Significance Analysis by Minimizing False Discovery Rate

Yuanzhe Bei and Pengyu Hong
Computer Science Department
Brandeis University, Waltham, MA 02453
E-mail: {beiyz, hongpeng}@brandeis.edu

***Abstract*—False discovery rate (FDR) control is widely practiced to correct for multiple comparisons in selecting statistically significant features from genome-wide datasets. In this paper, we present an advanced significance analysis method called miFDR that minimizes FDR when the number of the required significant features is fixed. We compared our approach with other well-known significance analysis approaches such as Significance Analysis of Microarrays [1-3], the Benjamini-Hochberg approach [4] and the Storey approach [5]. The results of using both simulated data sets and public microarray data sets demonstrated that miFDR is more powerful.**

Keywords: significant analysis, false discovery rate

I. Introduction

High-throughput technologies such as DNA microarray have made it possible to screen thousands of genomic features simultaneously. To identify a subset of interesting features for follow-up investigation, a large number of hypotheses are tested simultaneously. Hence, it is important to control false positives among the tests called "significant". False discovery rate (FDR) was first introduced by Benjamini-Hochberg [4] in 1995 to measure multiple-hypothesis testing errors, which was later improved by the Storey approach [5] in 2002. Both the Benjamini-Hochberg (BH) approach and the Storey approach estimate FDR values by taking the *p*-values of features calculated using some sorts of hypothesis tests (such as, Student's *t*-test [6] and the Wilcoxon ranksum test [7]). However, the calculation of *p*-values can be hampered by the number of samples and the deviation of true distributions from the one assumed by the *p*-value calculation methods. Significance Analysis of Microarrays (SAM) [1-3] offers a powerful alternative to the *p*-value based approaches (e.g., the BH and Storey approaches) for controlling FDR. Instead of calculating *p*-values first and then correcting them, SAM permutes the values of each feature under two different conditions and calculates a corrected *t*-statistics to measure the relative difference of the feature between two conditions. The statistics of a feature from all permutations are then used to calculate the expected statistics of the feature. SAM uses the difference between the observed statistics and the expected statistics to decide the cutoffs for calling significant features. Although SAM has been found to outperform the BH approach and the Storey approach in analyzing many datasets, our analysis and experiments suggest that its results are often not optimal due to the way it chooses cutoffs.

We therefore developed a more advanced significance analysis method to minimize FDR when the number of required significant features was fixed, and named this method miFDR. We compared miFDR, SAM, the BH approach, and the Storey approach using both simulated data sets and microarray gene expression data sets. The results showed that miFDR significantly outperformed the other three approaches. In particular, the results of simulation tests showed that miFDR did not under-estimate FDRs because its true FDRs were consistently bounded by its estimated FDRs. In addition, both the true FDRs and the estimated FDRs of miFDR were bounded by their counterparts of the other three approaches. When tested on the microarray data sets, miFDR was capable of identifying more biologically relevant genes than the other approaches as supported by literature evidence.

The rest of the paper is organized as follows: Our miFDR adopts the permutation method used in SAM. To better explain miFDR, we first describe SAM and point out an important problem in Section II. We then propose miFDR to address this problem in Sections III. The comparisons of miFDR, SAM, the BH approach, and the Storey approach are presented and discussed in Section IV. Finally, Section V concludes this paper.

II. SAM Approach to Control FDR

Without losing generality, we assume that the samples are divided into two groups: group C_1 with n_1 samples and group C_2 with n_2 samples. SAM [3] computes a corrected *t*-statistic (or *d*-value) $d_i = (\bar{x}_i^1 - \bar{x}_i^2)/(s_i + s_0)$ to measure the relative difference of the *i*-th feature between two groups, where:

- $\bar{x}_i^1$ and $\bar{x}_i^2$ are the average values of the *i*-th feature in C_1 and C_2, respectively.
- s_i is feature specific scatter $\sqrt{a[\sum_{x_{im} \in C_1}(x_{im} - \bar{x}_i^1)^2 + \sum_{x_{in} \in C_2}(x_{in} - \bar{x}_i^2)^2]}$, and

$a = (1/n_1 + 1/n_2)/(n_1 + n_2 - 2)$.

