# New Statistics for Testing Differential Expression of Pathways from Microarray Data

Hoicheong Siu[1,2], Hua Dong[1,2,3], Li Jin[1,2], and Momiao Xiong[1,3]

[1] Laboratory of Theoretical Systems Biology, School of Life Science, Fudan University,
Shanghai, 200433, China
[2] State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary
Anthropology, School of Life Sciences and Institutes of Biomedical Sciences,
Fudan University, Shanghai, 200433, China
[3] Human Genetics Center, University of Texas School of Public Health, Houston, TX 77225
{hoejohn,hdong0425,ljin007,momiao}@gmail.com

**Abstract.** Exploring biological meaning from microarray data is very important but remains a great challenge. Here, we developed three new statistics: linear combination test, quadratic test and de-correlation test to identify differentially expressed pathways from gene expression profile. We apply our statistics to two rheumatoid arthritis datasets. Notably, our results reveal three significant pathways and 275 genes in common in two datasets. The pathways we found are meaningful to uncover the disease mechanisms of rheumatoid arthritis, which implies that our statistics are a powerful tool in functional analysis of gene expression data.

**Keywords:** microarray, pathway, linear combination test, quadratic test, de-correlation test, rheumatoid arthritis.

## 1 Introduction

Understanding biological implication of gene expression profiles is important but challenging. A popular approach to gene expression data analysis is to identify a number of differentially expressed genes. Although such approach is useful for uncovering principles underlying biological processes, it has at least two limitations. First of all, in many cases, after Bonferroni correction, only a few individual genes may meet the threshold for statistical significance, because the relevant biological effects are modest relative to the noise inherent to the microarray technology. Second, one may be left with a long list of statistically significant genes without any unifying biological theme. Genes carry out their functions via intricate pathways of reactions and interactions. Pathways are sets of genes that act together to achieve certain cellular or physiologic functions. Prioritizing pathways relevant to a particular phenotype can help researchers to focus on the subset of most relevant genes, and generate further biological hypotheses. Genes belonging to the same pathway often exhibit subtle, coordinated changes in their expressions. Alternative to using a gene as a unit for analyzing expression profiles is to take a pathway as a unit for gene expression data

J. Zhou (Ed.): Complex 2009, Part I, LNICST 4, pp. 277 – 285, 2009.

analysis. Gene Set Enrichment Analysis (GSEA) which was developed by examining the overall differences in expression patterns between predefined gene sets and the whole gene list on the array is a useful tool for perform-ing pathway analysis [1]. Most methods for GSEA assume that the expressions of the genes within a pathway are independent [2,3]. However, in reality, the expressions of the genes within a pathway are correlated. Ignoring correlation among genes within a pathway may lead to misleading results.

Purpose of this paper is to develop statistics for GSEA which take correlations among gene expression into account. To accomplish this goal, we first investigated correlations among genes within the pathway and find that correlations among genes cannot be ignored. This motives us to develop three novel statistics which are able to combine dependent P-values of genes within the pathway. Finally, we apply the developed statistics to two rheumatoid arthritis gene expression datasets. Our method revealed pathways involved in rheumatoid arthritis disease, some of which have been independently validated by other microarray studies and by in vivo functional studies.

## 2 Methods

A pathway-based differential expression analysis is to use a pathway as the basic unit of analysis. Instead of testing differential expression of single gene between normal and abnormal tissues, pathway-based differential expression analysis is to jointly test for differential expressions of all genes within the pathway. Formally, suppose that there are k genes in the pathway. The null hypothesis for testing differential expression of the ith gene in the pathway is represented by:

$$H_{i0} : \theta_i = \theta_{i0} \tag{1}$$

where $\theta_i$ denotes the parameter, e. g., the difference of expression value between cases and controls. Then, the null hypothesis for testing differential expression of a pathway between normal and abnormal tissues is defined as testing for the combined null hypothesis:

$$H_{i0} : \theta_i = \theta_{i0}, i = 1,2,\ldots,k. \tag{2}$$

The alternative hypothesis is defined as $H_{ia} : \theta_i \neq \theta_{i0}$, for at least one gene.

In this report, we will focus on combining individual differential expression tests of genes. We developed methods for combining dependent P-values which take correlations among gene expressions into account.

