ($n = 7$) and ALL ($n = 13$). 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.

## 2.2 Background Drift Correction

For this type of proteomic data, there is often a drift in the apparent background values in raw m/z-intensity data (see top row of Figure 1 as an example). Optimally, we would like the minimum value for all non-peak m/z values to be at 0. In addition, a procedure should take advantage of the smoothness (in our case, the background declines in a linear fashion). That is, a reasonable low-dimensional model can be fit to this minimum. Our solution is to use quantile regression, which models the trend in the $p^{th}$ quantile of an outcome versus a predictor variable(s) (Koenker and Bassett, 1978): $F_{Y|X}^{-1}(p \mid X = x) = g(x \mid \beta)$, where $X$ is the explanatory variable, $p$ is the percentile $\in (0, 1)$, $F_{Y|X}(y \mid X = x) \equiv P(Y \leq y \mid X = x)$ and $g(x \mid \beta)$ is some function of $x$ and coefficients, $\beta$, for instance, $g(x \mid \beta) = \beta_0 + \beta_1 x$, $X$ is the explanatory variable (in our case, $m/z$) and $Y$ the outcome (intensity). One can not model the minimum, so we have chosen a very small quantile ($p = 0.02$) and in our case, we have modelled the background as a linear decline, but in practice models of arbitrary complexity can be applied, e.g., a high order polynomial basis. The background corrected intensities are simply $Y - (\hat{\beta}_0 + \hat{\beta}_1 X)$, where $Y$ is the original intensity and $X$ is the corresponding $m/z$ ratio. The results are shown on the second line of Figure 1. Because this procedure can borrow information from adjacent $m/z$ values when determining the baseline correction at a particular $m/z$, and because the baseline drift is typically quite smooth, this procedure should in theory provide a relatively robust method for baseline-correction.

## 2.3 Smoothed Intensity

Because one can assume that the observed profile of intensity versus $m/z$ has both an element of signal (the "true" profile) and noise, filtering can help to reduce the latter. Our method of filtering takes advantage of the replicate nature of our design. Therefore, each biologic replicate is analyzed twice resulting in two protein spectra per child. We want to filter, or smooth the data in a way that emphasizes reproducible peaks, and does the opposite for features that are unique to only one sample. To do so, we use an estimate of the underlying true (noiseless) $m/z$ curve on one sample to predict intensities on the other sample. To form this estimate we used a rectangular kernel smoother (Härdle, 1990). These smoothers estimating the curve at a particular point can be thought of as a simple, local weighted average of the intensities in a small neighborhood of $m/z$ ratios defined by the width of the neighborhood, referred to as the bandwidth. The nature of the weight (that is, how the weighted average declines with distance from the estimation point) is called a kernel. For very smooth functions, a typical kernel might be a Gaussian function and the bandwidth is thus the standard deviation of this function. However, our function is not smooth, but consist of a set of unpredictable peaks surrounded by flat areas with nearly no signal. Thus, a natural choice of a kernel is a simple, uniform weight over a small box or rectangular kernel. The width of the box is the bandwidth, and presumably the width chosen will represent the measurement error on the $m/z$ axis. Note, we used the function **ksmooth()** in R (using a box kernel) to estimate the kernel smooth.

The next problem is to choose the bandwidth, and that is where the replicate samples become very useful. We invoke recently developed theory for the optimality of cross-validation for choosing the "best" estimator from a set of candidate estimators. The kernel bandwidth is chosen by using a simple cross-validation technique on the replicates that attempts to minimize the mean-squared error of prediction. Specifically, the smoothing algorithm is trained, with a specific bandwidth, on one replicate of a biological sample (subject) and is 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 (MSE) 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 average MSE is chosen. In the case of the constant bandwidth method, the bandwidths of 1-10 m/z were tested and the minimum MSE was obtained with a bandwidth equal to 9.