- s_0 is a regularization factor chosen to minimize the dependency of *d*-values on the feature value. Details for computing s_0 are described in [3].

SAM can also use ranksum statistics. However, we found that SAM performed worse in the experiments carried out in this study when the ranksum statistics was used. Hence, we did not consider using the ranksum statistics with SAM in this paper.

SAM ranks features by their *d*-values so that the *d*-value of the *i*-th feature is the *i*-th largest. Then, SAM uses random permutation to generate a large number of controls. The relative differences (*i.e.*, *d*-values) from each permutation are ordered so that the *i*-th largest *d*-value in the permutation is assigned to the *i*-th feature. An expected *d*-value $E[d_i]$ is calculated for the *i*-th feature as the mean of the relative differences from all permutations. Note that $E[d_i]$ may be calculated using the permutations from other features. SAM then calculates $\Delta = d_i - E[d_i]$ for the *i*-th feature. Given a threshold $\tilde{\Delta}$, SAM finds the first feature indexed by *k* that satisfies $\Delta_k > \tilde{\Delta}$, and defines $cut_{up} = d_k$. The features with *d*-values > cut_{up} are called "significant positive" features. Similarly, $-\Delta_k > \tilde{\Delta}$ is used for deciding cut_{down} and calling "significant negative" features. The number of falsely called features in each permutation is computed by counting the number of features with their permutated *d*-values exceed the cutoffs (cut_{up} and cut_{down}). Finally, FDR is computed as the median of the number of falsely called features across all permutations multiplied by a factor π_0 and then divided by the total number of features called significant. The factor π_0 indicates the proportion of the true null features in the data set (details about how to calculate π_0 are described in [3]).

SAM decides both cut_{up} and cut_{down} based on Δ-values. We found that Δ-values did not always increase/decrease with the corresponding *d*-values. In some cases, the Δ-values of some features were positive even though their *d*-values were negative. Figure 1 shows an example using a Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) dataset GDS3661 [8]. Although the Δ-values increase monotonically with the *d*-values in most part of the curve, this trend disappears at both ends of the curve, where usually the most significant features are located. If we choose the absolute of the Δ-value indicated by the black horizontal lines in Figures 1a & 1b as $\tilde{\Delta}$ to decide cut_{down}, SAM will use the *d*-value indicated by the solid arrow in Figure 1b as cut_{down} to select significant negative features. This cut_{down} will exclude a feature set framed by the ellipse in Figure 1b. The *d*-values of these features do not concentrate within a narrow band, but spread out in a large region ranging from -6.3 to -7. Decreasing the chosen $\tilde{\Delta}$ by a very small amount to the next cutoff notch will change cut_{down} to the value indicated by the open arrow in Figure 1b. This new cut_{down} will call all features in the ellipse. There is no other intermediate $\tilde{\Delta}$ that allows us to call a subset of those features even though doing so may maintain or improve FDR (calling more features can sometimes result in a smaller FDR, see our simulation results in the Result section). This means, SAM sometimes fails to explore many options. Hence, we concluded that Δ-value is not always a reliable index for determining *d*-value cutoffs.

