Incorporating feature reliability in false discovery rateestimation improves statistical power to detect differentially expressed features

Elizabeth Chong, Yijian Huang, Hao Wu
Department of Biostatistics and Bioinformatics
Emory University
Atlanta, GA, USA

Tianwei Yu*
Department of Biostatistics and Bioinformatics
Emory University
Atlanta, GA, USA
Email: tianwei.yu@emory.edu

Nima Ghasemzadeh, Karan Uppal, Arshed A. Quyyumi, Dean P. Jones
Department of Medicine
Emory University
Atlanta, GA, US

Abstract—Feature selection is a critical step in translational omics research. False discovery rate (FDR) is anintegral tool of statistical inference in feature selection from high-throughput data. It is commonly used to screen features (SNPs, genes, proteins, or metabolites) for their relevance to the specific clinical outcome under study. Traditionally, all features are treated equally in the calculation of false discovery rate. In many applications, different features are measured with different levels of reliability. In such situations, treating all features equally will cause substantial loss of statistical power to detect significant features. Feature reliability can often be quantified in the measurements. Here we present a new method to estimate the local false discovery rate that incorporates feature reliability. We also propose a composite reliability index for metabolomics data. Combined with the new local false discovery rate method, it helps to detect more differentially expressed metabolites that are biologically meaningful in a real metabolomics dataset.

Keywords— false discovery rate, high-throughput data, genomics, metabolomics, feature selection, reliability score.

I. INTRODUCTION

In the analysis of high-throughput biological data, e.g. gene expression, proteomics, and metabolomics data, tens of thousands of features (genes, proteins, metabolites etc) are tested simultaneously for their association with certain clinical outcomes under study. This creates the well-known multiple testing problem and causes trouble in statistical inference and data interpretation. The concept and estimation procedures of false discovery rate (FDR) was developed to address this issue [1, 2], which provides a sound statistical framework for inference and feature selection. The FDR is the expected proportion of falsely rejected null hypotheses, i.e. false discoveries, among all features called significant. The concept of local false discovery rate (lfdr) went one step further to give a statistical statement at the single feature level [3], i.e. the probability a specific feature being null given the test statistics of all features in the study.

Over the years, a number of estimation procedures were developed for FDR and lfdr[2, 4-11]. Much effort has been invested in the estimation of the null distribution and proportion of differentially expressed features. Although different modeling approaches were used, all the methods share some common theme – the features are treated equally, certain statistics or p-values are computed for each feature, and the false discovery rates are computed based on the estimation of the distribution of null density from the observed test statistics or p-values.

In many high-throughput datasets, features are measured at different reliability levels. Here by "reliability" we refer to the confidence level we have on the point estimates of the expression values of a feature. In statistical terms, it can mean the size of the confidence interval relative to the measured values, which has a direct bearing on the statistical power to detect differential expression of the feature. In some other situations, it can also mean the probability that a detected feature is real (v.s. pure noise), either based on the measured values or some external information.

When different features are measured with different reliability, subjecting all features to the traditional false discovery rate procedures may yield sub-optimal results. We present two examples here. The first is detecting differentially expressed genes using RNA-seq data.Some genes are measured with low total read counts. For such genes, the measurement reliability, as well as the statistical power of detecting their differential expression is limited. As a result,

* Correspondence Author.
This research was partially supported by NIH grants P20HL113451 and U19AI090023.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

their p-values cannot reach very low when robust testing procedures are used[12-14]. When a false discovery rate procedure is applied to the test results of all genes, the low-read count genes mostlycontribute to the null (non-differentially expressed) distribution. Involving both high-read count and low-read count genes in the FDR or lfdr procedure will reduce the significance level of all the genes.

The second example is more extreme. In LC/MS metabolomics data, features are detected based on the data point patterns in the three dimensional space of mass-to-charge ratio (m/z), retention time (RT) and signal intensity [15, 16]. Certain signal to noise ratio (S/N) threshold and peak shape models are applied. The number of features detected relies on the stringency of the peak detection criterion. There is a trade-off between the detection of noise as features versus the loss of true features with low intensities [17]. Often lenient thresholds are used in order to capture as many real features as possible, and as a result, a large number of features are detected. Presumably some of them are derived from pure noise. The hope is such features will be filtered out in the feature selection process. However, the presence of such false features, for which we do not know how many there are, reduce the significance of all the features in false discovery rate calculation.

