Incorporating Genome-Wide Co-expression Information to Improve the Analysis of Gene Expression Data

Yinglei Lai
Department of Statistics, The George Washignton University, Washington, D.C., 20052, U.S.A.

* Corresponding author. Email: ylai@gwu.edu
Manuscript submitted January 10, 2015; accepted March 8, 2015.
doi: 10.17706/ijbbb.2015.5.3.149-156

Abstract: Although microarray and deep sequencing technologies allow us to monitor gene expression at a genomic level, they also generate a relatively high false discovery rate, which limits our further exploration of disease mechanism. Furthermore, it is usually difficult to detect disease related genes with weak differential expression. It is well known that genes interact with each other during cellular and molecular processes. Therefore, disease related genes are not isolated. With large scale gene expression data, it is possible to evaluate the genome-wide co-expression information. This information is valuable for understanding gene interactions. We have shown that differential expression analysis can be improved when this information can be incorporated. In this study, we demonstrate that genome-wide co-expression information can also clearly improve gene set enrichment analysis and classification analysis.

Key words: Expression, co-expression, differential expression, classification.

1. Introduction

Microarrays can be used to measure expression for tens of thousands of genes at the mRNA level for samples in normal and disease groups, and then statistical methods for two-sample comparison can be used to identify differentially expressed genes. Differentially expressed genes are potential disease related genes for clinical diagnoses and medical treatments. This approach has been successfully used in cancer studies [1] as well as diabetes studies [2]. Furthermore, disease related pathways or gene sets can be identified through microarray analysis [3]. With microarray data, we can understand unknown gene functions through the clustering analysis. Microarray data also allow us to perform disease classifications at a molecular level.

The recent deep sequencing technology (RNA-seq or digital gene expression) has shown a promising impact on biomedical studies. It can directly measure the amount of molecules at a genomic level. Compared to the microarray technology, this new technology can significantly reduce the noise in expression measurements and improve the detection range and accuracy. Nevertheless, as a well-developed technology, micorarrays have been continuously used for broad biomedical studies [4]. There is still a need for more efficient statistical and computational methods for analyzing these types of gene expression data. Furthermore, since the structures of data from different genomic technologies are basically similar, methods for analyzing genome-wide expression data can also be useful for analyzing other similar genomics data.

The traditional t/F-statistics as well as the nonparametric Wilcoxon/Kruskal-Wallis rank sum tests have

been widely used for two-sample or multi-sample comparisons. Due to the relatively small sample size of microarray data, it is difficult to achieve sufficient power from the nonparametric methods. Based on the exploration of microarray data, many statistical methods have been proposed to improve the detection of differential expression. It is well known that genes interact with each other during cellular and molecular processes. Disease related genes are not isolated. Furthermore, to understand gene interactions, we can use genome-wide expression data to measure the co-expression among genes at the mRNA level. Genome-wide co-expression information can be useful in the detection of disease related genes with relatively weak differential expression, since these genes are expected to co-express with many other differentially expressed genes. Therefore, we expect to further improve the detection of disease related genes if an efficient statistical method can be developed to incorporate the genome-wide co-expression information into the differential expression analysis. Storey *et al*. [5] have proposed an optimal discovery procedure (ODP) for large-scale significance testing. Their method evaluates the differential expression of a gene with the consideration of information from the other genes. However, the co-expression information is not explicitly considered in the procedure. Tibshirani and Wasserman [6] have proposed a correlation-sharing method for detecting differential expression. Their method evaluates differential expression through the correlation-sharing based maximization procedure: for a fixed gene X, its neighbors can be first defined as these genes (including X itself) with absolute correlations (with X) greater than a given threshold value; then, the average of differential expression measures of these neighbours can be obtained; the maximal average is reported after the threshold value is screened from 0 to 1. However, since the correlation is only used to rank genes, the magnitude of co-expression is not explicitly considered in the procedure. We have recently proposed to use a local regression technique for modelling the relationship between the differential expression and co-expression [7]. This method can be extended to other types of expression data analysis.