Let $n_A$ be the number of affected tissues and $n_G$ be the number of normal tissues. There are k genes in the pathway, define the mean expressions of *i*-th gene in cases and controls, respectively, as:

$$\overline{X}_i = \frac{1}{n_A}\sum_{j=1}^{n_A} x_{ij} \text{ and } \overline{Y}_i = \frac{1}{n_G}\sum_{j=1}^{n_G} y_{ij} \tag{3}$$

$x_{ij}$ is the expression level of gene *I* from the *j*-th abnormal tissue, and $y_{ij}$ is the expression level of the gene *I* from the *j*-th normal tissue. Let

$$\overline{X} = \begin{bmatrix} \overline{X}_1 \\ \vdots \\ \overline{X}_k \end{bmatrix}, \overline{Y} = \begin{bmatrix} \overline{Y}_1 \\ \vdots \\ \overline{Y}_k \end{bmatrix}, X_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{kj} \end{bmatrix} \text{ and } Y_j = \begin{bmatrix} y_{1j} \\ \vdots \\ y_{kj} \end{bmatrix} \tag{4}$$

then the sampling covariance matrix of the genes in the pathway defined as:

$$S = \frac{\sum_{j=1}^{n_A} (X_j - \overline{X})(X_j - \overline{X})^T + \sum_{j=1}^{n_G} (Y_j - \overline{Y})(Y_j - \overline{Y})^T}{n_A + n_G - 2} \tag{5}$$

where S has a form like $S = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1K} \\ S_{21} & S_{22} & \cdots & S_{2K} \\ \cdots & \cdots & \cdots & \cdots \\ S_{K1} & S_{K2} & \cdots & S_{KK} \end{bmatrix}$. Let $d_{ii} = \frac{1}{\sqrt{S_{ii}}}$, then

$$D = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & d_{KK} \end{bmatrix} \text{ and } \mathrm{R} = \mathrm{DSD} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix} \tag{6}$$

R is the correlation matrix of the genes in the pathway.

The statistic for testing differential expression of gene *i* is defined as:

$$T_i = \frac{\overline{X}_i - \overline{Y}_i}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_G}\right) S_{ii}}} \tag{7}$$

$T_i$ follows a student distribution with degree freedom $n_A + n_G - 2$. Denote its cumulative distribution by $F(T_i)$. Define the transformation as:

$$z_i = \Phi^{-1}(F(T_i)) \tag{8}$$

where $\Phi$ is a standard normal cumulative distribution and $z_i \sim N(0,1)$.

We define three statistics for testing differential expression of a pathway. Let

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_k \end{bmatrix} \text{ and } e = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \tag{9}$$

(1) Linear Combination Test:

The first test statistic is linear combination test, which is defined as:

$$T_L = \frac{e^T Z}{\sqrt{e^T \operatorname{Re}}} \tag{10}$$

$T_L$ follows a standard normal distribution.

(2) Quadratic Test

The second statistic is based on the quadratic form of Z and defined as:

$$T_Q = Z^T R^{-1} Z \tag{11}$$

$T_Q$ is asymptotically distributed as a central $\chi^2_{(k)}$ distribution, where k is the number of genes in the pathway.

(3) Decorrelation Test

We decompose the matrix R as $R = CC^T$ and let $T = [T_1 \quad ... \quad T_k]^T$, and define the de-correlated statistics $W = C^{-1}T = [W_1, \quad \cdots, \quad W_k]^T$, which are asymptotically distributed as a vector of independent standard normal variables. For each $W_i$, we calculate its P-value $P_i$. Define the statistic:

$$T_D = -2\sum_{i=1}^{k} \log P_i \tag{12}$$

$T_D$ follows a $\chi^2_{(2k)}$ distribution, where k is the number of genes in the pathway.

## 3 Results

### 3.1 Data Filtering and Statistical Analysis

We chose two gene expression datasets of rheumatoid arthritis by two standards: 1) case-control study: both rheumatoid arthritis patients and health controls are included; and 2) More than 45 samples are included. Datasets I was downloaded from Gene Expression Omnibus [4]. It contains 46 samples composed of 35 cases and 11 controls, using Affymetrix GeneChip Human Genome U95 Set HG-U95A array on which including 8685 genes [5]. Since it is one color DNA chip, we use the abstract intensity value for calculation. Datasets II is downloaded from Stanford MicroArray Database [6], which is composed of 35 cases and 15 controls, totally 14337 genes. As it is a two-color cDNA array, we extracted the log2 rations of gene expression values for our analysis [7]. 7434 genes shared between the two arrays. Gene represented more than once on the microarrays were averaged. The differential expression of the gene was tested by Mann-Whitney Test as the distribution of gene expression is unknown. Total 2069 genes in dataset I, 2003 genes in dataset II and 275 genes in both datasets showed mild differential expression with P-value < 0.05. However, after Bonferroni correction for multiple tests (P < 5.75705E-06 for dataset I and

**Table 1.** Significant differentially expressed genes in rheumatoid arthritis studies

| | P-value | Number of genes |
|---|---|---|
| Dataset I | <0.05 | 2069 |
| | <5.75705E-6* | 7 |
| Dataset II | <0.05 | 2003 |
| | <3.48748E-6* | 0 |
| Combined 2 datasets | <0.05 | 275 |
| | <6.72585E-6* | 0 |

* indicates Bonferroni correction was applied to adjust for multiple test corrections.