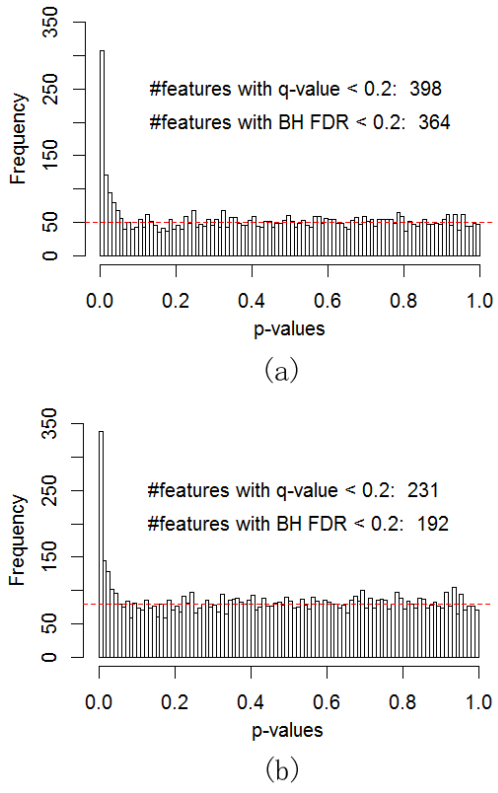

Fig. 1. Illustration of the impact of noise features on the calculation of false discovery rate using simulated p-values. P-values of 5000 non-differentially expressed features and 400 differentially expressed features were generated from the uniform distribution and exponential distribution respectively. The histograms of p-values are shown. Red dashed line: the part of histogram corresponding to the non-differentially expressed genes. (a) Without noise features. (b) With an additional 3000 noise features, whose p-values follow the uniform distribution.

A simple illustration of this issue is presented in Figure 1. In this simple simulation, we demonstrate the effect of involving pure noise features on the FDR adjustment. We simulated p-values of non-differentially expressed features from the uniform distribution, and the p-values of differentially expressed features from an exponential distribution. We applied two widely used FDR approaches – the Benjamini-Hochberg procedure [1], and the Storey q-value procedure [2]. As shown in Figure 1a, when no pure noise genes are present, ~400 features are claimed significant at the FDR level of 0.2, which is close to the hidden truth (Fig. 1a). However when pure noise features are present, they contribute to the null distribution, i.e. the uniform distribution in this case, and make all features less significant. Only ~200 features can be claimed significant at the FDR level of 0.2 (Fig. 1b).

Although the involvement of pure noise features is an extreme scenario which is only relevant in metabolomics data, similar effects can be caused by features measured with low reliability, e.g. low read count genes in RNAseq data, probesets with highly variant probe intensities in microarray data, and proteins with few matched peptides in proteomics data. Often the reliability of features can be partially quantified, not in rigorous statistical terms, but with good heuristic approximation that makes intuitive sense. For example, the log total read count of a gene can be used for RNAseq data.

In this study, we focus on metabolomics data, which has the biggest feature reliability issue among all omicsdata types. Several quantities can be used to indicate how reliable a metabolic feature is. They include the percentage of missing values, the magnitude of the signal, and within-subject variation when technical repeats are available.Using these quantities, we develop a composite reliability index for metabolomics data. Once the reliability is quantified, we devise a new lfdr procedure to incorporate reliability for better lfdr estimation. In simple terms, the method amounts to stratifying features based on their reliability levels. Each feature is compared to the null distribution derived fromall the features with similar reliability level to obtain the lfdr values. However no hard threshold is used. Rather, the null density is computed based on permutation without changing the reliability indices. Our estimation procedurebears some resemblance to themulti-dimensional local fdr by Ploner et al[8]. However it is more efficient, and the explicit use of reliability index makes it more useful than picking arbitrary pairs of test statistics.