2. A Gene Set Enrichment Analysis (GSEA)

We first briefly describe our method [7] as follows. Now consider a large number of genes measured by microarrays or RNA-seq: $\{X_1, X_2,..., X_m\}$. Given a gene X_j, we use the Pearson's correlation coefficient r_{jk} to measure the co-expression between X_j and another gene X_k, whose differential expression is calculated by the traditional two-sample *t*-test t_k (observed differential expression). Therefore, each gene X_k has a pair of measurements (r_{jk}, t_k). To gather more observations for the local regression, we consider a modification: $(r_{jk}', t_k') = (s_{jk}r_{jk}, s_{jk}t_k)$, where s_{jk} is the sign of r_{jk}. In this way, all the co-expression measures are non-negative. We consider $(r_{jj}' = 1, t_j')$ as a pair of outliers and exclude them from the local regression. We use LOWESS [8] to fit the rest *m*-1 pairs. To predict the differential expression measure of X_j, we linearly extend the fitted curve to the right and used the fitted value at $r_{jj}' = 1$ for prediction (predicted differential expression).

Several parameters need to be determined in LOWESS. Except the smoother span *f*, which is usually data-specific, the default values for all the other parameters can be well used for different data sets. The default value ($f = 2/3$ in R function lowess) of the smoother span is usually a good choice. However, based on our analysis experience, a well optimized smoother span *f* can significantly improve the results. We have proposed to determine *f* by maximizing the overall rank correlation between the observed and predicted differential expression measures [7]. The significance of a prediction can be evaluated by the permutation procedure that has been widely used in gene expresion data analysis.

Fig. 1 shows some results from a simulated data set with the configuration m=6000, π_0=0.85, n_1=n_2=15, b=20 and r=1.5 (see below for the detail of simulation configuration). The scatter plots demonstrate the prediction of differential expression for a truly differentially expressed gene and a truly non-differentially expressed gene (f=2/3).

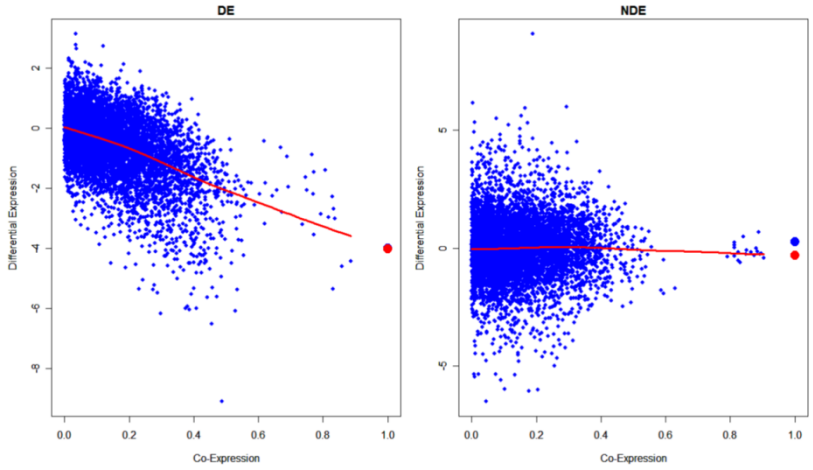

Fig. 1. Illustration of the genome-wide co-expression based prediction of differential expression (simulated data)(Blue and red colors represent observed and fitted measures).

Since the introduction of GSEA [2], [3], it has received much attention. Efron and Tibshirani [9] have recently showed that the *maxmean* is a powerful statistic for GSEA. For simplicity, we consider the *maxmean* statistic for a type 2 diabetes data set and compare the gene set ranking based on two different scoring methods for differential expression analysis: the Student's *t*-test and our recently published method [7]. For simplicity, the smooth span in LOWESS is still $f=2/3$. We observe interesting changes of gene set ranking. It has been shown that the oxidative phosphorylation pathway is associated with type 2 diabetes. There are two gene sets related with this pathway: "MOOTHA_VOXPHOS" and "OXIDATIVE_PHOSPHORYLATION". Their ranks based on the student's *t*-test are 8 and 23, respectively. Based on our method, their ranks are improved to be 5 and 17. Furthermore, we observe that the circadian pathway related gene set ("CIRCADIANPATHWAY") is ranked 14 based on our method. It has been discussed that the study of circadian pathway is promising for understanding diabetes and obesity [10]. However, based on the student's *t*-test, this circadian pathway related gene set is ranked 276 and it is difficult to identify it. Fig. 2 compares the differential expression measurements based on the student's *t*-test and our method. Overall, the scatter plot spreads around the diagonal line. However, the absolute score values of 6 circadian pathway related genes are all increased by our method. Furthermore, the scores of 3 genes are slightly positive based on the student's *t*-test, but our method adjusts them to be negative. Therefore, based on our method, these 6 genes are all coordinately down-regulated in the diabetic sample group, which makes this gene set highly ranked.