### 2.3.1 Variable Bandwidth

To make the procedure more flexible, we also consider a cross-validation based model selection routine that allows the bandwidth used from smoothing intensities to vary by $m/z$ value, based on the fact that the error in $m/z$ might not be constant, but itself have some drift. Although more complicated models can be used and also compete with simpler models, the simplest being a constant bandwidth as discussed above, we choose to examine bandwidths that changed linearly with $m/z$ value:

$$h = \beta_0 + \beta_1 m/z$$

where h is the bandwidth, and ($\beta_0$ and $\beta_1$) define the model. Both $\beta_0$ and $\beta_1$ are now chosen by cross-validation over a grid of possible values that include:

$\beta_0 = (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$

$\beta_1 = (0.00001, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.3).$

For each combination of $\beta_0$ and $\beta_1$ the smoothing algorithm is trained on one replicate of a biological sample (subject) and used to predict the intensities of its matched replicate, just as done above (the constant bandwidth choice is simply fixing $\beta_1 = 0$ and selecting over the $\beta_0$. 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 combination of $\beta_0$ and $\beta_1$. This is then repeated over all samples/replicates. The average MSE is calculated for each combination of $\beta_0$ and $\beta_1$ and the combination with the smallest MSE is chosen, in our case ($\beta_0 = 2, \beta_1 = 0.009$). The m/z vales are averaged within windows corresponding to the respective bandwidth. Finally, the original data is reduced to a set of unique m/z ratios. Although in our discussion focuses on a simple linear model (linear in m/z), competing models of greater complexity can be used (e.g., higher order polynomials) and the same technique can be used to choose the respective parameters of the model using cross validation.

## 2.4 Defining Peaks and Creating Protein Expression Data Matrix

The original data is reduced to a set of unique m/z ratios (that are non-zero in at least one biological sample) and this is done by condensing any set of unique m/z numbers within the chosen the bandwidth to a single value (the average of them) and also averaging all peak values on the same sample to get a single intensity per unique $m/z$. Finally, the two technical replicates for each biological sample are averaged to get a single set of intensities for a unique and common set of $m/z$ values.

Finally, after smoothing, the replicate profiles are averaged to get one protein expression/biologic replicate. For the leukemia data, the processing stream, using a constant bandwidth, results in a data matrix with 204 unique protein intensities (the rows) for each of the 20 biologic samples (the columns). The respective preprocessing steps are illustrated in Figure 1, which correspond to one AML biological sample

# 3 Multiple Testing

The final step after creating a data matrix that consists of processed protein intensites for each independent biological sample (the columns) and each unique $m/z$ value is to select the $m/z$ values that are significantly associated with some phenotypic trait. In our case, we are interested in associating the intensities of each $m/z$ value with the type of leukemia (ALL vs. AML). We want to chose those proteins for which we have relatively high confidence that they are truly different (i.e., different in mean intensities) between the two groups, using a multiple testing procedure (MTP). In general, MTP procedures consist of 1) choosing an appropriate parameter of interest (e.g., mean difference in intensities in the two groups); 2) specifying the null hypothesis that relates this parameter to the question of interest (e.g., the mean difference is 0); 3) specifying the test statistic for which the null distribution is known, at least asymptotically (e.g., the two-sample t-statistic); 4) performing the test for each row ($m/z$ value); 5) choosing an appropriate experimentwise error rate to control (e.g., the number of false positives or Type I errors); and 7) choosing the method to control this rate (e.g., Bonferroni). The parameters of interest, resulting null hypotheses, test statistic and Type I error rate are choices for the investigator. Once these are chosen, one can debate the merits of various MTPs, specifically, which provide accurate Type I error control under assumptions the investigator is willing to make and 2) among these, which have the greatest power.

There are several Type-I error rates: 1) The family wise error rate (FWER), which controls the probability of rejecting more than one false positive; 2) generalized family wise error rate (gFWER), which controls the probability of rejecting more than a user defined number, $k$, false positives; 3) tail probability of the proportion of false positives (TPPFP), which controls the proportion of false positives to total rejections at a user defined value $q$, $q \in (0, 1)$; 4) False Discovery Rate (FDR), or controlling the mean of the proportion of false positives to total rejections. FWER is a conservative error rate, and often too conservative for most biological applications; thus less stringent methods which will allow some false positives, but at a given number or proportion, may be more conducive to scientific application. 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 a data application of the E-Bayes/Bootstrap TPPFP approach, outlined 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 rationale 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 4.

## 3.1 TPPFP

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.

### 3.1.1 Augmentation Technique

An augmentation TPPFP procedure was also applied to the protein dataset (van der Laan et al., 2004b). 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 from the FWER procedure. As with 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).