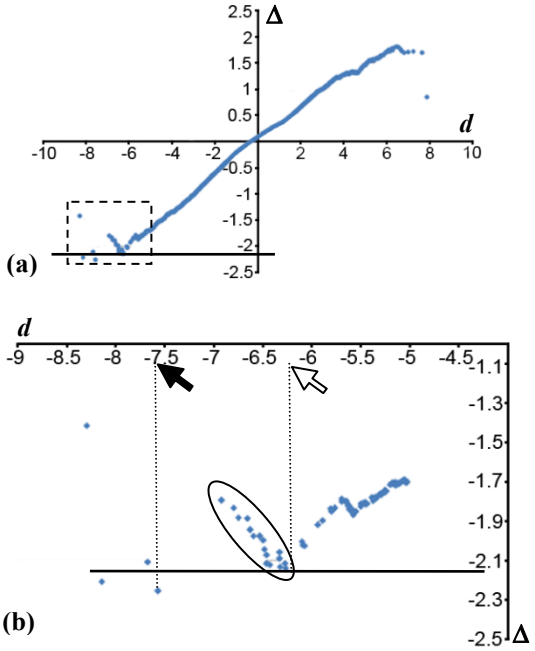

Figure 1. An example showing the problem of using Δ-values to decide the *d*-value cutoffs. The *x*-axis indicates *d*-values and the *y*-axis inciates Δ-values. (b) is the blow-out of the dashed rectangle region in (a). Two *d*-value cutoffs are indicated by the solid arrow and the open arrow respectively in (b) (see main text for detailed explanations). The solid arrow marks the *d*-value cut_{down} corresponding to the Δ-value indciated by the black line.

III. MINIMIZE FDR – MIFDR

We therefore propose to abandon Δ-values and rely on *d*-values in detecting significant features. Our idea is simple and goes as follows. We first sort all features by their *d*-values in descending order. If the *d*-values of two features have the same sign, the one with the larger absolute *d*-value is more significant. To call a fixed number of significant features, there can be many possible combinations of cut_{up} and cut_{down}. We would like to select one combination that minimizes the estimated FDR. More specifically, if we want to call *N* significant features = *r* positive significant features + *q* negative significant features, there can be *N*+1 possible choices for (*r*,*q*): (*0*,*N*), (1,*N*-1), (2,*N*-2), …, (*N*,0). Each choice of (*r*,*q*) has its own FDR. We designed the following straight-forward algorithm to find the combination with the minimal FDR:

Algorithm: [*fdr*, *pSig*, *nSig*] = ***miFDR***(*d*-values, *N*)
Input: *d*-values are sorted from the largest to the smallest; *N* is the number of significant features to be called.
1) Initialize the estimated false discovery rate *fdr* = inf, positive significant features *pSig* = [ ], and negative significant features *nSig* = [ ].
2) For $n \leftarrow 0$ to *N*
   2.1) Select the positive significant features as those with the *n* largest *d*-values.
   2.2) Select the negative significant features as those with the $N - n$ smallest *d*-values.
   2.3) Compute current false discovery rate cFDR using the permutated data in the same way SAM does (see details in Section II).
   2.4) If cFDR< *fdr*, then *fdr* = cFDR, update *pSig* and *nSig*.
Output: *fdr*, *pSig* and *pNeg*.

To call *N* significant features, the miFDR algorithm explores *N*+1 possible options, and hence its computational complexity is O(*N*). The options explored by miFDR include the one considered by SAM. Hence, the results of miFDR should have a lower estimated FDR than that of SAM. In addition, at the same FDR level, miFDR will identify at less the same number of genes as SAM does. It will takes a complexity of $O(N^2)$ to obtain a complete FDR *vs* (feature number) curve up to *N* significant features (such as those shown in Figure 2). The complexity of finding the optimal *N* under a particular FDR cut-off is O(*N*log*N*) when binary search is used. In practice, it might not be very useful to call a very large number of features because of the following two reasons. First, the corresponding FDR will very likely be unacceptably high when *N* is very large. Second, it will be too expensive to experimentally follow up a large number of significant features.

IV. Results

We compared the BH approach and the Storey approach, SAM (version 4.0), and miFDR on simulated datasets and two real microarray datasets. Two-side *t*-test was applied to calculate the *p*-values of features used by the BH approach and the Storey approach. We also tried the one-side *t*-test, one-side and two-side Wilcoxon ranksum test. However, the results were generally worse than those of the two-side *t*-test. The implementations of the BH and Storey approaches in the MATLAB Bioinformatics Toolbox were used. The results showed that miFDR was more powerful than the other three approaches at a wide spectrum of FDR cutoffs.