II. METHODS

A. The local fdr procedure

Following the consensus of the local fdr literature, we consider the density of the test statistic:

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

$$f(z) = \pi_0 f_0(z) + (1-\pi_0) f_1(z), \quad (1)$$

where f is the mixture density for the observed statistic $Z = z$, f_0 and f_1 are the respective densities of the test statisticof thenull (non-differentially expressed) and non-null (differentially expressed) metabolic features, and π_0 is the proportion of true null features.

The lfdr is then defined as

$$\text{fdr}(z) = \pi_0 \frac{f_0(z)}{f(z)} \quad (2)$$

at observed test statistic $Z = z$ and z is a k-dimensional statistic.

In this study, the test statistics were obtained from a metabolome-wide association study (MWAS). To identify metabolic features whose expression levels are associated with a certain clinical outcome or risk factor, simultaneous hypothesis testing is carried out. We use linear models with log-intensity of the features as the dependent variable, and the risk factor as the independent variable, adjusting for other confounders, e.g. age, gender, ethnicity and experimental batch effect. The regression is conducted one metabolic feature at a time. After obtaining the t-statistic and the corresponding p-values of all the metabolic features, different FDR and lfdr procedures can be applied to select significant metabolic features.

Here, we compare the 1-dimensional ($k = 1$; fdr1d) and 2-dimensional ($k = 2$; fdr2d) local fdr procedures. Density estimation is done non-parametrically.The fdr1d only uses the t-statistic (t) from the simultaneous hypothesis testing, while fdr2d takes both t and the composite reliability index, r, which will be further described in the next sub-section. Estimation of (2) is done via plug-in estimators of π_0, $f_0(z)$ and $f(z)$.

Let the observed statistics be denoted as $T = (t_1, \ldots, t_m)$ and $R = (r_1, \ldots, r_m)$, where m is the number of metabolic features.The null density $f_0(z)$ is estimated using the permutation method.We permute the risk factor (independent variable of the MWAS analysis) to obtain *K* sets of permuted variables. We run the MWAS procedure described earlier using each set of the permuted risk factor as the independent variable. The*K* sets of t-statistics$T^* = (T_1^*, \ldots, T_{10}^*)$ produced from the MWAS form the dataset for non-parametric estimation of f_0.In this study we used *K*=10.

Both f_0 and f are estimated using kernel density estimation methods, available in the R package KernSmooth [18, 19]. The bandwidth is selected using existing direct plug-in methodology[18, 20]. The observed density f is estimated using the observed data, T for fdr1d, and (T, R) for fdr2d. The null density f_0 is estimated using the permuted dataset, T^* for fdr1d and (T^*, R^*) for fdr2d, whereR^* is just *K* replicates of R.

We suggest that π_0 be estimated from (1), using the estimate obtained from Efron's 1D fdr procedure [5], which is more robust than basing the estimation on a 2-dimensional model fitting.In the presence of low-reliability features, the estimate of $\hat{\pi}_0$ using the 1D approach is an over-estimate of the truth, because low-reliability features only contribute to the null density. Using this over-estimate will result in slightly inflated lfdr estimates, which causes the overall lfdr procedure to be relatively conservative. However this inflation is minor. For example, an increase of $\hat{\pi}_0$ from 0.8 to 0.9 inflates the lfdr estimate by a factor of 1.1, which is well acceptable.

With the 3 estimates, $\hat{\pi}_0$, $\hat{f}_0^{2d}(t_i, r_i)$ and $\hat{f}^{2d}(t_i, r_i)$, we can plug in these estimators into (2) to get fdr2d. 2D density estimates of f_0 and f at the observed (t_i, r_i), denoted by $\hat{f}_0^{2d}(t_i, r_i)$ and $\hat{f}^{2d}(t_i, r_i)$, can be interpolated from their respective 2D kernel density estimates.

B. The composite reliability index for metabolomics data measured in triplicates.

Metabolic intensities are usually measured in triplicate, ie. for each feature, there are 3 readings per subject. The reliability index aims to account for within-subject variation and the intensity level. We define the reliability index for a single feature as

$$r = \sqrt{\frac{MSR}{MLC}}, \quad (3)$$

where *MSR* is the mean square residual obtained from the linear regression of log-intensity of a single feature against subject ID, and *MLC* is the mean log-intensity of the metabolic feature. This calculation is carried out for all of the *m* metabolic features in the dataset.