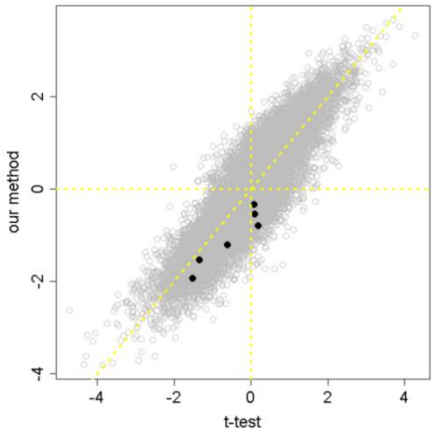

Fig. 2. Comparison of differential expression measures (Six circadian pathway genes are shown in black).

3. Adjustment for Classification Analysis

Our method [7] can be generalized to the classification analysis of microarray gene expression data. The idea is to adjust the observed expression measurements so that these well-developed classification methods [11] can still be used for the supervised learning.

We consider the following approach to adjust gene expression measurements. (Although gene expression measurements are collected for different sample groups, the sample group information is not used in the adjustment procedure. For each gene, all the expression measurements are pooled together in the adjustment procedure.) At the beginning, we perform a simple standardization procedure for each gene. Let X_{ij} be the expression measurement on the i-th array for gene X_j. It is first linearly transformed as: $x_{ji}^s = (x_{ji} - \bar{x}_j)/(s_j)$, where $\bar{x}_j$ and s_j are the sample mean and sample standard deviation of $\{x_{ji}\}_{i=1}^n$. After data standardization, we use the genome-wide co-expression information and the (transformed) expression measurements of other genes to adjust the transformed measurement x_{ji}^s. Based on the transformed data, we use the Pearson's correlation coefficient r_{jk} to measure the co-expression between X_j and another gene X_k, whose transformed expression measurement on the i-th array is x_{ki}^s. Therefore, each gene X_k has a pair of measurements (r_{jk}, x_{ki}^s). To gather more observations for the local regression, we consider a modification: $(r'_{jk}, x'^s_{ki}) = (s_{jk} r_{jk}, s_{jk} x_{ki}^s)$, where s_{jk} is the sign of r_{jk}. In this way, all the co-expression measures are non-negative. We consider $(r'_{jj} = 1, x'^s_{ji})$ as a pair of outliers and exclude them from the local regression. We use LOWESS [8] to fit the rest m-1 pairs. To obtain the adjusted expression measurement $\tilde{x}_{ji}^s$, we linearly extend the fitted curve to the right and used the fitted value at $r'_{jj} = 1$ for prediction (adjusted expression measurement).

We first evaluate the above approach with some simulated data. We considered the impact of different parameters in our simulation studies: (1) gene size, (2) the proportion of differentially expressed genes, (3) sample size, (4) distributions of expression measurements of differentially and non-differentially expressed genes, and (5) covariance structure. In our simulation studies, we consider the widely used block structure: genes are partitioned into many blocks; genes within the same block are positively dependent; and different blocks are independent. Multivariate normal distributions are used to simulate expression measurements in different blocks. We considered different values for the above parameters. Since relatively large sample sizes are usually required for microarray classification analysis, we set n_1=n_2=50. The block size is b=25 and the effect size factor is r=1.5. We simulate 900 (π_0=0.85) differentially expressed genes among m=6000 total genes. At this stage, we assume that 5 genes are known to be truly differentially expressed, and they are randomly selected from these 900 differentially expressed genes. They can be identified in our analysis. The predictive power of these individual genes and their combination (by the linear discriminant analysis, or LDA) can be evaluated by the widely used receiver operative characteristics (ROC) curve. Fig. 3 shows that the predictive power is clearly improved for the adjusted expression measurements based on the comparison of ROC curves. We then evaluate the above approach with an experimental data set. Singh *et al*. [1] accomplished a successful microarray study for prostate cancer. There are 50 normal and 52 cancerous subjects in their published data set. It is well-known that genes *hepsin*, *AMACR* and *GSTP1* are associated with prostate cancer [12]. Fig. 4 shows that our adjusted expression measurements can also clearly improve the predictive power based on the ROC curves.

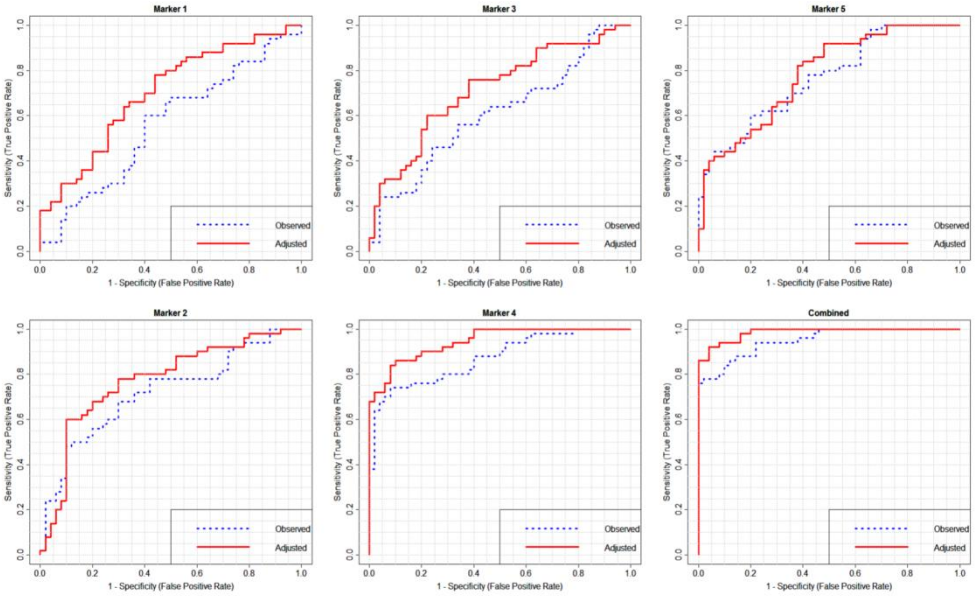

Fig. 3. Receiver operating characteristics (ROC) curves for 5 disease related genes and their combination (based on simulated data).

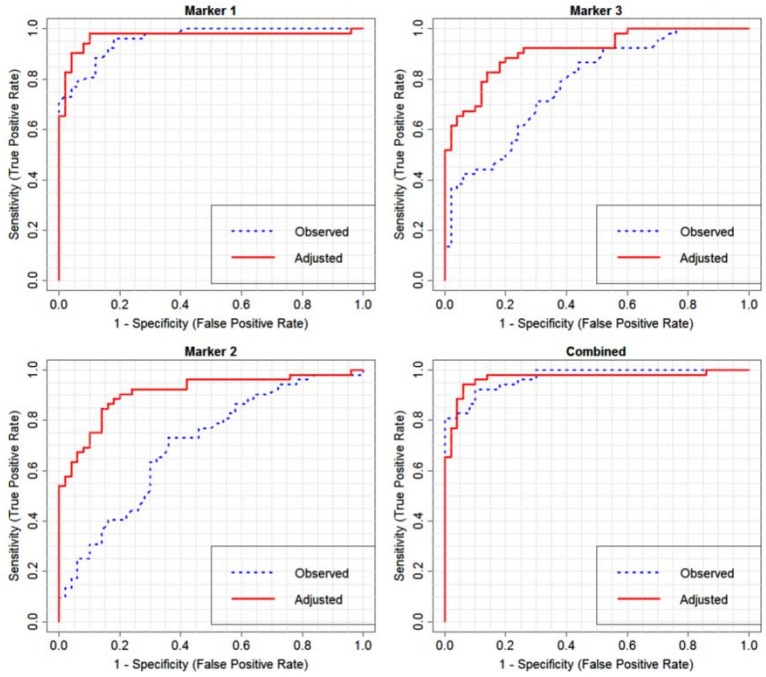

Fig. 4. Receiver operating characteristics (ROC) curves for 3 disease related genes and their combination (prostate cancer data).

In practice, the cross-validation procedure is a rigorous way to evaluate the classification performance. There will be a training set and a test set. The above procedure can be applied to the training set without any modification. But for the test set, we should only use the training set data to adjust the observations in the test set so that the selection bias can be avoided [13]. Therefore, the standardization to the test data will only use these means and standard deviations estimated by the training data; the correlations will also be estimated only based on the training data. We use the widely used support vector machine (SVM) as the classifier with 10-fold cross-validation to evaluate the classification performance of these known disease

related genes in the above simulated data set (five genes) as well as the experimental data set (three genes). (There is no selection of genes.) We observe a clear improvement in classification performance: for the simulated/experimental data, the classification accuracy is 80%/87.3% based on the observed expression measurements and is increased to 90%/91.2% based on the adjusted expression measurements.

4. Exploration of Gene Block Structures

Based on our experience, at least a simple block covariance structure is necessary for our method to achieve a satisfactory performance. To illustrate the applicability of out method to microarray gene expression data, we show the results from a hierarchical clustering analysis for two well-known microarray data sets: one for a prostate cancer study by Singh *et al*. [1] (50 normal and 52 cancerous subjects) and the other for a type 2 diabetes study by Mootha *et al*. [2] (17 normal and 18 diabetic subjects). The distance measure is one minus the Pearson's correlation coefficient, which has been widely used for microarray data analysis. The agglomeration method is "complete linkage" (farthest neighbor). For each sample group (normal or disease), a tree is generated based on the hierarchical clustering method. The height is set as 0.3 (the correlation is 0.7) to cut the trees. We count the number of genes in each partitioned cluster and explore its empirical distribution. Fig. 5 shows clearly that there are block structures in microarray data. The difference in these histograms implies different underlying block structures for different sample groups and different data sets. We also explore the distribution of number of genes in the gene set enrichment analysis. In the Molecular Signatures Database [3], over 1600 gene sets have been collected from various sources such as online pathway databases, publications in PubMed, and knowledge of domain experts (C2: curated gene sets, v2.1). Fig. 6 shows that most gene sets contain about 20~60 genes. Although genes in a gene set are not necessarily highly correlated, Fig. 6 is an indirect evidence to support the applicability of our method.

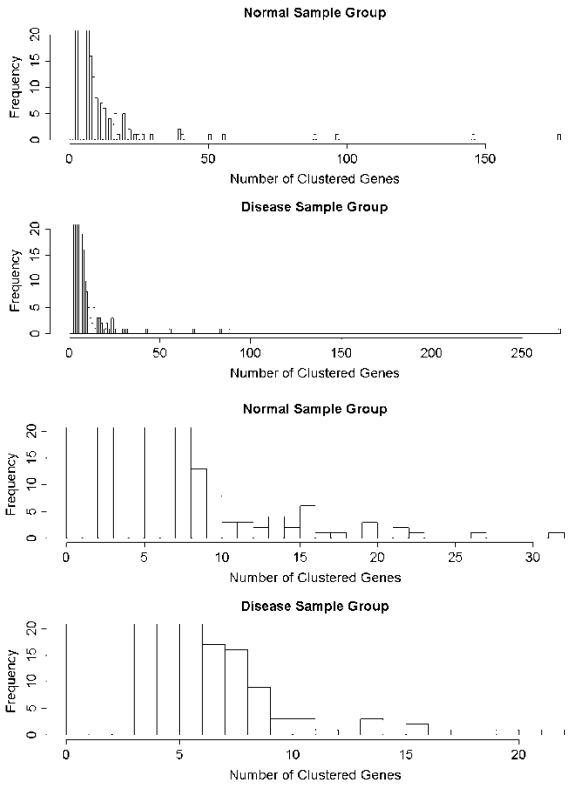

Fig. 5. Exploration of gene block structure based on the microarray data for a prostate cancer study (upper panel) and a type 2 diabetes study (lower panel).