P < 3.48748E-06 for dataset II), 7 genes were significant in dataset I but none gene was significant in dataset II.

## 3.2 Pathway Analysis

Since number of significantly differentially expressed genes identified were not large enough to explore disease mechanisms. Genes belonging to the same pathway often exhibit subtle, coordinated changes in the expression profile. Pathway analysis can detect genes which conferred small disease risk individually, but whose joint actions can be implicated in the development of disease. In this paper, we collected pathways of human and studied with 501 of them included more than 2 genes, 202 from KEGG [8] and 299 from BioCarta [9] (updated until Dec 2008). The three statistics: linear combination test, quadratic test, de-correlation test were used to identify the differentially expressed pathways.

### 3.2.1 Correlation Structure

To examine whether correlation among genes in a pathway can be ignored or not, for example we calculate the correlation among genes in the Expression Role of PPAR-gamma Coactivators in Obesity and Thermogenesis pathway (biocarta591) which were shown in Table 2. We can see that the correlations between the genes were quite large and cannot be ignored.

**Table 2.** Correlations among 6 genes within Biocarta591 in dataset I (upper triangular) and dataset II (lower triangular). Absolute value list as below:

| | CREBBP | EP300 | LPL | RXRA | NCOA1 | NCOA2 |
|---|---|---|---|---|---|---|
| CREBBP | 1 | 0.541728 | 0.099420 | 0. 278448 | 0.266486 | 0.225759 |
| EP300 | 0.751867 | 1 | 0.005371 | 0.349918 | 0.165928 | 0.112575 |
| LPL | 0.446183 | 0.362063 | 1 | 0.239493 | 0.063079 | 0.121320 |
| RXRA | 0.552236 | 0.560826 | 0.477218 | 1 | 0.157935 | 0.077518 |
| NCOA1 | 0.576320 | 0.618891 | 0.174420 | 0.486905 | 1 | 0.111558 |
| NCOA2 | 0.493967 | 0.658855 | 0.129232 | 0.236496 | 0.461819 | 1 |

To investigate how the correlations among genes affect the P-values, we present Table 3 to summarize P-values for testing differential expression of the pathway by combining independent P-values and dependent P-values. Independent P-values was got by replacing the correlation coefficient matrix R with identity matrix I. From Table 3, we can see that the P-value for testing pathway by combining independent P-values is much smaller than that by combining dependent P-values. This indicates that the statistic by combining independent P-values of genes may have high false positive rates. Thus we conclude that correlations have big impacts on the P-values of the statistics for testing differential expressions of the pathways and thus cannot be ignored.

**Table 3.** P-values of the pathways using linear combination of independent and dependent P-values of genes within pathways

| | independent | | dependent | |
|---|---|---|---|---|
| Pathway name | $T = \frac{e^T Z}{\sqrt{K}}$ | P-value | $T_L = \frac{e^T Z}{\sqrt{e^T \mathrm{Re}}}$ | P-value |
| Five pathways in dataset II | | | | |
| Glycolysis / Gluconeogenesis | -4.75595 | 1.98E-06 | -1.85403 | 0.063734 |
| Citrate cycle (TCA cycle) | -1.12204 | 0.261844 | -0.42385 | 0.671678 |
| Pentose phosphate pathway | -4.56911 | 4.90E-06 | -2.28524 | 0.022299 |
| Pentose and glucuronate interconversions | -1.41608 | 0.156753 | -1.05021 | 0.29362 |
| Fructose and mannose metabolism | -4.98735 | 6.12E-07 | -2.00286 | 0.045192 |
| Five significant pathways in dataset II | | | | |
| Cell Communication | -11.3023 | 0 | -5.05448 | 4.32E-07 |
| C21-Steroid hormone metabolism | -4.17168 | 3.02E-05 | -4.57937 | 4.66E-06 |
| Complement Pathway pathway | -6.78801 | 1.14E-11 | -4.39731 | 1.10E-05 |
| Complement and coagulation cascades | -10.5662 | 0 | -4.2202 | 2.44E-05 |
| Lectin Induced Complement Pathway | -4.47688 | 7.57E-06 | -4.05495 | 5.01E-05 |

### 3.2.2 Application to Rheumatoid Arthritis in Two Datasets

The proposed statistics were applied to two RA gene expression datasets. Table 4 showed the P-values of Dentatorubropallidoluysian atrophy (DRPLA) (hsa05050) and genes in the pathway. To evaluate the performance of the proposed statistics for testing differential expressions of the pathway, we also listed the P-value of the TAPPA method for comparison [3]. See supplementary Tables 3 for another pathway with more genes.