MSR is calculated from the analysis of variance of the linear model of log-intensity of a feature against the risk factor and adjusted for batch effect. That is,

$$MSR = \frac{\sum_{i=1}^{n}\sum_{j=1}^{3}\left[\widehat{log}(C_{ij}+1) - log(C_{ij}+1)\right]^2}{df}, \quad (4)$$

where $\widehat{log}(C_{ij}+1)$ and $log(C_{ij}+1)$ are the estimated and observed log-intensity of a single feature for the *j*-th replicate of the *i*-th subject respectively. The degrees of freedom (*df*) in the denominator of (4) is obtained from $df = 3n - n_m - p - 1$, where n is the total number of subjects, $n_m = \sum_{i=1}^{n}\sum_{j=1}^{3}(1-\omega_{ij})$ is the number of missing intensity values, with ω_{ij} as an indicator for a non-missing intensity, and p is the number of subjects with at least 1 observed replicate. In this way, *MSR* accounts for within-subject variation and the proportion of missing values.

MLC is the mean log-intensity of a single feature and is defined as

$$MLC = \frac{\sum_{i=1}^{n}\sum_{j=1}^{3}\omega_{ij} log(C_{ij}+1)}{\sum_{i=1}^{n}\sum_{j=1}^{3}\omega_{ij}}. \quad (5)$$

It only takes observed values into account.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

By this definition, we have a reliability index that takes smaller values when a feature is more reliably measured. The smallest possible value is zero.

III. RESULTS

A. The data used in this study

In this study, we used the metabolomics data generated from 494 subjects from the Emory Cardiovascular Biobank, which consists of patients who have undergone coronary angiography to document the presence/absence of coronary artery disease (CAD). Demographic characteristics, medical histories, behavioral factors and fasting blood samples have been documented and details about risk factor definitions and coronary angiographic phenotyping have been described previously[21, 22].

Each samplewas analyzed in triplicate with high-resolution liquid chromatography – mass spectrometry (LC-MS), using anion exchange and C18 chromatography combined with the Thermo Orbitrap-Velos (Thermo Fisher, San Diego, CA) mass spectrometer using an m/z range of 85 to 850. The data was pre-processed using xMSAnalyzer[23] in combination with apLCMS[16, 24]. For each feature, there were three readings per subject. An average metabolicintensity value was calculated from the non-zero readings for each subject. That is, an average reading of 0 was obtained only if all 3 readings for the subject were 0. This is the combined metabolic intensity data we used for subsequent analysis. This data was not corrected for batch effect. In our analysis, batch effect was accounted for linearly in the MWAS regression analysis as a confounder. There were 18,325 metabolic features detected.

B. Comparison between 2D lfdr and 1D lfdr

The reliability indices of each feature of this dataset of 18,325 features were calculated as described in the methods section. They range from 0.0042 to 0.26, with smaller values indicating more reliable measurements. The 10% most reliable features (n = 1841) had reliability indices of 0.0042 – 0.025.

In this proof-of-concept study, the risk factor of interest was High-Density Lipoprotein (HDL). The HDL levels ranged from 5 – 95 mg/dL, with mean 42.3 mg/dL and standard deviation 12.8 mg/dL. The MWAS was carried out as described in Section II.A.

Using our proposed fdr2d method, we found 301 significant features at the fdr cutoff of 0.2, while fdr1d found 210 significant features. By comparing the density plots, we clearly see the difference. As shown in Fig. 2(a), in fdr1d, Between the two lists, 174 of the significant features overlap, which means most of the features found by fdr1d were also found by fdr2d, and fdr2d found an extra ~100 features. Fig. 3 shows a comparison between the fdr levels as plotted against the test statistic. We see that for larger absolute values of the t-statistic, mostly the fdr2d procedure yielded lower fdr levels than fdr1d.

