[Frontiers in Bioscience 13, 2714-2720, January 1, 2008]

Dimension reduction and mixed-effects model for microarray meta-analysis of cancer

Tianwei Yu[1], Hui Ye[2], Zugen Chen[3], Barry L. Ziober[4], Xiaofeng Zhou[2,5,6]

[1] Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA, [2] Center for Molecular Biology of Oral Diseases, College of Dentistry, University of Illinois at Chicago, Chicago, IL, [3] Department of Human Genetics & Microarray Core, University of California at Los Angeles, Los Angeles, CA, [4] Department of Otorhinolaryngology-Head and Neck Surgery, University of Pennsylvania Health System, Philadelphia, PA, [5] UIC Cancer Center, Graduate College, University of Illinois Chicago, Chicago, IL, [6] Department of Oral and Maxillofacial Surgery, the Second Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Materials and Methods
3.1. Array Hybridization and Data Processing
3.2. Dimension reduction in the context of cancer-associated gene identification
3.3. Mixed effects model for the identification of cancer-associated genes
4. Results and Discussion
4.1. The inter-laboratory effects on microarray data
4.2. Merging the microarray datasets
4.3. Utilizing the mixed-effects model for analyzing merged microarray datasets
5. Acknowledgements
6. References

1. ABSTRACT

The rapid advances in high-throughput microarray technologies greatly facilitate the disease biomarker discovery. However, the potential of these microarray data has not yet been fully utilized. This is partly due to the limited sample sizes of each individual study. Combining microarray data from multiple studies improves the statistical power of detecting differentially expressed genes. Here we present a method for combining the microarray datasets at array probeset level. Using datasets from two commonly used array platforms, the Affymetrix Human Genome U133A and Human Genome U133 Plus 2.0 arrays, we found laboratory effects may be more influential than the platform effect. A visualization scheme for merging the array data from different array platforms was proposed to qualitatively judge the degree of agreement between datasets. A mixed-effects model was applied to identify differentially expressed genes from the merged array data.

2. INTRODUCTION

Microarray gene expression experiments are widely used for the identification of candidate genes and biomarkers for cancer detection, diagnosis, and prognosis. The number of publicly available microarray datasets is increasing rapidly. Combining datasets that address the same disease process offers the opportunity to reliably identify candidate genes across a larger sample set. Meta-analysis approaches that integrating multiple microarray datasets have gained more attention recently, especially for these aimed to identify cancer associated genes. Various studies have been done to compare array platforms (1-5), provide better cross-platform gene matching (6-10), perform cross-platform normalization (11, 12), and test dataset compatibility (13). Lin *et al* developed Reproducibility Probability Score from multi-lab experiments to improve gene selections in future studies (14). To identify differentially expressed genes by merging multiple experiments, methods were developed to combine

expression studies by generating summary statistics from individual p-values (15-17), false discovery rates (FDRs) (18), or gene ranks (19-21). In addition to merging summary statistics, some integrative modeling approaches were developed. Shen *et al.* used a mixture model approach to define an inter-study "meta-signature" (22). Huang *et al.* developed weighted regression methods to combine multiple datasets (23). Different variations of the effect-size approach, which models the effect-size from each study using a mixed effects model, were widely used (24-28). Park *et al.* proposed an ANOVA model that includes the lab effects and a two-stage method for model fitting (29).

In this study, we focused on two Affymetrix array platforms, the Human Genome U133A array and the Human Genome U133 Plus 2.0 array. We chose these array platforms as they have been widely utilized for expressional analyses, and massive amount of array datasets have been generated based on these array platforms. We combined in-house generated data with publicly available data from both platforms to study the inter-platform and the inter-laboratory variation. We then developed a dimension reduction scheme to visualize combined data that is resistant to inter-laboratory variation. A linear mixed effects model was used to identify differentially expressed genes from the combined data.