A. Simulation Tests

We conducted a simulation study, in which the ground truth was known. In each simulation, the simulated dataset contained 12 samples (6 in the 1st group, and the other 6 in the 2nd group). The number of features was 10400, which were simulated in the way shown in Table 1. All features in the 1st group are controls. The 2nd group contains 10000 features belonging to the null (categories 1 & 2 in Table 1) and 400 features (categories 3, 4, 5 & 6 in Table 1) belonging to the alternative. The datasets were simulated using a mixture of distributions so that they had the complexity typically seen in real data sets to fully test different approaches. Note that all uniform distributions used in the simulation test are $2\sqrt{3}$ in range, which makes the standard deviation of every feature equal to 1.

TABLE 1. SIMULATED ALTERNATIVE HYPOTHESIS

Category	# of features	1st Group	2nd Group
1	5000	Normal distribution (mean = 0, var = 1)	Normal distribution (mean = 0, var = 1)
2	5000	Uniform distribution in $[-\sqrt{3}, \sqrt{3}]$	Uniform distribution in $[-\sqrt{3}, \sqrt{3}]$
3	50	Normal distribution (mean = 0, var = 1)	Normal distribution (mean = –2, var = 1)
4	150	Normal distribution (mean = 0, var = 1)	Normal distribution (mean = 1, var = 1)
5	150	Uniform distribution in $[-\sqrt{3}, \sqrt{3}]$	Uniform distribution in $[-\sqrt{3}+1, \sqrt{3}+1]$
6	50	Uniform distribution in $[-\sqrt{3}, \sqrt{3}]$	Uniform distribution in $[-\sqrt{3}+1.5, \sqrt{3}+1.5]$

We ran the simulation 100 times. Each time we obtained a curve for each of the four approaches (miFDR, SAM, the BH approach, and the Storey approach) showing the estimated FDRs *vs.* the numbers of features called significant. We then calculated the mean curve of each approach with respect to the number of features called significant. The results are plot in Figure 2. Since we have the ground-truth, we can also compare the estimated FDRs and the true FDRs of those approaches. True FDR is the ratio between the number of null features falsely called significant and the total number of features called significant. The calculations of the estimated FDRs of SAM and miFDR are explained in Sections II and III, respectively.

We observed that miFDR consistently called more significant features than SAM did at the same estimated FDR levels (Figure 2a). We also compared the true FDRs of the features called by miFDR and SAM, respectively. The true FDR curve of miFDR was also consistently bounded by that of SAM (Figure 2b). That is, to identify any fixed number of significant features, the number of features falsely called significant by miFDR was smaller than that of SAM. Furthermore, the true FDR curve of miFDR was also bounded by the estimated FDR curve of miFDR (Figure 2c), indicating that miFDR did not under-estimate FDRs.

The BH and Storey approaches performed much worse than miFDR and SAM (Figure 2a & 2b). They consistently called fewer numbers of features than miFDR and SAM did at the same estimated FDR levels. In addition, their true FDRs are significantly higher than their estimated FDRs mainly because of the following reason. Fifty percent of features did not follow Gaussian distributions. However,

the BH and Storey approach using *t*-test to calculate *p*-values, which assumes Gaussian distributions. We also tried using the ranksum *p*-values in the BH and Storey approaches. Nevertheless, the performances were even worse (see Figure 3).

B. Results of Analyzing Real Microarray Data

We applied miFDR, SAM, the BH approach, and the Storey approach to two real microarray datasets GDS3661 and GDS3689. Both of them were publically available at Gene Expression Omnibus and are related to hypertension (or high blood pressure). Hypertension accounts for about one quarter of heart failure cases [9]. Uncontrolled hypertension can lead to various changes in the myocardial structure, coronary vasculature, and conduction system of the heart, which in turn can lead to the development of left ventricular hypertrophy, atherosclerosis, and several other complications. Both experimental animal studies and clinical studies have shown that left ventricular hypertrophy could lead to myocardial ischemia [10], which can result in large-scale programmed cell death and eventually heart failure. To investigate the molecular events underlying the onset of hypertensive heart failure, the gene expression data set GDS3661 is generated to profile the left ventricular samples from spontaneously hypertensive rats using Affymetrix Rat Genome 230 2.0 Array [8]. The dataset contains twelve samples in two groups: six without compensated hypertrophy *versus* six with compensated hypertrophy. In addition, extensive epidemiological evidence supports that links exist between diesel exhaust exposure and hypertension [11-15]. To investigate the underlying molecular mechanisms, Gottipolu et al. [16] generated the GEO dataset GDS3689 by exposing both healthy and hypertensive rats to diesel exhaust particles and then profiled gene expression levels by using Affymetrix Gene Chip Rat 230A microarray. This dataset has 16 samples, 8 hypertensive rats and 8 healthy rats. In each group, 4 out of 8 rats are exposed to diesel exhaust particles. In our analysis, we compared healthy rats without exposure (Control-rats) with healthy rats exposed to diesel exhaust particles (DE-rats).

For the first dataset GDS3661, at FDR < 0.05, neither the BH approach nor the Storey approach was able to identify any significant probe set no matter whether the *t*-test or the ranksum test was used to calculate the *p*-values of probe sets. Under the same FDR level, miFDR identified 210 probe sets while SAM identified 129 probe sets. We submitted the probe set lists called by SAM and miFDR respectively to DAVID [17, 18] for functional analysis using Gene Ontology terms [19]. The results showed that miFDR called more genes than SAM did in the functional categories closely associated with phenotypic changes from compensated hypertrophy to systolic heart failure, such as stress response (miFDR called 10 probe sets *vs.* SAM called 5 probe sets), positive regulation of cell communication (miFDR 10 *vs.* SAM 5), programmed cell

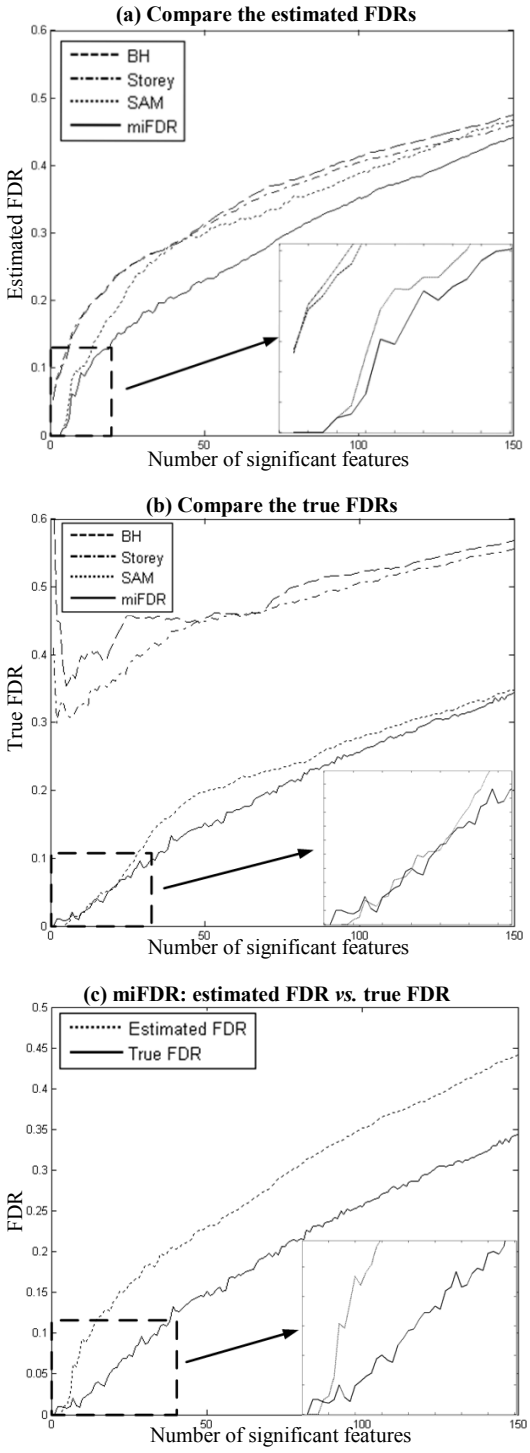

Figure 2. Compare the the average performances of BH, Storey, SAM and miFDR on 100 simulated datasets. In each plot, a blow-out of the curve segment in the dash rectangle is shown at the bottom-right corner for clearer illustration. **(a)** Compare the estimated FDRs. The performance of miFDR is the best. **(b)** Compare the true FDRs. Again, the performance of miFDR is the best. **(c)** Compare the estiamted FDRs and the true FDRs of miFDR. The true FDR of miFDR is bounded by its estimated FDR, indicating that miFDR does not under-estimate the number of features falsely called significant.

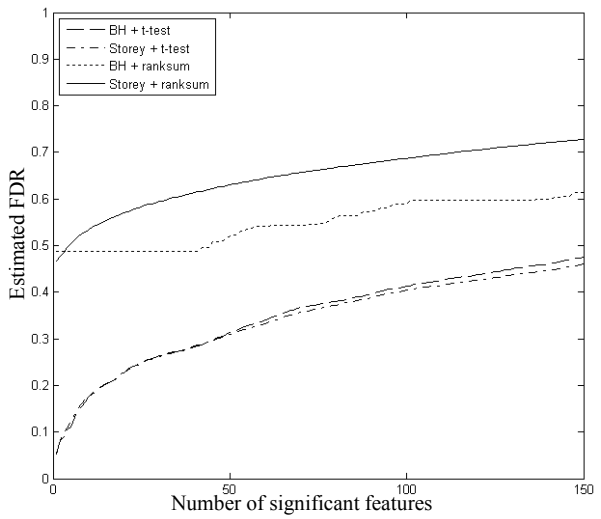

Figure 3. Compare the avearage performances of BH + *t*-test, Storey + *t*-test, BH + ranksum, Storey + ranksum on the simulated datasets. The performances of BH + *t*-test and Storey + *t*-test are much better than those of BH + ranksum and Storey + ranksum.

death (miFDR 9 *vs.* SAM 7), regulation of growth (miFDR 6 *vs.* SAM 0), TGF beta signaling pathway (miFDR 4 *vs.* SAM 3), and so on.

PubMed search revealed literature evidence suggesting that several genes called only by miFDR can indeed deepen our understanding about the molecular mechanisms underlying the deterioration of cardiac function and remodeling associated with hypertensive heart failure. For example, *Mmp*2 was up-regulated (1.88 folds) in heart failure rats. It was shown that angiotensin-converting enzyme inhibitors can suppress *Mmp*2 activity to prevent left ventricular remodeling in a rat model of heart failure [20]. Classic preconditioning can offer cardio-protection by inhibiting ischemia/reperfusion induced release and activation of *MMP*2 protein [21]. We hypothesize that *Mmp*2 plays an important role in the transition from hypertension to heart failure and inhibiting *Mmp*2 can help prevent such a progression from happening. *Pdlim*5 was up-regulated (2.96 folds) in heart failure rats, and it is a heart and skeletal muscle-specific protein that may play an important role in heart development [22]. *PDLIM*5 protein preferentially interacts with protein kinase *C beta* that is markedly activated in the cardiac hypertrophic signaling [23]. It was suggested that *PDLIM*5 protein scaffolded protein kinase *D*1, which played a central role in the response to stress signals in cardiomyocytes, to regulate the activity of the cardiac L-type voltage-gated calcium channel [24]. When over expressed in rat neonatal cardiomyocytes, *Pdlim*5 promoted the expression of hypertrophy markers and increased cell volume [25].

For the second dataset GDS3689, at FDR < 0.05, both the BH approach and the Storey approach identified 0 significant probe set if the ranksum *p*-values were used. Switching to the *t*-test *p*-values, the BH and Storey approaches identified 18 and 249 significant probe sets, respectively. Under the same FDR level, miFDR identified 640 significant probe sets while SAM identified 388 probe sets, which were significantly more than the results of the BH and Storey approaches. We submitted the probe set lists identified by SAM and miFDR for functional analysis at DAVID [17, 18] using Gene Ontology terms [19]. The results showed that miFDR identified more genes than SAM did in many functional categories closely associated with the response to diesel exhaust exposure and hypertension, such as, stress response (miFDR called 21 probe sets *vs.* SAM called 10 probe sets), inflammatory response (miFDR 12 *vs.* SAM 6), defense response (miFDR 17 *vs.* SAM 8), extracellular matrix organization (miFDR 3 *vs.* SAM 0), positive and negative signal transduction (miFDR 19 *vs.* SAM 11), and blood vessel morphogenesis (miFDR 8 *vs.* SAM 0).

We were also able to retrieve literature evidence suggesting that the genes called only by miFDR can shed new lights on the cellular and molecular links between hypertension and diesel exhaust exposure. One of the most serious health problems related to hypertension is atherosclerosis. Arteries of hypertensive animals have a greater mass of vascular smooth muscle than those of normotensive controls. Alteration in the differentiated state of vascular smooth muscle cells, such as increased proliferation, enhanced migration, and down-regulation of vascular smooth muscle cell differentiation marker genes, is known to play a key role in the development of atherosclerosis. Recent data have implicated air pollution (such as diesel exhaust particles) as one of the important risk factors for atherosclerosis. It has been the subject of extensive reviews [26, 27] and a consensus statement from the American Heart Association [28]. Diesel exhaust particles and oxidized phospholipids synergistically affect the expression profile of several gene modules that correspond to pathways relevant to vascular inflammatory processes such as atherosclerosis [29]. Several genes (*e.g.*, *Rab5a*, *Zeb*1, and *Hdac*2) called by miFDR alone have been reported to regulate vascular smooth muscle cell differentiation marker genes.

In addition, miFDR identified two other interesting genes: *Tgfbr*1 and *Plau*. *Tgfbr*1, which forms a heteromeric receptor complex with *TGF-beta* type II receptor that mediates *TGF-beta* signaling, was up-regulated (2.05 folds) in DE-rats. It is shown that diesel exhaust particles activate *p*38 *MAP* kinase to produce interleukin 8 and RANTES by human bronchial epithelial cells [30]. Since TGFBR1 is the upstream of p38 in MAPK signaling pathway (http://www.genome.jp/kegg-bin/show_pathway?hsa04010), our discovery suggests that diesel exhaust particles trigger *p*38 by activating *TGFBR*1. *Plau* was up-regulated (3.39 folds) in DE-rats. This gene encodes a serine protease that displays a post-transcriptional increase in enzyme levels in chemically induced mammary carcinoma and may play a

role in tumor invasion and metastasis (Entrez Gene summary, http://www.ncbi.nlm.nih.gov/gene/25619), which make it a potential molecular link between diesel exhaust exposure and cancers.

V. Conclusion

This paper presents miFDR, a new powerful method for controlling FDR. Our miFDR algorithm minimizes the estimated FDR when calling a fixed number of significant features. The results of analyzing two genome-wide microarray datasets and the simulated datasets demonstrate that miFDR is much more powerful than three widely used methods for controlling FDR (i.e., SAM, the BH approach, and the Storey approach). We also found literature evidence to support that some genes identified only by miFDR are indeed relevant to the underlying biology of interest. Since FDR control has been widely applied in genome-wide studies, we expect miFDR to benefit many such projects and generate broad impact in the future.