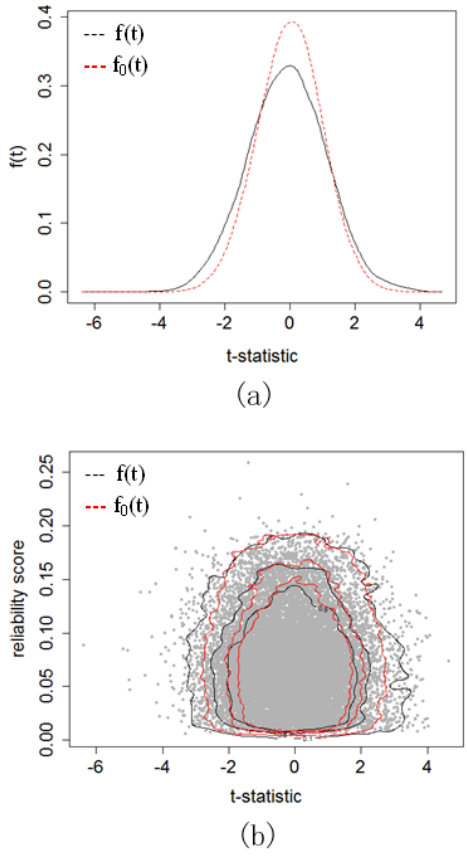

Fig. 2. Density plots of the test statistics. (a) test statistic density from fdr1d; (b) Joint density of test statistic and reliability score for fdr2d. For clarity, the t-statistics from permutations are not shown in the plot. Grey points: observed t-statistics and reliability scores. Three contours curves from inside out:density 0.9, 0.5 and 0.1. Grey: the density of f(t); red: the density of f0(t) based on permutation.

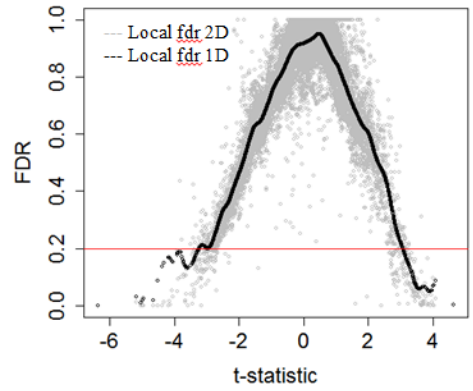

Fig. 3. Comparing the fdr levels generated from fdr1d and fdr2d. Black curve: local fdr versus t-statistic based on fdr1d; grey points: local fdr versus t-statistic based on fdr2d.

We next looked at pathway analysis for these significant features. We used R package xMSanalyzer [23]to annotate the features and MetaboAnalyst 2.0 [25]to see which of the pathways might be of interest. The annotated results match the feature m/z values to multiple known possible compounds in the Kyoto Encyclopedia of Genes and Genomes (KEGG), which can then be used by MetaboAnalyst 2.0 to determine possible significant pathways.

The 301 significant features under the fdr2d method had 451 annotated matches to known KEGG compounds with 109 unique m/z values. The pathways of possible interest include steroid hormone biosynthesis, primary bile acid biosynthesis, D-Glutamine and D-glutamate metabolism, and linoleic acid metabolism (Fig. 4a).

The 210 significant features under the fdr1d method had 310 annotated matches to known compounds with 78 unique m/z values. The pathways of potential interest includeD-glutamine and D-glutamate metabolism, limonene and pinene degradation, and alanine, aspartate and glutamate metabolism (Fig. 4b).

As the risk factor in this study is the level of High-Density Lipoprotein (HDL), clearly the fdr2d selected pathways, which are largely related to lipid metabolism, make better biological sense. For example, the class B type I scavenger receptor, SR-BI, is an HDL receptor that provides substrate cholesterol for steroid hormone synthesis[26]. Hepatocyte nuclear factor-1alpha (TCF1) regulates both bile acid and HDL metabolism [27]. In addition, the top pathway found by fdr1d, D-Glutamine and D-glutamate metabolism, is shared with the fdr2d result. More detailed biological interpretation will be conducted in a separate manuscript that focuses on the biomedical aspects of the Biobank metabolomics data.