## 3.2 Adjusted $p$-values

A convenient way to display the results of a MTP is by reporting the adjusted $p$-values in a ordered list corresponding to their relative significance. 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 makes them easier to compare.

# 4 Data Applications

In the following section, 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 results of the multiple testing application to the dataset preprocessed with the constant bandwidth method. This will be followed with the variable bandwidth results.

## 4.1 Application to AML/ALL data: Constant Bandwidth Preprocessing

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

Table 1: Constant Bandwidth: 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 |

$(\mu_{AML} - \mu_{ALL}) = 0$ 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 ($\alpha = 0.05$).

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, USA) also found these masses to be significant, based on another algorithm, not accounting for multiple testing. 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 identify the most significant mass to charge ratios by purification and MS/MS.

Table 2: Variable Bandwidth: Adjusted $p$-values; Top 6 m/z Ratios:

| m/z | E-Bayes/Bootstrap TPPFP ($q = 0.1$) | Augmentation ($q = 0.1$) |
|---|---|---|
| 4967.375 | <0.0001 | 0.001 |
| 3336.293 | 0.051 | 0.089 |
| 2908.006 | 0.092 | 0.122 |
| 3201.008 | 0.156 | 0.291 |
| 5174.152 | 0.171 | 0.312 |
| 9956.193 | 0.238 | 0.340 |

## 4.2 Application to AML/ALL data: Variable Bandwidth Preprocessing

The variable bandwidth preprocessing and multiple testing procedures were also applied to the same AML/ALL dataset used with the constant bandwidth method. The test statistics are created in the same manner, with the only difference being the preprocessing steps. In total, there are 109 m/z ratios which are tested between the AML and ALL samples.

There are 9 m/z values out of the 109 with an unadjusted $p$-value less than $\alpha = 0.05$. With the tppfp augmentation method one m/z is rejected at an $\alpha = 0.05$ and two are rejected at an $\alpha = 0.1$ level. The E-Bayes/Bootstrap TPPFP rejects two m/z ratios at an $\alpha = 0.05$ and also four are rejected at an $\alpha = 0.1$ level. The results are displayed in Table 2. Similarly with the previous example, the m/z ratios which are found to be significant with this procedure are also found significant using the Biomarker Wizard® software, though this software does not adjust for the multiplicity of the tests performed. (Note that the variable bandwidth method as compared to the constant bandwidth method has proven to be more consistent in finding peaks similar to those peaks found using the Biomarker Wizard® software (Ciphergen Biosystems, Fremont, CA, USA), with the accuracy increasing as m/z increases. This can be attributed to the fact that the mass accuracy of the machine is dependent on the mass of the protein and therefore the variable bandwidth method incorporates the mass value when determining the appropriate window over which to average).

# 5 Discussion

Unless one knows the true underlying data-generating mechanism for their particular technology and experimental design, it is hard to argue that one set of processing steps yields universally superior results relative to a competitor. However, we have proposed a series of processing steps and multiple testing procedures that are flexible, take advantage of technical replicates and have some optimal properties (e.g., cross-validation for bandwidth selection and the empirical Bayes approach for controlling TPPFP). In addition, the TPPFP is an appropriate Type-I error rate to control in many biological applications. This error rate is less conservative than the family-wise error rate. The application of the E-Bayes/Bootstrap TPPFP approach resulted in rejecting more m/z values as compared to the augmentation approach. We suggest that the applied example as well as the simulations presented in van der Laan et al. (2005) demonstrate 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. Finally, the significant m/z values found with this analysis were also seen as significant peaks using the Biomarker Wizard® software, though the latter procedure does not take into account the multiplicity of the tests being performed. Again, we can not argue for the universal optimality of our approach, but it has both worked well in practice, has theoretical justification and controls for multiple testing without being overly conservative.

## Acknowledgements

Merrill D. Birkner is a trainee of the U.C. Berkeley Superfund Basic Research program. We thank Professor Patricia Buffler, Principal Investigator of the Northern California Childhood Leukemia Study, for access to the biological samples. This study was supported by NIH grant P42ES04705.

# References

K.A. Baggerly, J.S. Morris, and K.R. Coombes. Reproducibility of SELDI-TOF Protein Patterns in Serum: Comparing Datasets from Different Experiments. *Bioinformatics*, 22;20(5):777–785, 2004.