3. MATERIALS AND METHODS

3.1. Array Hybridization and data processing

The RNA samples from skin fibroblasts GM00302, GM05386 and GM04522 (Coriell Cell Repositories/NIGMS) were utilized to generate the array dataset using both Human Genome U133A and Human Genome U133 Plus 2.0 GeneChip arrays. A total of 150 to 200 ng of purified total RNA was utilized to generate cRNA probes according to standard Affymetrix protocols. The quantity and the purity of biotinylated cRNA were determined by spectrophotometry, and an aliquot of the sample was checked by gel electrophoresis. The sample was hybridized to the Affymetrix Human Genome U133A or Human Genome U133 Plus 2.0 GeneChip arrays (Affymetrix, Santa Clara, CA) according to Affymetrix protocols. The arrays were scanned with a GeneChip Scanner 3000. The scanned array images were processed with GeneChip Operating software (GCOS). The data were processed using Robust Multiarray Analysis (RMA) (30). Two additional datasets from previously published studies of oral tongue squamous cell carcinomas (OTSCC) were also used, including 1) the U133A array dataset of 7 OTSCC samples and 7 normal samples from Ziober *et al.* study (31), and 2) the U133 Plus 2.0 array dataset of 15 OTSCC samples and 7 normal samples from Zhou *et al.* study (32).

3.2. Dimension reduction in the context of cancer-associated gene identification

From the data generation point of view, the hidden factors which contributing to the data structure may contain several categories. The major categories include: (1) gene expression dependencies unrelated to disease status, i.e. normal physiological dependencies; (2) gene expression variation caused by laboratory/experiment–specific sample treatment, i.e. tissue preservation, RNA extraction reagents/procedure etc; (3) gene expression changes related to disease process (e.g., cancer); and (4) gene expression dependencies specific to the person (contributions of individual genetic background). In microarray data analysis, personal effects are implicitly absorbed by the error term in various models, unless paired data is available. In the current discussion we assumed the data is generated from unpaired samples. Considering the first three categories of factors, for each gene i, we can describe its expression level in sample j of experiment k by a factor model. Here a linear additive relationship is assumed between the categories of factors for simplicity. Depending on the transformations performed on the *y* values, it can mean multiplicative or other relationships.

$$y_{ijk} = \mu_{ik} + f_i(\boldsymbol{u}) + g_i(\boldsymbol{v}(k)) + h_i(x_j) + \varepsilon_{ijk}$$
$$\varepsilon_{ijk} \sim N(0, \sigma_i)$$
$$x_j = \begin{cases} 1, & \text{cancer} \\ 0, & \text{normal} \end{cases}$$

In this model, the vector $\boldsymbol{u}$ represents regular physiological factors; the vector $\boldsymbol{v}(k)$ represents lab-related factors that are specific to the lab (or experiment batch within the same lab) *k*; x_j represents the cancer factor; and ε is the error term. After proper data normalization is performed, we assume $h_i()$ and σ_i to be the same across labs/experiments. It is not our goal to conduct estimation and inference on the factor model. Rather, this model is only used for a conceptual explanation of our procedure. Suppose in the k^{th} experiment, there are N_k normal samples, the mean expression level of gene *i* in the normal group will follow:

$$\bar{y}_{ik,normal} = \mu_{ik} + f_i(\boldsymbol{u}) + g_i(\boldsymbol{v}(k)) + h_i(\boldsymbol{0}) + \varepsilon'_{ijk},\ \varepsilon'_{ijk} \sim N(0, \sigma_i/\sqrt{N_k})$$

If we subtract $\bar{y}_{ik,normal}$ from every expression value of gene *i* in experiment k, for normal samples, we have:

$$\Delta y_{ijk} = y_{ijk} - \bar{y}_{ik,normal} = \varepsilon''_{ijk},\ \varepsilon''_{ijk} \sim N(0, \sigma_i\sqrt{(N_k-1)/N_k})$$

For cancer samples, we have:

$$\Delta y_{ijk} = y_{ijk} - \bar{y}_{ik,normal} = h_i(1) - h_i(0) + \varepsilon'''_{ijk},$$
$$\varepsilon'''_{ijk} \sim N(0, \sigma_i\sqrt{(N_k+1)/N_k})$$

After the subtraction within each experiment, we eliminated the lab/experiment–dependent term $\boldsymbol{v}(k)$. The ratio between $\sqrt{(N_k-1)/N_k}$ and $\sqrt{(N_k+1)/N_k}$ reaches 0.8 when N_k is 5, and 0.9 when N_k is 10. We can ignore

their difference for the purpose of dimension reduction and visualization. By subtracting the normal mean on a gene-by-gene basis from the same experiment, the normal data becomes noise, and the cancer data contains the information of the cancer–dependent function that is not reliant on the lab/experiment. We can combine the data from different labs/experiments at this stage, and use dimension reduction techniques to extract information about the relationship between the genes' expression with cancer. Up to now we have assumed *hi()* to be arbitrary and gene-specific. In the current study, where there are only two groups "cancer" and "normal", $h_i(1)-h_i(0)$ reveals the difference in first-order. If more groups are under study, all groups are normalized against the normal group. This would preserve all the between-group differences. For the purpose of cancer-associated gene identification, our major interest is to find genes that linearly associate with cancer status. Thus, the remaining signal in the cancer data, $h_i(1)-h_i(0)$, retains the desired information. In this study we performed the Principal Component Analysis (PCA) on the combined data.

3.3. Mixed effects model for the identification of cancer-associated genes

Park *et al.* (29) proposed using the ANOVA model with hospital/laboratory effects. They applied a two step fitting procedure, the first step of which is to remove the uninteresting effects. When the design is unbalanced, i.e. there are unequal numbers of normal/cancer samples from each laboratory, this step may lead to biased estimates. Consistent with our visualization scheme, we take two steps to identify differentially expressed genes: (1) quantile–normalizing the array data at the gene/probeset level; (2) fitting a mixed effect model for each gene *i*:

$$y_{ijk} = \mu_i + \alpha_{ik} + \beta_i x_j + \varepsilon_{ijk}, \quad \alpha_{ik} \sim N(0,\delta_i), \varepsilon_{ijk} \sim N(0,\sigma_i)$$

$$x_j = \begin{cases} 1, \text{cancer} \\ 0, \text{normal} \end{cases}$$

In the model, the random effect α_{ik} is the physiological and laboratory effect combined, and β_i is the first-order cancer effect, which is our major interest in the identification of cancer–associated genes. The purpose of the quantile-normalization is to make the effect size similar between different datasets. The model fitting can be easily done in most statistical softwares, providing the overall effect size estimate $\hat{\beta}_i$ and its significance.

4. RESULTS AND DISCUSSION

In this study, we focused on two Affymetrix microarray platforms, Human Genome U133A and Human Genome U133 Plus 2.0. The probesets in U133A are a subset of those in U133 Plus 2.0. Hence the merged data contains 100% of the probesets in the U133A array and 41% of the probesets in the U133 Plus 2.0 array. It has been shown that gene matching between different platforms (e.g., oligo array vs. cDNA array) may have substantial influence on the level of correlations (9). By using data from the two Affymetrix platforms, for which gene matching is trivial, we eliminated the possible gene-matching effect and focused on the analysis of the combined data.

4.1. The inter-laboratory effects on microarray data

The gene expression analyses were performed on 3 skin fibroblast primary cultures, GM00302, GM05386 and GM04522, using both array platforms. GM00302 was analyzed twice independently using both platforms (2 U133A arrays, and 2 U133 Plus 2.0 arrays). By plotting the correlation coefficients (Figure 1a), we found that between these two array platforms, when sample preparation was well-controlled, the inter-platform consistency is comparable to intra-platform consistency. We then analyzed two datasets from published studies of oral tongue squamous cell carcinomas (OTSCC) of the tongue. The U133A array dataset of 7 OTSCC samples and 7 normal samples from Ziober *et al.* study (31), and the U133 Plus 2.0 array dataset of 15 OTSCC samples and 7 normal samples from Zhou *et al.* study (32) were used. Although some of the samples were paired, we performed analyses as if they were unpaired for simplicity. We compared the correlation structure across the samples (Figure 1b). Comparatively, the inter-study correlations were much lower, although they were still high in absolute terms (~0.7). By comparing Figure 1a and Figure 1b, we observed that the inter-laboratory difference was a crucial factor affecting the analyses of expression data. It was not a simple baseline-shifting problem, as a baseline-shift would not have substantially influenced the correlations. Rather, different probesets were influenced differently by the laboratory effects. Since the array hybridization protocols are relatively standardized, the observed laboratory effects were most likely caused by differences in experimental procedures, such as tissue procurement, RNA isolation, and sample preparation. In situations where datasets are generated from completely different microarray platforms (e.g., cDNA microarray vs. oligo array), we would expect higher platform effects. Nevertheless, the platform effects and the lab effects would still be confounded. If we can assume the two effects have a close to additive relationship, the same data processing procedure can still be applied.

4.2. Merging the microarray datasets

We merged the two datasets by extracting overlapping probesets and performed Principle Component Analysis (PCA) (Figure 2a). We found that the second principal component (PC) 2 summarized the inter-laboratory difference, while PC3 contained some information of cancer/normal differences. After quantile-normalization, the inter-laboratory difference still dominated PC1 and PC2

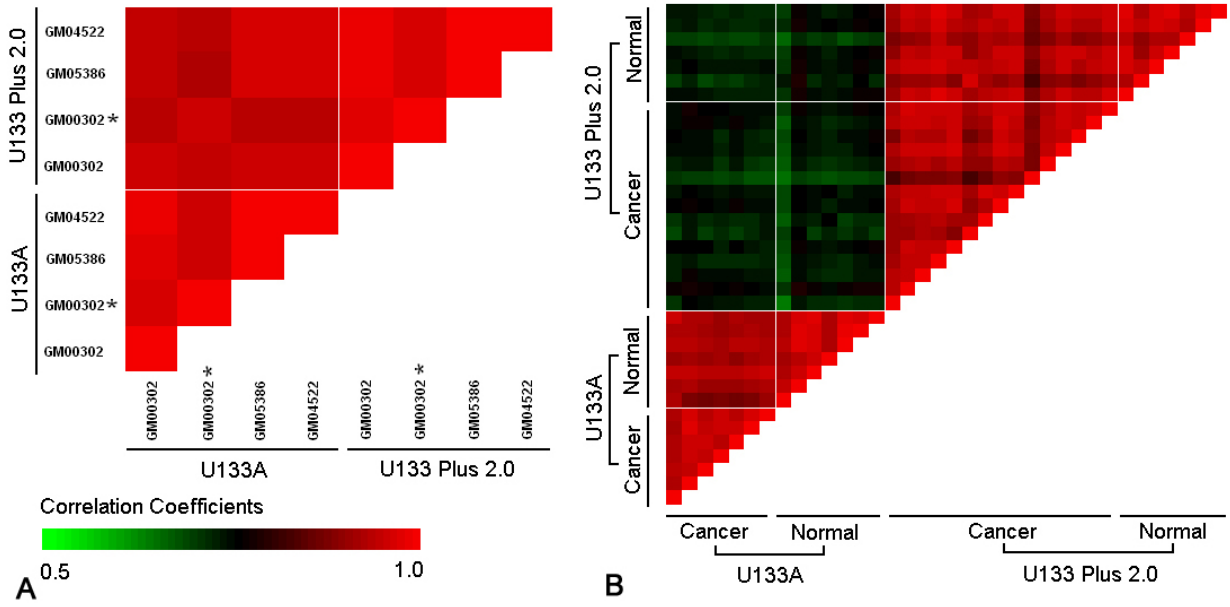

Figure 1. Correlation between datasets from Human Genome U133A and Human Genome U133 Plus 2.0 arrays. All array datasets were processed with Robust Multiarray Analysis (RMA) to generate probeset-level expression values. Correlation coefficients from overlapping probesets between arrays were visualized using heatmap. (a) Three cell lines measured by both U133A and U133 Plus 2.0 arrays. GM00302 was analyzed twice independently using both platforms. The second analysis of GM00302 is labeled with a star. (b) Cancer/normal tissue samples measured by U133A (Ziober study) and U133 Plus 2.0 (Zhou study) arrays.

(Figure 2b). We then removed the normal group mean within each dataset on a probeset-by-probeset basis. As discussed in the Methods section, this operation removed most of the non-cancer physiological effects and laboratory effects. The first PC was dominated by cancer effect after the mean removal (Figure 2c). In addition, we found that the samples from the two laboratories (triangles and spheres) were mingled together, which indicated the two datasets contained similar cancer-related information. To confirm this point, we first permuted the sample labels within the U133A dataset, before subjecting the combined data to quantile-normalization and normal-mean removal. After permuting the sample labels, the normal-mean becomes an estimate of the overall mean. All the U133A samples, regardless of cancer/normal status, grouped with the normal samples in the U133 Plus 2.0 dataset (Figure 2d). Further, we permuted half of the probeset names in the U133A dataset before combining the two datasets. The purpose of this step was to artificially create the scenario that the two datasets contained different cancer-related genes, with a certain amount of overlap. The resulting PCA plot showed that in addition to cancer/normal separation, cancer samples from the two datasets were also separated, while all the normal samples were mingled together (Figure 2e). Overall, the results showed that the separation of cancer samples in the PCA plot could be an indication of strong disagreement between datasets.

4.3. Utilizing the mixed-effects model for analyzing merged microarray datasets

We applied the mixed-effects model to identify OTSCC associated genes from the two datasets, and compared the results to those obtained by t-test from individual datasets (Figure 3). All the p-values were adjusted to FDR based on the method described previously (33). Genes were selected at the FDR level of 0.05 and effect size $\geq$ 2 fold. Using the U133A data*set alone*, 91 genes were selected, 48 of which were also identified in the meta-analysis. Using the U133 Plus 2.0 data*set alone*, 33 genes were selected, 30 of which were also identified in the meta-analysis. A total of 343 extra genes were selected by pooling the two datasets, indicating a much higher statistical power by combining datasets. The better agreement between the U133 Plus 2.0 dataset and the pooled data was caused partially by the larger size of the U133 Plus 2.0 dataset (22 arrays, compared to 14 arrays from the U133 dataset). Details on these identified candidate genes will be presented in a manuscript in preparation.

In summary, we presented a simple dimension reduction procedure for the visualization of combined microarray data. After quantile-normalization, the mean normal-group expression vector of each dataset was subtracted from every

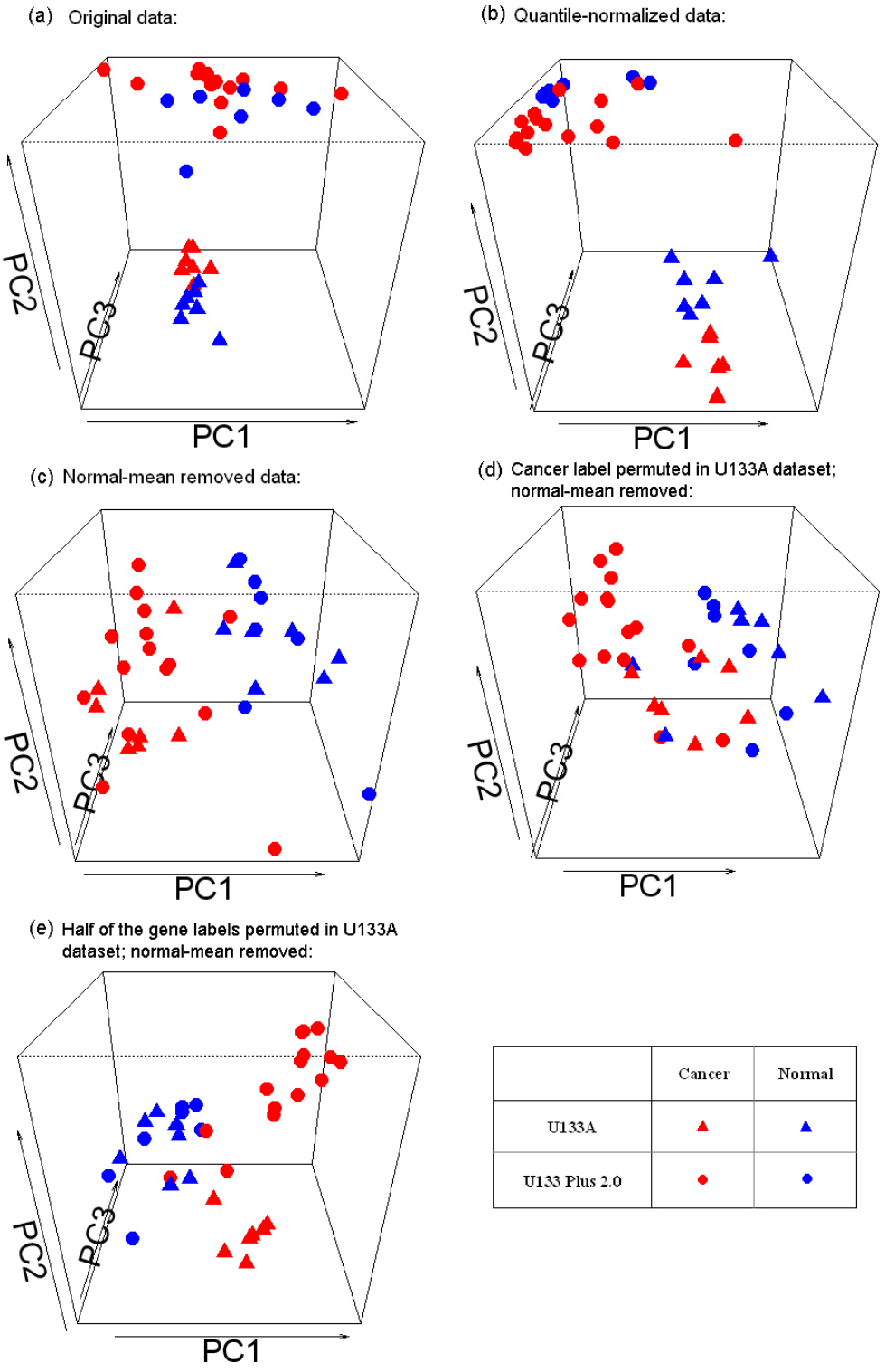

Figure 2. Principal component analysis on original and transformed data. The first three principal components (PC) of the merged data from the U133A dataset (Ziober study) and the U133 Plus 2.0 dataset (Zhou study) were plotted. (a) Original data; (b) quantile-normalized data; (c) the data after quantile-normalization and the removal of normal group mean; (d) tissue type labels were permuted within each dataset, then quantile-normalization and normal-mean removal were performed; (e) half of the probeset labels were permuted in the U133A dataset, then quantile-normalization and normal-mean removal were performed.

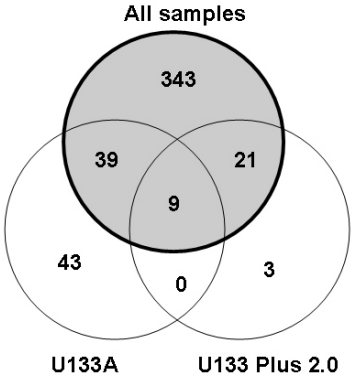

Figure 3. The differentially expressed genes in OTSCC identified by different array datasets. The differentially expressed genes were selected based on the U133A dataset (Ziober study) and the U133 Plus 2.0 dataset (Zhou study) using the criteria of FDR=0.05 and fold change ≥ 2. All samples: genes identified by combining both datasets; U133A: genes identified using the U133A data alone; U133 Plus 2.0: genes identified using the U133 Plus 2.0 data alone.

sample within that dataset before data merging. When the agreement between the datasets is poor, the visualization scheme could help reveal the discrepancy. In accordance with the visualization scheme, we used the mixed effects model including random lab effects to find cancer-associated genes. Such a model is well-established and the fitting is straight-forward.

5. ACKNOWLEDGEMENTS

This work was supported in part by NIH PHS grants DE014847, DE016569, CA114688 (to X. Zhou). We thank Ms. Katherine Long for excellent editorial assistance.

6. REFERENCES

1. Chris Cheadle, Kevin G. Becker, Yoon S. Cho-Chung, Maria Nesterova, Tonya Watkins, William Wood, 3rd, Vinayakumar Prabhu & Kathleen C. Barnes: A rapid method for microarray cross platform comparisons using gene expression signatures. *Mol Cell Probes*, 21, 35-46(2007)

2. MAQC Consortium: The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, 24, 1151-61(2006)

3. Rafael A. Irizarry, Daniel Warren, Forrest Spencer, Irene F. Kim, Shyam Biswal, Bryan C. Frank, Edward Gabrielson, Joe G.N. Garcia, Joel Geoghegan, Gregory Germino, Constance Griffin, Sara C. Hilmer, Eric Hoffman, Anne E. Jedlicka, Ernest Kawasaki, Francisco Martinez-Murillo, Laura Morsberger, Hannah Lee, David Petersen, John Quackenbush, Alan Scott, Michael Wilson, Yanqin Yang, Shui Qing Ye & Wayne Yu: Multiple-laboratory comparison of microarray platforms. *Nat Methods*, 2, 345-50(2005)

4. Joerg Schlingemann, Negusse Habtemichael, Carina Ittrich, Grischa Toedt, Heidi Kramer, Markus Hambek, Rainald Knecht, Peter Lichter, Roland Stauber & Meinhard Hahn: Patient-based cross-platform comparison of oligonucleotide microarray expression profiles. *Lab Invest*, 85, 1024-39(2005)

5. Leming Shi, Weida Tong, Hong Fang, Uwe Scherf, Jing Han, Raj K. Puri, Felix W. Frueh, Federico M. Goodsaid, Lei Guo, Zhenqiang Su, Tao Han, James C. Fuscoe, Alex Xu, Tucker A. Patterson, Huixiao Hong, Qian Xie, Roger G. Perkins, James J. Chen & Daniel A. Casciano: Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics*, 6 Suppl 2, S12(2005)

6. Jussi Paananen, Markus Storvik & Garry Wong: CROPPER: a metagene creator resource for cross-platform and cross-species compendium studies. *BMC Bioinformatics*, 7, 418(2006)

7. Scott L. Carter, Aron C. Eklund, Brigham H. Mecham, Isaac S. Kohane & Zoltan Szallasi: Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics*, 6, 107(2005)

8. Laura L. Elo, Leo Lahti, Heli Skottman, Minna Kylaniemi, Riitta Lahesmaa & Tero Aittokallio: Integrating probe-level expression changes across generations of Affymetrix arrays. *Nucleic Acids Res*, 33, e193(2005)

9. Yuan Ji, Kevin Coombes, Jiexin Zhang, Sijin Wen, James Mitchell, Lajos Pusztai, Fraiser Symmans & Jing Wang: RefSeq refinements of UniGene-based gene matching improve the correlation of expression measurements between two microarray platforms. *Appl Bioinformatics*, 5, 89-98(2006)

10. Brigham H. Mecham, Gregory T. Klus, Jeffrey Strovel, Meena Augustus, David Byrne, Peter Bozso, Daniel Z. Wetmore, Thomas J. Mariani, Isaac S. Kohane & Zoltan Szallasi: Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res*, 32, e74(2004)

11. Patrick Warnat, Roland Eils & Benedikt Brors: Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6, 265(2005)

12. Bjorn Nilsson, Anna Andersson, Mikael Johansson & Thoas Fioretos: Cross-platform classification in microarray-based leukemia diagnostics. *Haematologica*, 91, 821-4(2006)

13. Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Scholkopf & Alex J. Smola: Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, 22, e49-57(2006)

14. Guixian Lin, Xuming He, Hanlee Ji, Leming Shi, Ronald W. Davis & Sheng Zhong: Reproducibility Probability Score--incorporating measurement variability across laboratories for gene selection. *Nat Biotechnol*, 24, 1476-7(2006)

15. Yves Moreau, Stein Aerts, Bart De Moor, Bart De Strooper & Michal Dabrowski: Comparison and meta-

analysis of microarray data: from the bench to the computer desk. *Trends Genet*, 19, 570-7(2003)
16. Daniel R. Rhodes, Terrence R. Barrette, Mark A. Rubin, Debashis Ghosh & Arul M. Chinnaiyan: Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res*, 62, 4427-33(2002)
17. Chiara Romualdi, Cristiano De Pitta, Lucia Tombolan, Stefania Bortoluzzi, Francesca Sartori, Angelo Rosolen & Gerolamo Lanfranchi: Defining the gene expression signature of rhabdomyosarcoma by meta-analysis. *BMC Genomics*, 7, 287(2006)
18. Daniel R. Rhodes, Jianjun Yu, K Shanker, Nandan Deshpande, Radhika Varambally, Debashis Ghosh, Terrence Barrette, Akhilesh Pandey & Arul M. Chinnaiyan: Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A*, 101, 9309-14(2004)
19. Lei Xu, Aik Choon Tan, Daniel Q. Naiman, Donald Geman & Raimond L. Winslow: Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, 21, 3905-11(2005)
20. Robert P. DeConde, Sarah Hawley, Seth Falcon, Nigel Clegg, Beatrice Knudsen & Ruth Etzioni: Combining results of microarray experiments: a rank aggregation approach. *Stat Appl Genet Mol Biol*, 5, Article15(2006)
21. Fangxin Hong, Rainer Breitling, Connor W. McEntee, Ben S. Wittner, Jennifer L. Nemhauser & Joanne Chory: RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22, 2825-7(2006)
22. Ronglai Shen, Debashis Ghosh & Arul M. Chinnaiyan: Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics*, 5, 94(2004)
23. Xiaohong Huang, Wei Pan, Xinqiang Han, Yingjie Chen, Leslie W. Miller & Jennifer Hall: Borrowing information from relevant microarray studies for sample classification using weighted partial least squares. *Comput Biol Chem*, 29, 204-11(2005)
24. Jung Kyoon Choi, Jong Young Choi, Dae Ghon Kim, Dong Wook Choi, Bu Yeo Kim, Kee Ho Lee, Young Il Yeom, Hyang Sook Yoo, Ook Joon Yoo & Sangsoo Kim: Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Lett*, 565, 93-100(2004)
25. Jung Kyoon Choi, Ungsik Yu, Sangsoo Kim & Ook Joon Yoo: Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19 Suppl 1, i84-90(2003)
26. Robert Grutzmann, Hinnerk Boriss, Ole Ammerpohl, Jutta Luttges, Holger Kalthoff, Hans Konrad Schackert, Gunter Kloppel, Hans Detlev Saeger & Christian Pilarsky: Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene*, 24, 5079-88(2005)
27. Pingzhao Hu, Celia M. Greenwood & Joseph Beyene: Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinformatics*, 6, 128(2005)
28. John R. Stevens & RW Doerge: Combining Affymetrix microarray results. *BMC Bioinformatics*, 6, 57(2005)
29. Taesung Park, Sung-Gon Yi, Young Kee Shin & Seung Yeoun Lee: Combining multiple microarrays in the presence of controlling variables. *Bioinformatics*, 22, 1682-9(2006)
30. Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs & Terence P. Speed: Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, 31, e15(2003)
31. Amy F. Ziober, Kirtesh R. Patel, Faizan Alawi, Phyllis Gimotty, Randall S. Weber, Michael M. Feldman, Ara A. Chalian, Gregory S. Weinstein, Jennifer Hunt & Barry L. Ziober: Identification of a gene signature for rapid screening of oral squamous cell carcinoma. *Clin Cancer Res*, 12, 5960-71(2006)
32. Xiaofeng Zhou, Stephane Temam, Myungshin Oh, Nisa Pungpravat, Bau Lin Huang, Li Mao & David T. Wong: Global expression-based classification of lymph node metastasis and extracapsular spread of oral tongue squamous cell carcinoma. *Neoplasia*, 8, 925-32(2006)
33. Yoav Benjamini & Yosef Hochberg: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57, 289–300(1995)

Abbreviations: PCA: Principal Component Analysis; OTSCC: oral tongue squamous cell carcinomas

Key Words: Dimension reduction, Principal Component Analysis, Mixed effects model, Microarray

Send Correspondence to: Dr Tianwei Yu, Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, Tel: 404-727-7671, Fax: 404-727-1370, E-mail: tyu8@emory.edu

http://www.bioscience.org/current/vol13.htm