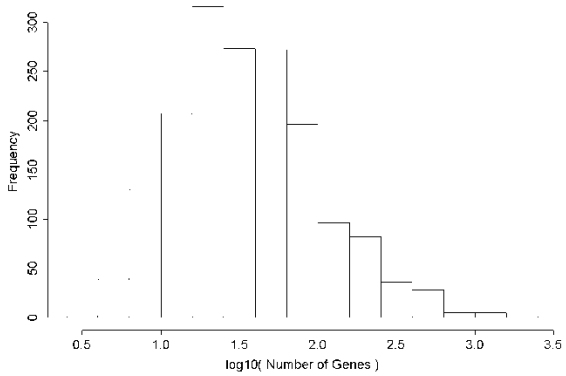

Fig. 6. The distribution of gene set size.

5. Conclusions

Microarrays and deep sequencing technologies have been widely used to understand gene regulations at a genomic scale. Although there are many excellent methods available for different aspects of large scale data analysis, it is still necessary to develop more efficient approaches. For differential expression analysis, it is important to achieve a better control of false discoveries; for gene set enrichment analysis, it is crucial to achieve a better distinction of disease related gene set; for clustering analysis, it is meaningful to observe a clearer separation among different gene clusters; for classification analysis, it is desirable to build a better classifier with less genes and a simpler model. The genome-wide co-expression based analysis can be a solution since this approach efficiently utilizes the genome-wide interaction information.

Although it is well known that genes interact with each other during cellular and molecular processes, there is a lack of efficient multivariate methods for analyzing gene expression data, especially the incorporation of genome-wide interaction information. In this study, we have demonstrated that other types of expression data analysis (gene set enrichment analysis and classification analysis) can be improved when the genome-wide co-expression information is incorporated. It is interesting to further understand the theoretical properties of this approach so that more efficient approaches can be developed. However, this task can be difficult due to the complicated correlation structure in experimental data and also the limited choices of multivariate statistical distributions.

Acknowledgements

The research work was supported by the research fund in the Department of Statistics at The George Washington University.

References

[1] Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., *et al*. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, *1*, 203-209.

[2] Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., *et al*. (2003). PGC-1α-response genes involved in oxidative phos-phorylation are coordinately downregulated in human diabetes. *Nature Genetics*, *34*, 267-273.

[3] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., *et al*. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, *102*, 15545-15550.

[4] Cancer Genome Atlas Research Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*, 1061-1068.

[5] Storey, J. D., Dai, J. Y., & Leek, J. T. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, *8*, 414-432.

[6] Tibshirani, R., & Wasserman, L. (2006). Correlation-sharing for detection of differential gene expression. Technical Report.

[7] Lai, Y. (2008). Genome-wide co-expression based prediction of differential expressions. *Bioinformatics*, *24*, 666-673.

[8] Cleveland, W. S. (1979), Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*, 829-836.

[9] Efron, B., & Tibshirani, R. (2007). On testing the significance of sets of genes. *Annals of Applied Statistics*, *1*, 107-129.

[10] Ramsey, K. M., Marcheva, B., Kohsaka, A., & Bass, J. (2007). The clockwork of metabolism. *Annu. Rev. Nutr.*, *27*, 219-240.

[11] Hastie, T., Tibshirani, R., & Firedman, J. (2001). The elements of statistical learning. *Springer Texts in Statistics*, *6*, 15-57.

[12] DeMarzo, A. M., Nelson, W. G., Isaacs, W. B., & Epstein, J. I. (2003). Pathological and molecular aspects of prostate cancer. *Lancet*, *361*, 955-964.

[13] Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA*, *99*, 6562-6566.

Yinglei Lai received his B.S. degrees in information & computation sciences and business administration in 1999 from the University of Science and Technology of China, and his Ph.D. degree in applied mathematics in 2003 from the University of Southern California. After the postdoctoral training during 2003-2004 at Yale University School of Medicine, he joined the George Washington University as a faculty member. His research areas are statistical problems in the fields of bioinformatics, computational biology and statistical genetics.