References

1. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p.5116-21.
2. Chu, G., et al., *Significance Analysis of Microarrays -- Users guide and technical document ver 3.0*, 2007.
3. Chu, G., et al., *Significance Analysis of Microarrays – Users guide and technical document ver 4.0* 2011.
4. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: A practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society, 1995. **57**: p. 289-300.
5. Storey, J.D., *A direct approach to false discovery rates.* Journal of the Royal Statistical Society, 2002. **64**(3): p. 479-498.
6. Student, *The Probable Error of a Mean.* Biometrika, 1908. **6**(1): p. 1-25.
7. Wilcoxon, F., *Individual Comparisons by Ranking Methods.* Biometrics Bulletin, 1945. **1**(6): p. 80-83.
8. Brooks, W.W., et al., *Transition from compensated hypertrophy to systolic heart failure in the spontaneously hypertensive rat: Structure, function, and transcript analysis.* Genomics, 2010. **95**(2): p. 84-92.
9. Kannel, W.B. and J. Cobb, *Left ventricular hypertrophy and mortality--results from the Framingham Study.* Cardiology, 1992. **81**(4-5): p. 291-8.
10. Dunn, F.G. and S.D. Pringle, *Left ventricular hypertrophy and myocardial ischemia in systemic hypertension.* Am J Cardiol, 1987. **60**(17): p. 19I-22I.
11. Pope, C.A., 3rd, et al., *Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease.* Circulation, 2004. **109**(1): p. 71-7.
12. Brook, R.D., et al., *Inhalation of fine particulate air pollution and ozone causes acute arterial vasoconstriction in healthy adults.* Circulation, 2002. **105**(13): p. 1534-6.
13. Ibald-Mulli, A., et al., *Effects of air pollution on blood pressure: a population-based approach.* Am J Public Health, 2001. **91**(4): p. 571-7.
14. Ibald-Mulli, A., et al., *Effects of particulate air pollution on blood pressure and heart rate in subjects with cardiovascular disease: a multicenter approach.* Environ Health Perspect, 2004. **112**(3): p. 369-77.
15. Zanobetti, A., et al., *Ambient pollution and blood pressure in cardiac rehabilitation patients.* Circulation, 2004. **110**(15): p. 2184-9.
16. Gottipolu, R.R., et al., *One-month diesel exhaust inhalation produces hypertensive gene expression pattern in healthy rats.* Environ Health Perspect, 2009. **117**(1): p. 38-46.
17. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
18. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.
19. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
20. Brower, G.L., S.P. Levick, and J.S. Janicki, *Inhibition of matrix metalloproteinase activity by ACE inhibitors prevents left ventricular remodeling in a rat model of heart failure.* Am J Physiol Heart Circ Physiol, 2007. **292**(6): p. H3057-64.
21. Lalu, M.M., et al., *Preconditioning decreases ischemia/reperfusion-induced release and activation of matrix metalloproteinase-2.* Biochem Biophys Res Commun, 2002. **296**(4): p. 937-41.
22. Nakagawa, N., et al., *ENH, containing PDZ and LIM domains, heart/skeletal muscle-specific protein, associates with cytoskeletal proteins through the PDZ domain.* Biochem Biophys Res Commun, 2000. **272**(2): p. 505-12.
23. Kuroda, S., et al., *Protein-protein interaction of zinc finger LIM domains with protein kinase C.* J Biol Chem, 1996. **271**(49): p. 31029-32.
24. Maturana, A.D., et al., *Enigma homolog 1 scaffolds protein kinase D1 to regulate the activity of the cardiac L-type voltage-gated calcium channel.* Cardiovasc Res, 2008. **78**(3): p. 458-65.
25. Yamazaki, T., et al., *Splice variants of enigma homolog, differentially expressed during heart development, promote or prevent hypertrophy.* Cardiovasc Res, 2010. **86**(3): p. 374-82.
26. Bhatnagar, A., *Environmental cardiology: studying mechanistic links between pollution and heart disease.* Circ Res, 2006. **99**(7): p. 692-705.
27. Brook, R.D., *Cardiovascular effects of air pollution.* Clin Sci (Lond), 2008. **115**(6): p. 175-87.
28. Brook, R.D., et al., *Air pollution and cardiovascular disease: a statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association.* Circulation, 2004. **109**(21): p. 2655-71.
29. Gong, K.W., et al., *Air-pollutant chemicals and oxidized lipids exhibit genome-wide synergistic effects on endothelial cells.* Genome Biol, 2007. **8**(7): p. R149.
30. Hashimoto, S., et al., *Diesel exhaust particles activate p38 MAP kinase to produce interleukin 8 and RANTES by human bronchial epithelial cells and N-acetylcysteine attenuates p38 MAP kinase activation.* Am J Respir Crit Care Med, 2000. **161**(1): p. 280-5.