Often jointly studied with HDL is the low-density lipoprotein (LDL). LDL itself is not measured by the LC/MS data, because LC/MS metabolomics measures small molecules. Nonetheless, because LDL was measured by a traditional method in this study, we tried to add it to the metabolite table and conduct the analysis. After obtaining its test statistics and adjusting it together with all other metabolites, we found that LDL is significant by the 2D lfdr method, with an lfdr value of 0.077. On the other hand, the 1D method assigned it an lfdr value of 0.475. Because HDL and LDL are known to associate with heart disease risk in a reverse manner in the population under study, this serves as a positive control and validates our new method.

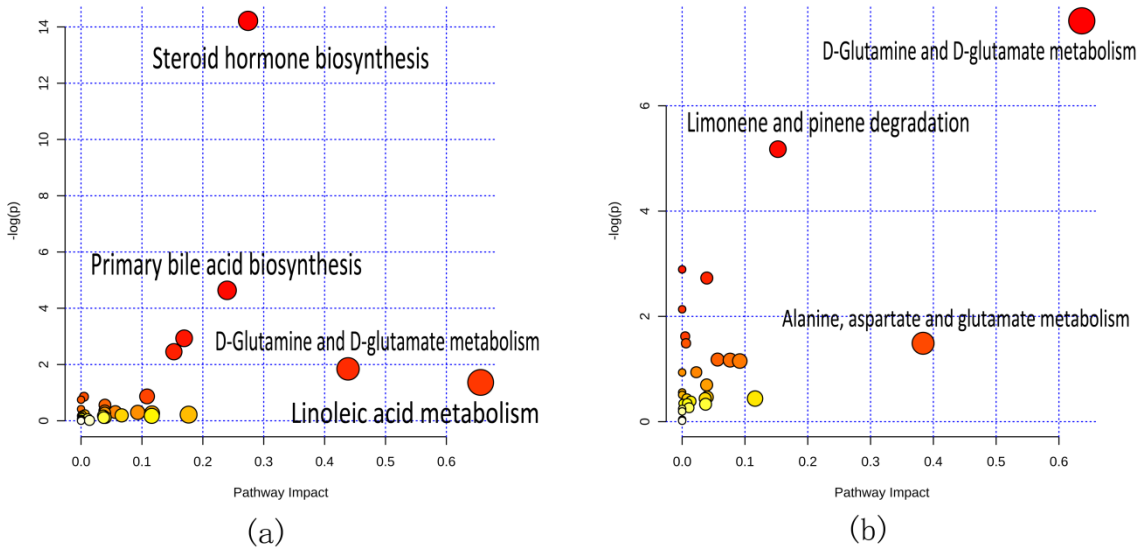

Fig. 4. Pathway analysis results using MetaboAnalyst 2.0 [25]. (a) Important pathways from the 301 significant features identified by the fdr2d method. (b) Important pathways from the 210 significant features identified by the fdr1d method. Pathways with large impact scores and/or small p-values are labeled.

IV. DISCUSSION

In the fdr2d approach, the permutation procedure ensures that the reliability index values of the metabolic features, $(r_1, \ldots, r_m)$ don't change. Thus the marginal density of the data points on the reliability index axis doesn't change in the permutations. Only the marginal density of the t-statistic, and conditional densities of t-statistic given reliability index changes. As a result, the t-statistic of each feature is effectively compared to the distribution of t-statistics of all features with similar reliability values. Thus in a sense, the fdr2d approach can be approximated by a 1D fdr estimation in which the features

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

are stratified based on their reliability indices. We can think of this as horizontal bands in Fig. 2b.

If we can assume that most differentially expressed features are concentrated in the most reliable range, another way the reliability indices can be used is to reduce the number of features that enter into the FDR/lfdr estimation. The fewer the number of features that are involved in the simultaneous hypothesis testing, the smaller adjustment to individual p-values. This approach amounts to taking only the stratum of the most reliable features. It is similar to the pre-filtering of genes based on variation criterion in gene expression data analysis.

For the Biobank dataset, when we consider only the 1841 features in the top 10% most reliable group, the Benjamini-Hochberg method [1]produced 95 significant features at the FDR level of 0.2. Applying the Benjamini-Hochberg method to all features, there were 269 features that were considered significant. Of these 269 features, 39 of them were in the top 10% most reliable group.