Table 5 listed pathways with both P-values < 0.01 which were obtained by novel linear combination test (LCT) in pathway-based gene expression studies of rheumatoid arthritis. We found three pathways showing significance in the two gene expression datasets. Obviously, these pathways are involved in inflammatory process. See supplementary Tables 2 for more pathways with P-values < 0.05.

**Table 4.** P-values of genes in Dentatorubropallidoluysian atrophy (hsa05050)

| Dataset I | | Dataset II | |
|---|---|---|---|
| Method | P-value | Method | P-value |
| LCT | 5.76E-04 | LCT | 0.001985 |
| QT | 1.39E-09 | QT | 1.52E-04 |
| DT | 9.27E-10 | DT | 2.57E-04 |
| TAPPA | 0.347244 | TAPPA | 0.006951 |
| Gene | P-value | Gene | P-value |
| ATN1 | 0.273753 | ATN1 | 0.002935 |
| BAIAP2 | 0.511389 | BAIAP2 | 0.024157 |
| CASP1 | 0.00022 | CASP1 | 0.147013 |
| CASP3 | 0.562305 | CASP3 | 0.60399 |
| CASP8 | 0.652234 | CASP7 | 0.439692 |
| GAPDH | 0.334194 | INS | 0.215546 |
| INSR | 0.202407 | INSR | 0.50486 |
| ITCH | 0.231124 | ITCH | 0.294672 |
| MAGI1 | 0.708848 | MAGI1 | 0.248591 |
| MAGI2 | 0.388306 | MAGI2 | 0.223494 |
| RERE | 0.000894 | RERE | 0.799462 |
| WWP1 | 0.00044 | WWP1 | 0.200264 |
| WWP2 | 0.000179 | WWP2 | 0.379637 |

**Table 5.** Three Pathways with P-values < 0.01 in rheumatoid arthritis studies obtained by novel linear combination test (LCT)

| | | Dataset I | | Dataset II | |
|---|---|---|---|---|---|
| Pathway name | Gene$^{\#1}$ | Gene$^{\#2}$ | P-value | Gene$^{\#3}$ | P-value |
| Dentatorubropallidoluysian atrophy (DRPLA) | 15 | 13 | 0.000576 | 13 | 0.001985 |
| Axon guidance | 128 | 95 | 0.000710 | 106 | 0.008328 |
| Tetrachloroethene degradation | 3 | 2 | 0.001051 | 3 | 0.000479 |

Gene$^{\#1}$, Gene$^{\#2}$, Gene$^{\#3}$ mean in the specific pathway, the total number of genes, the number of genes contained in dataset I and the number of genes contained in dataset II.

## 4 Discussion

Despite great success in microarray technology, traditional strategies for gene expression analysis have focused on identifying individual genes that exhibit differences in expressions between abnormal and normal samples. Although useful, single gene differential expression analysis will miss many genes with moderate genetic effects and fail to detect biological processes which play an important role in disease development. To overcome these limitations, several pathway-based data analysis methods have been proposed for gene expression data analysis[10-14]. However, most statistical methods for GSEA have ignored correlations among the genes in the pathway.

To investigate whether correlations exist among the genes in the pathway and if there are, whether the correlations have impact on the results of pathway differential expression analysis, we calculated correlations among genes in the pathway and test statistics without consideration of the correlations among genes using real RA gene expression datasets. We found that the correlations among the genes in the pathway were quite large and cannot be ignored when we design test statistics.

In this report, we proposed new statistics for GSEA which take correlations among the genes in the pathway into account. The newly developed statistics were applied to two RA gene expression datasets. We found three common pathways in two datasets which are significantly different between case and control samples. Our results were not very consistent with other published literatures due to the following reasons[5,7,15-17]: 1) Sample sizes in both two datasets are very small. 2) The patients with different rheumatoid arthritis subtypes may fall ill by different disease mechanisms thus we find different related pathways instead of shared common pathways. Dataset I includes 35 patients, 25 of which are polyarticular rheumatoid arthritis and 10 are pauciarticular rheumatoid arthritis. There are gene expression differences in two RA subtypes. Dataset II also contains two rheumatoid arthritis subtypes signature by IFN-induced gene. 3) Different microarray platform definitely would affect the results. One-color affymetrix DNA chip use the abstract intensity value while two-color cDNA chip get the log2 ratio of intensity value of test sample by reference sample. The different design principle in two microarray platform causes noise and variances of results. The merits of our statistic should be further validated in more trusted datasets.