J.W. Cui, J. Wang, K. He, B.F. Jin, H.X. Wang, W. Li, L.H. Kang, M.R. Hu, H.Y. Li, M. Yu, B.F. Shen, G.J. Wang, , and X.M. Zhang. Proteomic Analysis of Human Acute Leukemia Cells: Insight into their Classification. *Clinical Cancer Research*, 10(20):6887–6896, 2004.

J.W. Cui, J. Wang, K. He, B.F. Jin, H.X. Wang, W. Li, L.H. Kang, M.R. Hu, H.Y. Li, M. Yu, B.F. Shen, G.J. Wang, and X.M. Zhang. Two-dimensional Electrophoresis Protein Profiling as an Analytical Tool for Human Acute Leukemia Classification. *Electrophoresis*, 26(1):268–279, 2005.

S. Dudoit, M. J. van der Laan, and K. 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.

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 386(5439): 531–537, 1999.

M. Greaves. Childhood Leukaemia. *British Medical Journal*, 324(7332):283–287, 2002.

W. Härdle. *Applied Nonparametric Regression*. Economic Series Monographs, No. 19. Cambridge University Press, Cambridge, U.K., 1990.

C.M. Hegedus, C.F. Skibola, L. Zhang, R. Shiao, S. Fu, E.A. Dalmasso, C. Metayer, G.V. Dahl, P.A. Buffler, and M.T. Smith. Proteomic Analysis of Childhood Leukemia. *Leukemia*, 19:1713–1718, 2005.

H.J. Issaq, T.D. Veenstra, T.P. Conrads, and D. Felschow. The SELDI-TOF MS Approach to Proteomics: Protein Profiling and Biomarker Identification. *Biochemical and Biophysical Research Communications*, 292(3): 587–592, 2002.

R. Koenker and G.S. Bassett. Regression Quantiles. *Econometrica*, 46:33–50, 1978.

A. Kohlmann, C. Schoch, S. Schnittger, M. Dugas, W. Hiddemann, W. Kern, and T. Haferlach. Pediatric Acute Lymphoblastic Leukemia (ALL) Gene Expression Signatures Classify an Independent Cohort of Adult ALL Patients. *Leukemia*, 18(1):63–71, 2004.

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

X. Ma, P.A. Buffler, R.B. Gunier, G. Dahl, M.T. Smith, K. Reinier, and P. Reynolds. Critical Windows of Exposure to Household Pesticides and Risk of Childhood Leukemia. *Environmental Health Perspective*, 110(9): 955–960, 2002.

K. Ohmine, J. Ota, M. Ueda, S. Ueno, K. Yoshida, Y. Yamashita, K. Kirito, S. Imagawa, Y. Nakamura, K. Saito, M. Akutsu, K. Mitani, Y. Kano, N. Komatsu, K. Ozawa, and H. Mano. Characterization of Stage Progression in Chronic Myeloid Leukemia by DNA Microarray with Purified Hematopoietic Stem Cells. *Oncogene*, 20(57):8249–8257, 2001.

K. S. Pollard and M. 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`.

M.E. Ross, X. Zhou, G. Song, S.A. Shurtleff, K. Girtman, W.K. Williams, H.C. Liu, R. Mahfouz, S.C. Raimondi, N. Lenny, A. Patel, and J.R. Downing. Classification of Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profiling. *Blood*, 102(8):2951–2959, 2003.

M.E. Ross, R. Mahfouz, M. Onciu, H.C. Liu, X. Zhou, G. Song, S.A. Shurtleff, S. Pounds, C. Cheng, J. Ma, R.C. Ribeiro, J.E. Rubnitz, K. Girtman, W.K. Williams, S.C. Raimondi, D.C. Liang, L.Y. Shih, C.H. Pui, and J.R. Downing. Gene Expression Profiling of Pediatric Acute Myelogenous Leukemia. *Blood*, 104(12):3679–3687, 2004.

M.T. Smith, C.M. McHale, J.L. Wiemels, L. Zhang, J.K. Wiencke, S. Zheng, L. Gunn, C.F. Skibola, X. Ma, and P.A. Buffler. Molecular Biomarkers for the Study of Childhood Leukemia. *Toxicology and Applied Pharmacology*, 206(2):237–245, 2005.

P.J. Valk, R.G. Verhaak, M.A. Beijen, C.A. Erpelinck, S. Barjesteh van Waalwijk van Doorn-Khosrovani, J.M. Boer, H.B. Beverloo, M.J. Moorhouse, P.J. van der Spek, B. Lowenberg, and R. Delwel. Prognostically Useful Gene-expression Profiles in Acute Myeloid Leukemia. *New England Journal of Medicine*, 350(16):1617–1628, 2004.

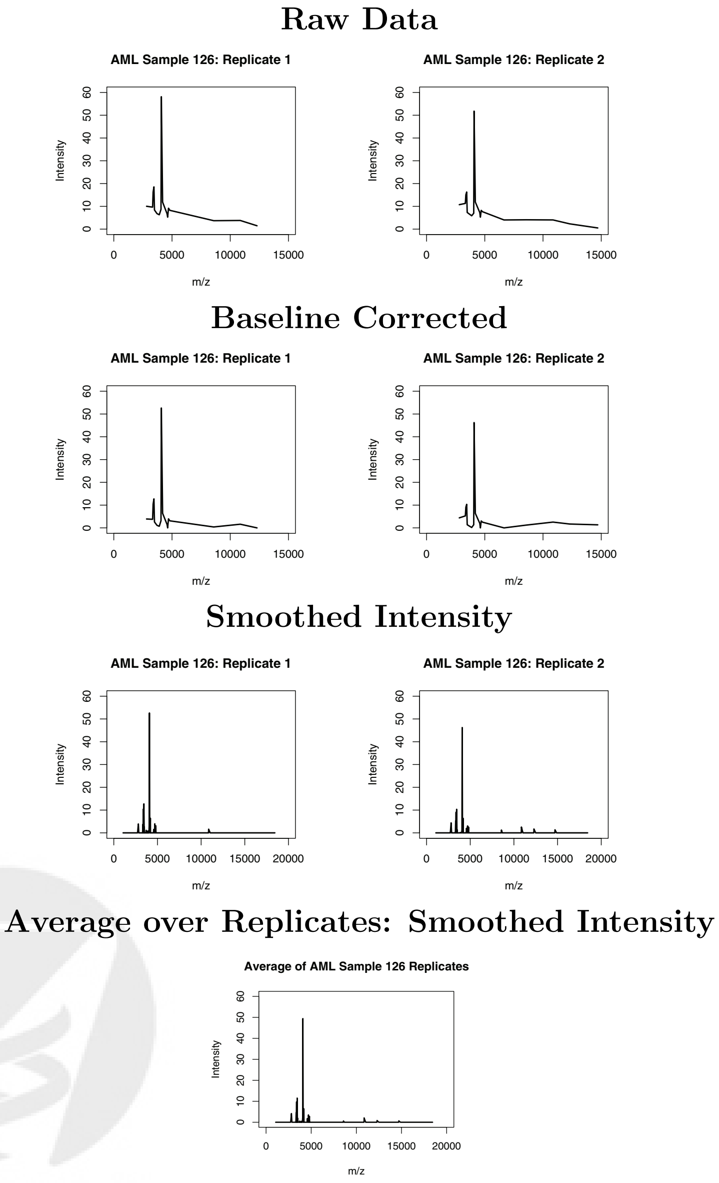
M. J. van der Laan, S. Dudoit, and K. S. Pollard. Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004a. URL `http://www.bepress.com/sagmb/vol3/iss1/art15`.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

M. J. van der Laan, S. Dudoit, and K. 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.

M. J. van der Laan, M. D. Birkner, and A. E. Hubbard. Resampling Based Multiple Testing Procedure Controlling Tail Probability of the Proportion of False Positives. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. URL `http://www.bepress.com/sagmb/vol4/iss1/art29`. Article 29.

E.J. Yeoh, M.E. Ross, S.A. Shurtleff, W.K. Williams, D. Patel, R. Mahfouz, F.G. Behm, S.C. Raimondi, M.V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.H. Pui, W.E. Evans, C. Naeve L. Wong, and J.R. Downing. Classification, Subtype Discovery, and Prediction of Outcome in Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profiling. *Cancer Cell*, 1(2):133–143, 2002.

Figure 1: **Preprocessing Steps**



**Raw Data**

**Baseline Corrected**

**Smoothed Intensity**

**Average over Replicates: Smoothed Intensity**