We then compared the FDR levels for the features in the most reliable 10% group. When only the 10% of features were involved in the FDR calculation, the FDR values were smaller than those obtained when involving all features (Fig 5). This result is expected as the Benjamini-Hochberg method is based on p-values and ranks[1].

We do not argue for substantially reducing the number of features under study, just to obtain better FDR or lfdr estimates and hence find more significant features. This simple demonstration mainly shows that a stratification approach may be used as a simpler version of the 2D fdr procedure when reliability measure is available, especially when the number of features under study is limited and cannot support robust 2D density estimation. When sufficient data points are available, still the fdr2d is preferred, as it avoids hard thresholding to stratify the data.

In broad terms, variation in the measurements of a metabolite can be dissected into biological variation (diet, diurnal variation, etc) and technical measurement noise. Biologically highly variant metabolites should not be confused with those with high measurement noise. In this manuscript, by reliability we mean technical reliability, i.e. a quantity that reflects technical measurement noise. This is possible to estimate in metabolomics data measured in replicates. Our reliability score is obtained by a regression model. It can be seen as a pooled coefficient of variation (CV) among the samples. It measures the technical measurement noise, but not the biological variation.

In non-targeted metabolomics, there is a risk of generating large number of fake metabolites if peak detection is carried out in an overly lenient manner. This is often done when the interest is to detect environmental impacts, because all environmental chemicals exist in the human blood in very low concentrations. When large numbers of fake metabolites are present, they certainly contain no biological variation at all. At the same time their technical variation is very big because they are just noise in LC/MS profiles, hence low reliability. While traditional methods allow them to contaminate the null distribution, our method suppresses their impact in the statistical inference to obtain more accurate local false discovery rate estimates.

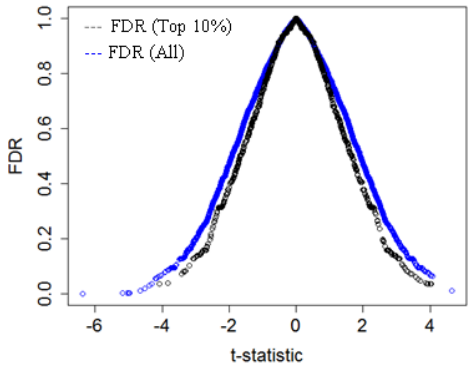

Fig. 5. Comparing the Benjamini-Hochberg FDR results between using all features v.s. using only the 10% most reliable features. The plot shows the FDR estimate versus the test statistic.

In conclusion, we presented a method for the computation of local false discovery rate that incorporates reliability index.In situations where substantial noise features are present, the method improves the statistical power of detecting differentially expressed features by minimizing the influence of noise features, because such features tend to have worse reliability values. One major aspect of this procedure is to quantify the reliability using a single variable. As we have shown, a composite reliability index for metabolic features worked well. Similar measures can be derived for other data types.

ACKNOWLEDGMENT

The authors thank three anonimous reviewers whose comment helped to improve the quality of this manuscript.

REFERENCES

[1] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," J. R. Statist. Soc. B, vol. 57, pp. 289-300, 1995.

[2] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," Proc Natl Acad Sci U S A, vol. 100, pp. 9440-5, Aug 5 2003.

[3] B. Efron and R. Tibshirani, "Empirical Bayes methods and false discovery rates for microarrays," Genetic Epidemiology, vol. 23, pp. 70-86, Jun 2002.

[4] A. Celisse and S. Robin, "A cross-validation based estimation of the proportion of true null hypotheses," J. Statist. Plan. Infer., vol. 140, pp. 3132-3147, 2010.

[5] B. Efron, "Large-scale simultaneous hypothesis testing: The choice of a null hypothesis," Journal of the American Statistical Association, vol. 99, pp. 96-104, Mar 2004.

[6] M. Guedj, S. Robin, A. Celisse, and G. Nuel, "Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation," BMC Bioinformatics, vol. 10, p. 84, 2009.

[7] J. G. Liao, Y. Lin, Z. E. Selvanayagam, and W. J. Shih, "A mixture model for estimating the local false discovery rate in DNA microarray analysis," Bioinformatics, vol. 20, pp. 2694-701, Nov 1 2004.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

[8] A. Ploner, S. Calza, A. Gusnanto, and Y. Pawitan, "Multidimensional local false discovery rate for microarray studies," Bioinformatics, vol. 22, pp. 556-65, Mar 1 2006.