Although in most cases QT and DT are much powerful than the LCT method, the QT and DT methods are not reliable due to singularity of the correlation matrix of the expressions of the genes in the pathway. In the future, we need to design a strategy to ensure that the correlation matrices of the gene expressions are positive definite.

## Acknowledgments

Hoicheong Siu, Hua Dong, Li Jin, Momiao Xiong are supported by National Science Foundation of China (30890034), Hi-Tech Research and Development Program of China(863) (2007AA02Z312) and Shanghai Commission of Science and Technology (04dz14003). Li Jin is also supported by Shanghai Leading Academic Discipline Project (B111) and the Center for Evolutionary Biology.

## References

1. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545–15550 (2005)
2. Wang, K., Li, M., Bucan, M.: Pathway-Based Approaches for Analysis of Genome-wide Association Studies. Am. J. Hum. Genet. 81, 1278–1283 (2007)
3. Gao, S., Wang, X.: TAPPA: topological analysis of pathway phenotype association. Bioinformatics 23, 3100–3102 (2007)

4. Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/
5. Barnes, M.G., Aronow, B.J., Luyrink, L.K., Moroldo, M.B., et al.: Gene expression in juvenile arthritis and spondyloarthropathy: pro-angiogenic ELR+ chemokine genes relate to course of arthritis. Rheumatology (Oxford) 43(8), 973–979 (2004)
6. Stanford MicroArray Database, http://genome-www5.stanford.edu/
7. van der Pouw Kraan, T.C., Wijbrandts, C.A., van Baarsen, L.G., Voskuyl, A.E., Rustenburg, F., Baggen, J.M., Ibrahim, S.M., Fero, M., Dijkmans, B.A., Tak, P.P., Verweij, C.L.: Rheumatoid arthritis subtypes identified by genomic profiling of peripheral blood cells: assignment of a type I interferon signature in a subpopulation of patients. Ann. Rheum. Dis. 66(8), 1008–1014 (2007)
8. KEGG (Kyoto Encyclopedia of Genes and Genomes), http://www.genome.jp/kegg
9. BioCarta, http://www.biocarta.com
10. Benfey, P.N., Mitchell-Olds, T.: From genotype to phenotype: systems biology meets natural variation. Science 320, 495–497 (2008)
11. Curtis, R.K., Oresic, M., Vidal-Puig, A.: Pathways to the analysis of microarray data. Trends Biotechnol 23, 429–435 (2005)
12. Werner, T.: Bioinformatics applications for pathway analysis of microarray data. Curr. Opin. Biotechnol. 19(1), 50–54 (2008)
13. Curtis, R.K., Oresic, M., Vidal-Puig, A.: Pathways to the analysis of microarray data. Trends Biotechnol. 23(8), 429–435 (2005)
14. Adewale, A.J., Dinu, I., Potter, J.D., Liu, Q., Yasui, Y.: Pathway analysis of microarray data via regression. J. Comput. Biol. 15(3), 269–277 (2008)
15. Olsen, N., Sokka, T., Seehorn, C.L., Kraft, B., Maas, K., Moore, J., Aune, T.M.: A gene expression signature for recent onset rheumatoid arthritis in peripheral blood mononuclear cells. Annals of the Rheumatic Diseases 63, 1387–1392 (2004)
16. Szodoray, P., Alex, P., Frank, M.B., Turner, M., Turner, S., Knowlton, N., Cadwell, C., Dozmorov, I., Tang, Y., Wilson, P.C., Jonsson, R., Centola, M.: A genome-scale assessment of peripheral blood B-cell molecular homeostasis in patients with rheumatoid arthritis. Rheumatology 45, 1466–1476 (2006)
17. Chang, M., Rowland, C.M., Garcia, V.E., Schrodi, S.J., Catanese, J.J., van der Helm-van Mil, A.H., Ardlie, K.G., Amos, C.I., Criswell, L.A., Kastner, D.L., Gregersen, P.K., Kurreeman, F.A., Toes, R.E., Huizinga, T.W., Seldin, M.F., Begovich, A.B.: A large-scale rheumatoid arthritis genetic study identifies association at chromosome 9q33.2. PLoS Genet. 27, 4(6), e1000107 (2008)