[9] S. Scheid and R. Spang, "twilight; a Bioconductor package for estimating the local false discovery rate," Bioinformatics, vol. 21, pp. 2921-2, Jun 15 2005.

[10] K. Strimmer, "A unified approach to false discovery rate estimation," BMC Bioinformatics, vol. 9, p. 303, 2008.

[11] B. Wu, "On the geometric modeling approach to empirical null distribution estimation for empirical Bayes modeling of multiple hypothesis testing," Comput Biol Chem, vol. 43, pp. 17-22, Apr 2013.

[12] S. Anders and W. Huber, "Differential expression analysis for sequence count data," Genome Biol, vol. 11, p. R106, 2010.

[13] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," Bioinformatics, vol. 26, pp. 139-40, Jan 1 2010.

[14] H. Wu, C. Wang, and Z. Wu, "A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data," Biostatistics, vol. 14, pp. 232-43, Apr 2013.

[15] M. Katajamaa and M. Oresic, "Data processing for mass spectrometry-based metabolomics," J Chromatogr A, vol. 1158, pp. 318-28, Jul 27 2007.

[16] T. Yu, Y. Park, J. M. Johnson, and D. P. Jones, "apLCMS--adaptive processing of high-resolution LC/MS data," Bioinformatics, vol. 25, pp. 1930-6, Aug 1 2009.

[17] T. Yu, Y. Park, S. Li, and D. P. Jones, "Hybrid Feature Detection and Information Accumulation Using High-Resolution LC-MS Metabolomics Data," J Proteome Res, Feb 12 2013.

[18] M. P. Wand and M. C. Jones, Kernel smoothing. London: Chapman and Hall, 1995.

[19] M. P. Wand, "Fast Computation of Multivariate Kernel Estimators," Journal of Computational and Graphical Statistics, vol. 3, pp. 433-445, 1994.

[20] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," J. R. Statist. Soc. B, vol. 53, pp. 683-690, 1991.

[21] R. S. Patel, Y. V. Sun, J. Hartiala, E. Veledar, S. Su, S. Sher, et al., "Association of a genetic risk score with prevalent and incident myocardial infarction in subjects undergoing coronary angiography," Circ Cardiovasc Genet, vol. 5, pp. 441-9, Aug 1 2012.

[22] R. S. Patel, S. Su, I. J. Neeland, A. Ahuja, E. Veledar, J. Zhao, et al., "The chromosome 9p21 risk locus is associated with angiographic severity and progression of coronary artery disease," Eur Heart J, vol. 31, pp. 3017-23, Dec 2010.

[23] K. Uppal, Q. A. Soltow, F. H. Strobel, W. S. Pittard, K. M. Gernert, T. Yu, et al., "xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data," BMC Bioinformatics, vol. 14, p. 15, 2013.

[24] T. Yu and D. P. Jones, "Improving peak detection in high-resolution LC/MS metabolomics data using preexisting knowledge and machine learning approach," Bioinformatics, Jul 7 2014.

[25] J. Xia, R. Mandal, I. V. Sinelnikov, D. Broadhurst, and D. S. Wishart, "MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis," Nucleic Acids Res, vol. 40, pp. W127-33, Jul 2012.

[26] A. Rigotti, E. R. Edelman, P. Seifert, S. N. Iqbal, R. B. DeMattos, R. E. Temel, et al., "Regulation by adrenocorticotropic hormone of the in vivo expression of scavenger receptor class B type I (SR-BI), a high density lipoprotein receptor, in steroidogenic cells of the murine adrenal gland," J Biol Chem, vol. 271, pp. 33545-9, Dec 27 1996.

[27] D. Q. Shih, M. Bussen, E. Sehayek, M. Ananthanarayanan, B. L. Shneider, F. J. Suchy, et al., "Hepatocyte nuclear factor-1alpha is an essential regulator of bile acid and plasma cholesterol metabolism," Nat Genet, vol. 27, pp. 375-82, Apr 2001.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE