Duke University
Duke Biostatistics and Bioinformatics (B&B) Working Paper Series

Year 2011 *Paper* 15

Multiple Testing for Gene Sets from Microarray Experiments

Insuk Sohn[*] Kouros Owzar[†] John Lim[‡]
Stephen George[**] Stephanie Mackey Cushman[††] Sin-Ho Jung[‡‡]

[*]Samsung Medical Center
[†]Duke University
[‡]Seoul National University
[**]Duke University
[††]Duke University
[‡‡]Duke University, sinho.jung@duke.edu

http://biostats.bepress.com/dukebiostat/art15

Multiple Testing for Gene Sets from Microarray Experiments

Insuk Sohn, Kouros Owzar, John Lim, Stephen George, Stephanie Mackey Cushman, and Sin-Ho Jung

Abstract

Background: A key objective in many microarray association studies is the identification of individual genes associated with clinical outcome. It is often of additional interest to identify sets of genes, known a priori to have similar biologic function, associated with the outcome.

Results: In this paper, we propose a general permutation-based framework for gene set testing that controls the false discovery rate (FDR) while accounting for the dependency among the genes within and across each gene set. The application

of the proposed method is demonstrated using three public microarray data sets. The performance of our proposed method is contrasted to two other existing Gene Set Enrichment Analysis (GSEA) and Gene Set Analysis (GSA) methods.

Conclusions: Our simulations show that the proposed method controls the FDR at the desired level. Through simulations and case studies, we observe that our method performs better than GSEA and GSA, especially when the number of prognostic gene sets is large.

Multiple Testing for Gene Sets from Microarray Experiments

Insuk Sohn[1] , Kouros Owzar[2], Johan Lim[3], Stephen L. George[2], Stephanie Mackey Cushman[4]and Sin-Ho Jung[*2]

[1]Biostatistics and Bioinformatics Center, Samsung Cancer Research Institute, Samsung Medical Center, Seoul, 137-710, Republic of Korea
[2]Department of Biostatistics and Bioinformatics, Duke University Medical Center, NC 27710, USA
[3]Department of Statistics, Seoul National University, Seoul 151-747, Republic of Korea
[4]Department of Medicine, Division of Medical Oncology, Duke University, NC 27710, USA

Email: Insuk Sohn - insuk.sohn@samsung.com; Kouros Owzar - kouros.owzar@duke.edu; Johan Lim -johanlim@snu.ac.kr; Stephen L. George - stephen.george@duke.edu; Stephanie Mackey Cushman - stephanie.mackey@duke.edu; Sin-Ho Jung*- sinho.jung@duke.edu;

*Corresponding author

Abstract

Background: A key objective in many microarray association studies is the identification of individual genes associated with clinical outcome. It is often of additional interest to identify sets of genes, known a priori to have similar biologic function, associated with the outcome.

Results: In this paper, we propose a general permutation-based framework for gene set testing that controls the false discovery rate (FDR) while accounting for the dependency among the genes within and across each gene set. The application of the proposed method is demonstrated using three public microarray data sets. The performance of our proposed method is contrasted to two other existing Gene Set Enrichment Analysis (GSEA) and Gene Set Analysis (GSA) methods.

Conclusions: Our simulations show that the proposed method controls the FDR at the desired level. Through simulations and case studies, we observe that our method performs better than

GSEA and GSA, especially when the number of prognostic gene sets is large.

Background

One of the primary objectives in microarray association studies is the identification of individual genes that are associated with clinical endpoints such as disease type, toxicity or time to death. It is also of interest to examine the association between known biological categories or pathways and outcome. To this end, gene sets a priori believed to have similar biological functions from databases including KEGG [1] and Gene Ontology [2] are used. In recent years, a number of statistical methods have been proposed for the identification of significant gene-sets based on microarray experiments. Ackerman and Strimmer [3] list 36 methods, including [4–13], while outlining a general framework for formulating the hypothesis and analysis method for gene set inference.

In this paper, we propose a gene set analysis framework that utilizes classical theory for estimating equations to assess the association between each gene set and the outcome of interest. One of the statistical challenges in this setting is that there is dependency within each gene set, by virtue of co-regulated genes belonging to the same gene set, as well as dependency across the gene sets since gene sets are not mutually exclusive. Our method will account for both intra-gene set and inter-gene set dependencies. Furthermore, given the large number of gene sets, one has to address the issue of multiple testing. The sampling distribution of our proposed procedure is approximated using permutation resampling to simultaneously address the dependency and multiple testing issues by controlling the false discovery rate (FDR; [14]). In the framework described by Ackerman and Strimmer [3], gene set analysis methods are broadly categorized as univariate or as global and multivariate procedures. Generally speaking, our method belongs to the latter category. The novelty of our proposed approach is that it leverages the flexibility of estimating equations to conduct inference for a variety of endpoints including binary, continuous, censored or longitudinal outcomes.

After presenting the theoretical and computational details for the proposed method, we summarize the results from a simulation study evaluating its statistical properties. We then apply the proposed method to analyze a number of microarray data sets. Finally, we provide a brief discussion to compare the performance of our method to those of two other methods: GSEA [4] and GSA [6]. For notational brevity, we will refer to transcripts on microarrays as genes, even though this may not be technically correct.

All analyses are carried out using the `R` statistical environment [15]. The code is available from www.duke.edu/~is29/GeneSet. Generalized inverses are computed using the `pinv` function from the `maanova` [16] extension package. The inverse of linear shrinkage covariance matrix is computed using `invcov.shrink` function in `corpcor` [17] extension package. The `R` extension packages `R-GSEA` [5] and `GSA` [18] are used to implement the GSEA and GSA methods respectively. The `qvalue` [19] extension package is used for calculating FDR adjusted P-values. For gene set and probe set annotation, Bioconductor [21] annotation packages (e.g., `hu6800.db` [20]) and Molecular Signature Database (MSigD; http://www.broad.mit.edu/gsea) annotation files are used.

Methods

In these discussions, we will assume that RNA expression levels for m genes have been measured for n patients. Let us denote the set of genes on the microarray by $G = \{G_1, \ldots, G_m\}$. For patient $i(= 1, ..., n)$, let y_i denote the clinical outcome and z_{ij} denote the measured gene expression level for G_j. Let $G_j \perp Y$ denote that expression of gene j is not associated with outcome. For each gene the marginal inference of interest will be canonically presented as testing $H_j : G_j \perp Y$ versus $\bar{H}_j : G_j \not\perp Y$.

Suppose that for gene j, the hypotheses of independence can be quantified using a parameter say θ_j. We assume that $\theta_j = 0$ indicates that G_j and the outcome are independent. Thus, the hypotheses of interest can be expressed as testing $H_j : \theta_j = 0$ against $\bar{H}_j : \theta_j \neq 0$. We consider testing these marginal hypotheses within the context of general estimating functions, which for large n, are expressible in the form

$$U_j(\theta_j) = \sum_{i=1}^{n} U_{ij}(\theta_j),$$

where $U_{ij}(\theta_j)$ is a function of the data for subject i only so that $U_{1j}, ..., U_{nj}$ are independent. The corresponding test statistic for H_j will be $U_j(0)$. Let $\mu_{ij}(\theta_j) = \mathrm{E}(U_{ij})$ and $\mu_j(\theta_j) = \sum_{i=1}^{n} \mu_{ij}(\theta_j)$. If $\mathrm{E}\{U_j(\theta)\}$ is a smooth function and $\mathrm{E}\{U_j(\theta)\} = 0$ has a unique solution, then the solution $\widehat{\theta}_j$ to $U_j(\theta) = 0$ is a consistent estimator of θ_j. The family of score statistics [22] is a special case of this type of estimating equation.

A gene set is defined as a subset of G. We will assume that there are K pre-specified gene sets say $\mathcal{G}_1, \ldots, \mathcal{G}_K$ based on a given annotation database such as KEGG or GO. Note that $\mathcal{G}^* := \mathcal{G}_1 \cup \ldots \cup \mathcal{G}_K$ is usually a proper subset of G as not all genes are annotated. Let m_k $(k = 1, \ldots, K)$ denote the number of genes in gene set $\mathcal{G}_k$. We consider a gene set to be associated with the outcome of interest if at least one of its member genes is associated with the outcome. Let $\mathcal{G}_k \perp Y$ denote that gene set $\mathcal{G}_k$ is not associated with the outcome Y. The hypotheses of interest from gene set k can then be denoted as testing $\mathcal{H}_k : \mathcal{G}_k \perp Y$ versus $\bar{\mathcal{H}}_k : \mathcal{G}_k \not\perp Y$.

For notational convenience, for the remainder of this section we will focus on the first gene set $\mathcal{G}_1$ and assume that it consists of the first m_1 genes, $G_1, \ldots, G_{m_1}$. Then the hypotheses of interest can be presented as testing $\mathcal{H}_1 = \cap_{j=1}^{m_1} H_j$ against $\bar{\mathcal{H}}_1 = \cup_{j=1}^{m_1} \bar{H}_j$. For testing this hypothesis, consider the vector $\boldsymbol{U}_1 = (U_1, ..., U_{m_1})^T$, of the first m_1 marginal statistics, which is approximately normal with marginal means $\mu_j(\theta_j)$ and covariances $\sigma_{jj'}$ $(j, j' = 1, \ldots, m_1)$. These quantities can be consistently estimated by $\widehat{\mu}_j = \mu_j(\widehat{\theta}_j)$ and

$$\widehat{\sigma}_{jj'} = \sum_{i=1}^{n} (U_{ij} - \widehat{\mu}_{ij})(U_{ij'} - \widehat{\mu}_{ij'}),$$

respectively, where $\widehat{\mu}_{ij} = \mu_{ij}(\widehat{\theta}_j)$. Let $\sigma_{jj} = \sigma_j^2$ and $\widehat{\sigma}_{jj} = \widehat{\sigma}_j^2$.

In the marginal testing setting, we have $\mu_j(0) = 0$ under H_j, so we reject H_j in favor of $\bar{H}_j$ if the realized value of $U_j^2/\widehat{\sigma}_j^2$ is large. The test statistic $U_j^2/\widehat{\sigma}_j^2$ has an asymptotic χ^2 distribution with 1 degree of freedom under the null distribution. For gene set $\mathcal{G}_1$ we will consider the test statistic

$$W_1 = \boldsymbol{U}_1^T V_1^{-1} \boldsymbol{U}_1,$$

where $V_1 = (\widehat{\sigma}_{jj'})_{m_1 \times m_1}$. If n is large and $m_1 < n$, the distribution of W_1 under $\mathcal{H}_1$ is approximately χ^2 with m_1 degrees of freedom. Similarly, we can compute $\boldsymbol{U}_k, V_k$ and W_k for any gene set $\mathcal{G}_k$.

In many cases, the sample size for a microarray study may not be large enough for the null sampling distribution to be well approximated by the theoretical limiting distribution. To address this issue, we propose calculating the P-values by approximating the exact null sampling distribution using permutation resampling. Note that the permutation distribution is generated under the hypothesis $\mathcal{H}_1 \cap \ldots \cap \mathcal{H}_K$. That is, none of the K gene sets are associated with the outcome. This hypothesis is equivalent to the hypothesis $\cap_j H_j$ (i.e., none of the genes are associated with outcome). Note that the latter intersection is restricted to G^*, the set of annotated genes. A permutation replicate sample is obtained by randomly shuffling the the clinical outcomes $\{y_1, ..., y_n\}$ while holding the gene expression matrix in place. This ensures that the intra-gene dependency structure is preserved while breaking the association between the genes and the outcome.

If m_k is large, V_k many not be reliably inverted numerically and, in the case where it exceeds n, is not invertible. For these cases, we consider the Moore-Penrose (MP) generalized inverse or the inverse of the linear shrinkage covariance matrix estimate V_{LW} [13, 44–46]. Here, we remark that the Hotelling's tests with the MP generalized inverse (g-inverse) and that with the inverse of V_{LW} have been previously studied [13, 44]. The MP g-inverse (of the sample covariance matrix) uses $V_k^- = P_k D_k^{-1} P_k^T$ to derive a test statistic $W_k = \boldsymbol{U}_k^T V_k^- \boldsymbol{U}_k$, where P_k is the eigen matrix and $D_k^{-1} = \text{diag}(1/\nu_1, ..., 1/\nu_d, 0, ..., 0)$, $\nu_1, ..., \nu_d$ are the d positive eigenvalues of V_k. The asymptotic distribution of W_k when m_k is larger than n has been investigated extensively (e.g., [23–25]). The linear shrinkage estimate (LW) of V is $V_{\text{LW}} = \lambda V + (1 - \lambda)E$, where E is a well conditioned target matrix and λ is the tuning parameter. The tuning parameter λ is chosen to minimize the Frobenius risk along with several candidates of target matrices [45, 46].

Two-Sample Tests

Suppose that there are two groups with n_g subjects in group $g(= 1, 2)$, $n = n_1 + n_2$. Let $\boldsymbol{z}_{gi} = (z_{gi1}, ..., z_{gim})^T$ denote the gene expression measurements from subject $i(= 1, ..., n_g)$ in group $g(= 1, 2)$, and $\bar{\boldsymbol{z}}_k = n_g^{-1} \sum_{i=1}^n \boldsymbol{z}_{gi}$ the vector of sample means. Kong et al. [10] consider the Hotelling's T^2 statistic

$$T^2 = \frac{n_1 n_2}{n} (\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)^T S^{-1} (\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2),$$

where $S = (n-2)^{-1}\sum_{g=1}^{2}\sum_{i=1}^{n_g}(\boldsymbol{z}_{gi} - \bar{\boldsymbol{z}}_g)(\boldsymbol{z}_{gi} - \bar{\boldsymbol{z}}_g)^T$ is the pooled variance-covariance matrix. For $\theta_j = \mathrm{E}(z_{1ij}) - \mathrm{E}(z_{2ij})$ and $\mu_{ij}(\theta_j) = \theta_j$, T^2 asymptotically has a χ^2_m distribution under H_0.

For $U = \bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2$, our method gives

$$V = \frac{\widetilde{S}_1}{n_1} + \frac{\widetilde{S}_2}{n_2},$$

where $\widetilde{S}_k = n_g^{-1}\sum_{i=1}^{n_g}(\boldsymbol{z}_{gi} - \bar{\boldsymbol{z}})(\boldsymbol{z}_{gi} - \bar{\boldsymbol{z}})^T$ and $\bar{\boldsymbol{z}} = n^{-1}\sum_{g=1}^{2}\sum_{i=1}^{n_g}\boldsymbol{z}_{gi}$. Since T^2 is asymptotically equivalent to $W = \boldsymbol{U}^T V^{-1}\boldsymbol{U}$ under H_0, we use the more popular Hotelling's T^2 statistic in this paper.

As a rank test alternative to the t-test, it is easy to show that the Wilcoxon rank sum test can be expressed as T^2 with z_{gij} the rank of the gene j expression level for subject i in the pooled data $\{z_{gij}, 1 \le i \le n_g, g = 1, 2\}$. In this case, $\theta_j = \mathrm{P}(z_{1ij} \ge z_{2ij}) - 1/2$ and $\mu_{ij}(\theta_j) = \theta_j$.

Linear Regression Case

Suppose that we want to relate the gene expression z_{ij} for gene j with a continuous outcome y_i through a linear regression

$$\mathrm{E}(y_i) = a_j + \theta_j z_{ij}.$$

No association between y and the expression of gene j implies that $\theta_j = 0$. In this case, we use $U_j = \widehat{\theta}_j$, the least square estimator of the slope θ_j,

$$U_{ij} = \frac{(z_{ij} - \bar{z}_j)y_i}{\sum_{i'=1}^{n}(z_{i'j} - \bar{z}_j)^2},$$

and $\widehat{\mu}_{ij} = (z_{ij} - \bar{z}_j)(\widehat{a}_j + \theta_j z_{ij})/\sum_{i'=1}^{n}(z_{i'j} - \bar{z}_j)^2$, where $\bar{z}_j = n^{-1}\sum_{i=1}^{n} z_{ij}$, $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ and $\widehat{a}_j = \bar{y} - \widehat{\theta}_j\bar{z}_j$.

Cox Regression Case

For right-censored time to event data, the outcome data are pairs of the form $y_i = (t_i, \delta_i)$, where t_i is the minimum of survival and censoring times, and δ_i is the event indicator. Let $\lambda_i(t)$ denote the hazard function of patient i. Then the Cox proportional hazards model

relates the expression of gene j, z_{ij}, with the survival time of patient i using the model $\lambda_{ij}(t) = \lambda_{0j}(t)\exp(\theta_j z_{ij})$, where $\lambda_{0j}(t)$ is an unknown baseline hazard function. We propose using the partial score statistic [26] $U_j = U_j(0)$, where

$$U_j(\theta_j) = \sum_{i=1}^{n} \int_0^{\infty} \Big(z_{ij} - \frac{\sum_{i'=1}^{n} z_{i'j} Y_{i'}(t) e^{\theta_j z_{i'j}}}{\sum_{i'=1}^{n} Y_{i'}(t) e^{\theta_j z_{i'j}}} \Big) dN_i(t),$$

$Y_i(t) = I(t_i \geq t)$, and $N_i(t) = \delta_i I(t_i \leq t)$. Let $\widehat{\theta}_j$ denote the partial MLE of θ_j solving the partial score equation $U_j(\theta) = 0$. In this case, we have

$$\widehat{\mu}_{ij} = \int_0^{\infty} \Big(z_{ij} - \frac{\sum_{i'=1}^{n} z_{i'j} Y_{i'}(t)}{\sum_{i'=1}^{n} Y_{i'}(t)} \Big) Y_i(t) e^{\widehat{\theta}_j z_{ij}} d\widehat{\Lambda}_{0j}(t),$$

where

$$\widehat{\Lambda}_{0j}(t) = \int_0^{t} \frac{\sum_{i=1}^{n} dN_i(s)}{\sum_{i'=1}^{n} Y_{i'}(s) e^{\widehat{\theta}_j z_{i'j}}}.$$

The resulting variance estimator is equivalent to the robust estimator under the possible violation of the proportional hazards model proposed by Lin and Wei [27].

Results

Simulation Study

We investigate the performance of our proposed method with respect to FDR control through a simulation study. Let z_{ijk} denote the expression level of gene $j(= 1, \ldots, m_k)$ from subject i $(= 1, \ldots, n_1 + n_2)$ in the group $g(= 1, 2)$ for gene set $\mathcal{G}_k$. We consider the following model :

$$z_{ijk} = \begin{cases} s_i \delta_j + \sqrt{\rho_1}\, a_k + \sqrt{\rho_2}\, b + \sqrt{1-\rho_1-\rho_2}\, \epsilon_{ijk} & \text{for } 1 \leq j \leq D, 1 \leq k \leq K_1 \\ \sqrt{\rho_1}\, a_k + \sqrt{\rho_2}\, b + \sqrt{1-\rho_1-\rho_2}\, \epsilon_{ijk} & \text{for } D+1 \leq j \leq m_k, 1 \leq k \leq K_1 \\ \sqrt{\rho_1}\, a_k + \sqrt{\rho_2}\, b + \sqrt{1-\rho_1-\rho_2}\, \epsilon_{ijk} & \text{for } 1 \leq j \leq m_k, K_1+1 \leq k \leq K, \end{cases}$$

where $s_i = 0$ if subject i belongs to group 1 and $s_i = 1$ otherwise. Here, for gene j, δ_j is the treatment effect, D is the number of prognostic genes, K_1 is the number of prognostic gene sets, a_k is the gene set effect, b is the array effect, (ρ_1, ρ_2) are the correlation coefficients, and ϵ_{ijk} is the error term. The gene set effect a_k, the array effect b, and the error term ϵ_{ijk} are generated from independently and identically distributed $N(0, 1)$ random variate.

At first, we investigate the performance of the test statistic using the MP inverse generalized inverse. We consider $m = 1,000$ genes and $n = 100$ samples, each with non-overlapping $K = 50$ or 20 gene sets of $m_k = 20$ or 50 genes, respectively, $(\rho_1,\ \rho_2) = (0,\ 0)$, (0.2, 0.2) or (0.4, 0.4), $D/m_k = 0.2, 0.5$ or 0.8, $\delta = 0.4$ and $K_1 = 1$ or 5. We conduct $N = 1,000$ simulations under each setting, and approximate the null distribution of the test statistic using $B = 10,000$ random permutations for each simulation. The q-values [28] are obtained from the resulting unadjusted permutation P-values by setting $\lambda = 0.5$. Results are presented in Table 1 where $\widehat{q}$ denotes the empirical FDR and $\widehat{r}_1$ denotes the mean number of true rejections, i.e. the mean number of prognostic gene sets that are discovered by testing. These results illustrate that the proposed method accurately controls the FDR at the desired level q^*. The observed true rejection rate is high when the proportion of prognostic genes within each gene set is large (i.e., $D/m_k = 0.8$).

We proceed by investigating the case with small n but large m_k. We set the sample size $n = 20$, and consider $K = 20$ and $m_k = 50$. All other parameters are identical to those used in the simulation study reported in Table 1. We conduct $N = 500$ simulations and apply the test using both MP and LW generalized inverses. The results reported in Table 2 show that both tests control the FDR at the desired level q^*. Similar to the results presented in Table 1, for both tests the observed true-rejection rate $(\widehat{r}_1)$ increases in the proportion of prognostic genes within each gene set (D/m_k). However, the test with the LW inverse has generally higher true-rejection rate than that with the MP generalized inverse.

We compare the performance of our method to GSEA and GSA within the simulation framework described above. We choose, for GSEA, the weighted Kolmogorov Smirnov-like statistic as enrichment correlation-based weighting, while for GSA we choose the maxmean statistic along with restandardization. The technical details are provided in [5]) and in [6] respectively.

We generate $m = 1,000$ genes and $n = 100$ samples, each with non-overlapping $K = 50$ gene sets of $m_k = 20$ genes, $(\rho_1,\ \rho_2) = (0,\ 0)$, D/m_k=1, and $\delta = 0.4$ as in [6]. The first ($n_1 = 50$) and second ($n_2 = 50$) samples will constitute the control and treatment groups respectively. Next, we will discuss two scenarios similar to those considered by [6]:

- One-sided shifts: The mean expression level for the $m_k = 20$ genes in each of the K_1

prognostic gene sets is $\delta = 0.4$ units higher in the treatment group.

- Two-sided shifts: The mean expression level for the first 10 genes in each of the K_1 prognostic gene sets is $\delta = 0.4$ units higher, while the mean expression level for the next 10 genes is $\delta = 0.4$ units lower.

Each scenario is simulated 100 times using 1000 permutation replicates. The P-values for the first gene set is shown against the number of prognostic gene sets in Figure 1. Overall, our method gives lower mean P-values under both scenarios. In the one-sided shift case, the three methods are comparable when the number of prognostic gene sets is at most thirteen. For the cases with a large number of prognostic gene sets or a two-sided shift, our method is consistently better.

Case Studies

Two-Sample Case

We analyze two microarray data sets available from the GSEA website (www.broad.mit.edu/gsea). The first data set, called the Gender data set, consists of profiles of $m = 15,056$ genes from male ($n_1 = 15$) and female ($n_2 = 17$) lymphoblastoid cell lines. The second data set consists of transcriptional profiles of $m = 10,100$ genes from p53 positive ($n_1 = 17$) and p53 mutant ($n_2 = 33$) cancer cell lines. The pathways from MSigDB are currently organized into five catalogs. We use the Positional gene sets (C_1; 319 gene sets), which correspond to each human chromosome and each cytogentic band, for the Gender data set and the Curated gene sets (C_2; 522 gene sets), which are derived from online pathway databases and publications, for the p53 data set. For the analyses, we limit our attention to gene sets which consist of a minimum of 15 and a maximum of 500 genes. Each analysis is based on 10,000 permutation replicates. The performance of our method is compared to those of GSEA and GSA.

For comparison of the three methods, we compare the number of prognostic gene sets identified by each of the three methods. The analysis results for the two data sets are shown in Figure 2. For both data sets, our method consistently identifies more prognostic gene sets than GSEA and GSA for any q-value threshold. For the Gender data set, at the FDR level of $q^* = 0.2$, our method identifies 8 gene sets compared to only 4 for the other

two methods [see Additional file 1]. There are 4 prognostic gene sets identified in common among the three methods, consisting of gene sets found on ChrY, ChrYp11, ChrYq11, and ChrXp22. Our method identifies 4 other gene sets not identified by the other two methods, which include gene sets for ChrX, ChrXp11, Chr3q25, and Chr6q25. Genes expressed on the Y chromosome are expected to be differentially expressed between genders, while gene expression from the X chromosome is more similar between genders due to X chromosome inactivation in females [29, 30]. However, ChrXp22 and ChrXp11 gene sets have been previously been shown to be overrepresented in females likely caused by escape of X inactivation [31]. Furthermore, several genes within the Chr3q25 and Chr6q25 gene sets have also been shown to be differentially expressed between males and females, including ACAT2 [32], MAP3K4 [33], NOX, PTX3 [34], SGEF, and SOD2 [35]. Thus, our method for identifying overrepresented genes in gene set lists can provide biologically relevant and important information that may be overlooked by other common methods such as GSA and GSEA.

For the p53 data set, at the same FDR level, our method identifies 87 prognostic gene sets while GSA and GSEA identify 5 and 9 prognostic gene sets, respectively [see Additional file 1]. There are 5 prognostic gene sets common among the three methods, including the p53 pathway, hsp27 pathway, radiation sensitivity pathway, ceramide pathway, and the ras pathway. However, our method identifies 78 gene sets not identified by the other two methods. The supplementary material provides a list of gene sets that are identified only by our method [see Additional file 2]. p53 is a tumor suppressor protein that is activated in response to DNA damage. p53 can induce growth arrest by halting the cell cycle at the G1/S phase transition to allow DNA repair or it can induce apoptosis if the DNA damage cannot be repaired. p53 acts as a transcription factor regulating the expression of many genes involved in its functions [36]. Thus many of the gene sets identified by our method can be directly linked to p53 functions, such as cell cycle arrest, ATM pathway, tumor suppressor, bcl2 family and network, death pathway, etc [36]. Additionally, several cytokine and growth factor signaling pathways are represented in our list of gene sets differentially expressed between p53 positive and mutant cell lines, including the IL-4 [37], EGF [38], NGF [39], CXCR4 [40], IL-7 [41], and PDGF [42] pathways, which have all

shown roles for p53 in their regulation and signaling. The method that we describe here for identifying prognostic gene sets can provide a more inclusive list of gene sets that provide further insight into the biology of two sample case studies from microarray experiments.

Cox Regression Case

We carry out gene set analysis of the lung cancer microarray data set [43] using the KEGG pathway (175 gene sets) provided by the `hu6800.db` Bioconductor package. The data set consists of gene expressions of m = 4, 966 genes from n = 86 stage I or III lung cancer patients. As in the analyses for the previous data sets, we include gene sets consisting of 15 to 500 genes each in the analysis, and use 10,000 permutations to derive the null distribution of the test statistics. For this analysis, we will compare our method to GSA only since the R-GSEA extension package does not provide the functionality for analyzing right censored data. The results are shown in Figure 3 suggest that our method generally identifies a larger number of prognostic gene sets compared to GSA.

Conclusion

In this paper, we have presented a multiple testing procedure to identify prognostic gene sets from a microarray experiment correlated with common types of binary, continuous and time to event clinical outcomes. We calculate the marginal P-values using a permutation method accounting for dependency among the genes within and across each gene set, and account for multiple testing by controlling the FDR. Our simulations show that our proposed method controls the FDR at the desired level. Through extensive simulations and real case studies, we observe that our method performs better than GSEA and GSA, especially when the number of prognostic gene sets is large.

Authors contributions

IS and KO performed statistical analysis and wrote the manuscript. JL supported technical aspects of the research. SC provided biological interpretation of the gene sets found to be significant by the proposed method. SHJ and SLG proposed the research project. SHJ developed the methodological framework. All authors read and approved the

final manuscript.

Acknowledgements

Partial support for this research was provided by a grant from the National Cancer Institute (CA142538).

References

1. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27**:29-34.

2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. : **Gene Ontology: tool for the unification of biology.** *Nat. Genet* 2000, **25**:25-29.

3. Ackermann M, Strimmer K.: **A general modular framework for gene set enrichment analysis.** *BMC Bioinformatics.* 2009, **88**:365-411.

4. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** *Nat. Genet.* 2003, **34**:267-273.

5. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.** *Proc. Natl. Acad. Sci.* 2005, **102**:15545-15550.

6. Efron B and Tibshirani R: **On testing the significance of sets of genes.** *Ann. Appl. Stat.* 2007, **1**:107-129.

7. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome.** *Bioinformatics* 2004, **20**:93-99.

8. Mansmann U, Meister R: **Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach.** *Methods of Inf. Med.* 2005, **44**:449-453.

9. Barry WT, Nobel AB, Wright F: **Significance analysis of functional categories in gene expression studies: a structured permutation approach.** *Bioinformatics* 2005, **21(9)**:1943-9.

10. Kong SW, Pu WT, Park PJ: **A multivariate approach for integrating genome-wide expression data and biological knowledge.** *Bioinformatics* 2006, **22**:2373-2380.

11. Nettleton D, Recknor J, Reecy JM: **Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis.** *Bioinformatics* 2008, **24**:192-201.

12. Barry WT, Nobel AB, Wright F: **A statistical framework for testing functional categories in microarray data.** *Annals of Applied Statistics* 2008, **2(1)**:286-315.

13. Tsai C-A, Chen, JJ.: **Multivariate analysis of variance test for gene set analysis.** *Bioinformatics.* 2009, **25(7)**:897-903.

14. Benjamini Y and Hochber Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *JR Statist Soc B* 1995, **57**:289-300.

15. Development Core Team: `R`: A Language and Environment for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. `http://www.R-project.org`. 2009.

16. Wu H, modified by Yang, H, Sheppard,K with ideas from Churchill G, Kerr K, Cui, X.: **maanova: Tools for analyzing Micro Array experiments**, 2010. R package version 1.20.0.

17. Schaefer J, Opgen-Rhein R, Strimmer, K.: **corpcor: Efficient Estimation of Covariance and (Partial) Correlation**, 2010. R package version 1.5.7.

18. Efron B, Tibshirani, R.: **GSA: Gene set analysis**, 2010. R package version 1.03.

19. Dabney A, Storey JD with assistance from Warnes GR.: **qvalue: Q-value estimation for false discovery rate control**, 2010. R package version 1.24.0.

20. Carlson M, Falcon S, Pages H, Li N.: **hu6800.db: Affymetrix HuGeneFL Genome Array annotation data (chip hu6800)**, 2010. R package version 2.4.5.

21. Gentleman R, Carey V, Bates D, et al.: **Bioconductor: Open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **5(10)**:R80.

22. de Boor C: *Theoretical Statistics* hapman and Hall. London. 1974.

23. Box,G.E.P.: **Some theorems on quadratic forms applied in the study of analysis of variance problems, I: Effect of inequality of variance in the one-way classification.** *The Annals of Mathematical Statistics* 1995, **25**:290-302.

24. Brunner E: **Asymptotic and approximate analysis of repeated measuresd esigns under heteroscedasticity. In: mathematical statistics with applications in biometrys** Eds.: Kunert, J. and G. Trenkler. Josef Eul Verlag, Lohmar. 2001.

25. Chen SX and Qin YL: **A two sample test for high dimensional data with application to gene-set testing.** *The Annals of Statistics*, to appear.

26. Cox DR: **Regression models and life-tables (with discussion.** *J. R. Stat. Soc. Ser. B* 1972, **34**:187-220.

27. Lin DY and Wei LJ: **The robust inference for the Cox proportinal havards model.** *J. Am. Stat. Assoc.* 1989, **84**:1074-1078.

28. Storey JD: **A direct approach to false discovery rates.** *JR Statist Soc B* 2002, **64(1)**, 479-498.

29. Barakata TS, Jonkers I, Monkhorst K and Gribnau J: **X-changing information on X inactivation.** *Exp Cell Res.* 2010, **316(5)**, 679-687.

30. Prothero KE, Stahl JM, and Carrel L: **Dosage compensation and gene expression on the mammalian X chromosome: one plus one does not always equal two.** *Chromosome Res.* 2009, **17(5)**, 637-648..

31. Zhang W, et al.: **Gene set enrichment analyses revealed differences in gene expression patterns between males and females.** *In Silico Biol* 2009, **9(3)**, 55-63.

32. Parini P, et al.: **ACAT2 and human hepatic cholesterol metabolism: identification of important gender-related differences in normolipidemic, non-obese Chinese patients.** *Atherosclerosis* 2009, **207(1)**, 266-271.

33. Bogani D, et al.: **Loss of mitogen-activated protein kinase kinase kinase 4 (MAP3K4) reveals a requirement for MAPK signalling in mouse sex determination.** *PLoS Biol.* 2009, **7(9)**, e1000196.

34. Yamasaki K, et al.: **Determination of physiological plasma pentraxin 3 (PTX3) levels in healthy populations.** *Clin Chem Lab Med.* 2009, **47(4)**, 471-477.

35. Khymenets O, et al.: **Role of sex and time of blood sampling in SOD1 and SOD2 expression variability.** *Clin Biochem.* 2008, **41(16-17)**, 1348-1354.

36. Tomasini R, Mark TW and Melino G: **The impact of p53 and p73 on aneuploidy and cancer.** *Trends Cell Biol* 2008, **18(5)**, 244-252.

37. Pesch J, et al.: **Repression of interleukin-2 and interleukin-4 promoters by tumor suppressor protein p53.** *J Interferon Cytokine Res.* 1996, **16(8)**, 595-600.

38. Sheikh MS, et al.: **Identification of an additional p53-responsive site in the human epidermal growth factor receptor gene promotor.** *Oncogene* 1997, **15(9)**, 1095-1101.

39. Brynczka C, Labhart P and Merrick BA: **NGF-mediated transcriptional targets of p53 in PC12 neuronal differentiation.** *BMC Genomics* 2007, **8**, 139.

40. Mehta SA, et al.: **Negative regulation of chemokine receptor CXCR4 by tumor suppressor p53 in breast cancer cells: implications of p53 mutation or isoform expression on breast cancer cell invasion.** *Oncogene* 2007, **26(23)**, 3329-3337.

41. Costello, PS, et al.: **The GTPase rho controls a p53-dependent survival checkpoint during thymopoiesis.** *J Exp Med* 2000, **192(1)**, 77-85.

42. Yang W, et al.: **Kinetics of repression by modified p53 on the PDGF beta-receptor promoter.** *Int J Cancer* 2008, **123(9)**, 2020-2030.

43. Beer DG, Kardia SL, Huang CC, et al.: **Gene-expression profiles predict survival of patients with lung adenocarcinoma.** *Nat Med.* 2002, **25**:25-29.

44. Warton DI .: **Penalized normal likelihood and ridge regularization of correlation and covariance matrices.** *Journal of the American Statistical Association* 2009, **103**:340-349.

45. Schäfer J, Strimmer, K.: **A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.** *Statistical Applications in Genetics and Molecular Biology.* 2005, **4(1)**:32.

46. Ledoit O, Wolf, M.: **A Well-conditioned estimator for large-dimensional covariance matrices.** *Journal of Multivariate Analysis.* 2004, **88**:365-411.

Figures

Figure 1 - Mean P-value against the number of prognostic gene sets

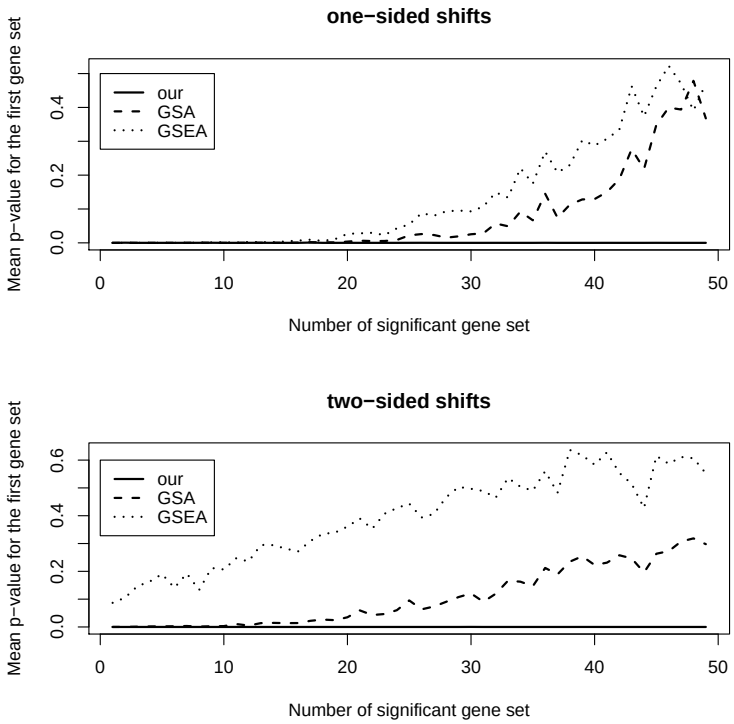


Figure 2 -The number of prognostic gene sets, at a given q-value threshold, identified by all three methods are shown for the Gender and p53 data sets.

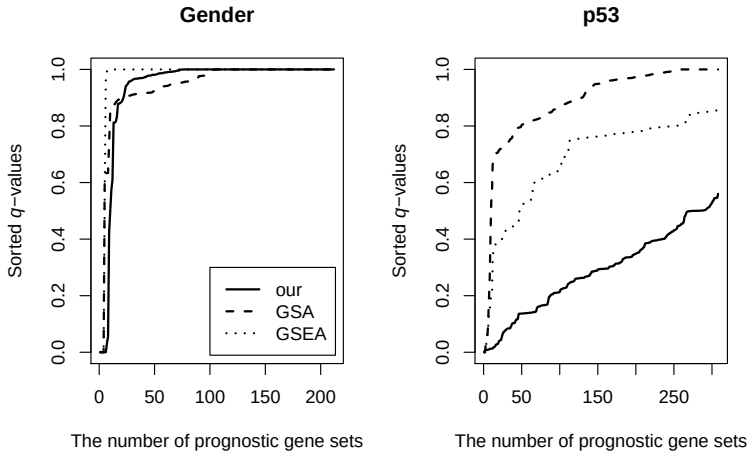
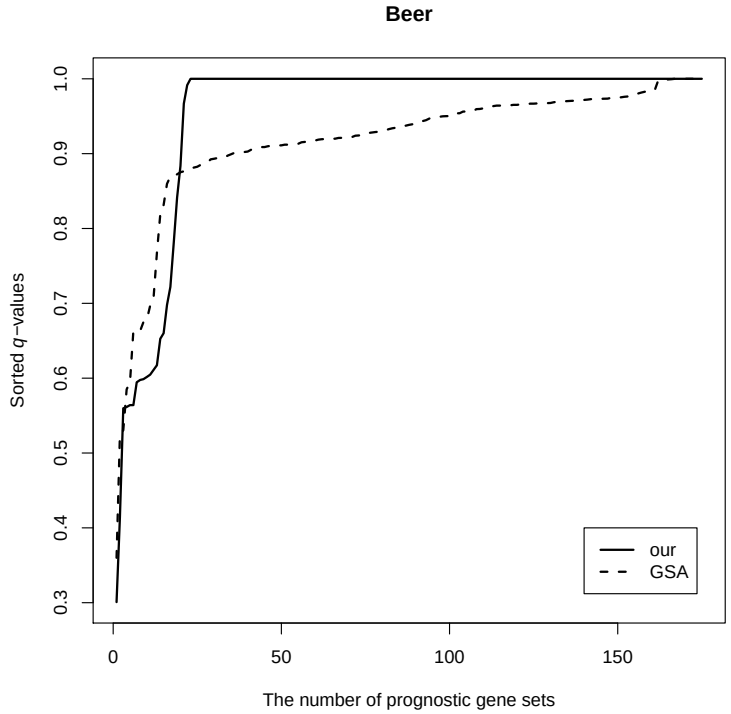


Figure 3 -The number of prognostic gene sets, at a given q-value threshold, identified by our and the GSA method are shown for the Beer Lung Cancer data set.



Tables

Table 1 - Empirical FDR and mean true rejections

				$q^*=0.01$		$q^*=0.05$		$q^*=0.1$		$q^*=0.2$	
(m_k, K)	K_1	D/m_k	(ρ_1, ρ_2)	$\widehat{q}$	$\widehat{r}_1$	$\widehat{q}$	$\widehat{r}_1$	$\widehat{q}$	$\widehat{r}_1$	$\widehat{q}$	$\widehat{r}_1$
(50,20)	1	0.2	(0, 0)	0.010	0.10	0.063	0.25	0.105	0.35	0.189	0.46
			(0.2, 0.2)	0.015	0.08	0.053	0.28	0.128	0.38	0.231	0.45
			(0.4, 0.4)	0.013	0.12	0.045	0.25	0.087	0.36	0.215	0.47
		0.5	(0, 0)	0.011	0.72	0.060	0.88	0.107	0.95	0.220	0.98
			(0.2, 0.2)	0.015	0.71	0.051	0.89	0.122	0.94	0.212	0.97
			(0.4, 0.4)	0.005	0.74	0.071	0.89	0.115	0.94	0.266	0.95
		0.8	(0, 0)	0.013	0.97	0.057	1.00	0.106	1.00	0.215	1.00
			(0.2, 0.2)	0.013	0.98	0.065	1.00	0.138	1.00	0.235	1.00
			(0.4, 0.4)	0.018	0.97	0.078	1.00	0.124	1.00	0.253	1.00
	5	0.2	(0, 0)	0.025	0.72	0.067	1.70	0.136	2.51	0.239	3.56
			(0.2, 0.2)	0.009	0.68	0.056	1.79	0.127	2.63	0.235	3.59
			(0.4, 0.4)	0.011	0.74	0.058	1.82	0.118	2.46	0.232	3.55
		0.5	(0, 0)	0.007	4.44	0.058	4.94	0.124	4.97	0.227	5.00
			(0.2, 0.2)	0.013	4.46	0.056	4.90	0.118	4.97	0.214	4.99
			(0.4, 0.4)	0.012	4.44	0.074	4.87	0.137	4.96	0.239	5.00
		0.8	(0, 0)	0.012	4.99	0.066	5.00	0.120	5.00	0.235	5.00
			(0.2, 0.2)	0.011	5.00	0.056	5.00	0.105	5.00	0.200	5.00
			(0.4, 0.4)	0.011	4.98	0.053	5.00	0.101	5.00	0.218	5.00
(20,50)	1	0.2	(0, 0)	0.008	0.04	0.055	0.13	0.092	0.21	0.185	0.25
			(0.2, 0.2)	0.015	0.04	0.051	0.10	0.128	0.15	0.219	0.21
			(0.4, 0.4)	0.005	0.06	0.050	0.10	0.110	0.14	0.198	0.22
		0.5	(0, 0)	0.013	0.45	0.048	0.63	0.118	0.72	0.240	0.82
			(0.2, 0.2)	0.010	0.43	0.077	0.64	0.122	0.72	0.223	0.80
			(0.4, 0.4)	0.010	0.49	0.072	0.68	0.112	0.76	0.214	0.83
		0.8	(0, 0)	0.010	0.86	0.043	0.97	0.113	0.99	0.222	0.99
			(0.2, 0.2)	0.015	0.89	0.048	0.98	0.115	0.98	0.208	0.98
			(0.4, 0.4)	0.013	0.86	0.052	0.96	0.102	0.99	0.201	1.00
	5	0.2	(0, 0)	0.013	0.31	0.054	0.76	0.121	1.08	0.210	1.64
			(0.2, 0.2)	0.010	0.28	0.039	0.57	0.102	0.89	0.224	1.56
			(0.4, 0.4)	0.015	0.18	0.062	0.57	0.103	0.94	0.195	1.59
		0.5	(0, 0)	0.011	3.03	0.055	4.03	0.107	4.43	0.201	4.75
			(0.2, 0.2)	0.008	3.01	0.054	4.11	0.104	4.44	0.218	4.77
			(0.4, 0.4)	0.016	3.22	0.058	4.16	0.103	4.50	0.203	4.72
		0.8	(0, 0)	0.010	4.74	0.054	4.91	0.112	4.95	0.224	4.99
			(0.2, 0.2)	0.011	4.76	0.054	4.94	0.111	4.97	0.212	4.98
			(0.4, 0.4)	0.012	4.73	0.054	4.93	0.110	4.96	0.201	4.99

Table 2 - Empirical FDR and mean true rejections on simulation data with small n large p values. Here, $n = 20$, $m_k = 50$, and $K = 20$.

				$q^* = 0.01$		$q^* = 0.05$		$q^* = 0.1$		$q^* = 0.2$	
K_1	D/m_k	(ρ_1, ρ_2)	method	$\widehat{q}$	$\widehat{r}_1$	$\widehat{q}$	$\widehat{r}_1$	$\widehat{q}$	$\widehat{r}_1$	$\widehat{q}$	$\widehat{r}_1$
1	0.2	(0, 0)	MP	0.0160	0.004	0.0633	0.018	0.1102	0.036	0.2328	0.058
			LW	0.0060	0.018	0.0523	0.040	0.1022	0.060	0.2294	0.098
		(0.2 0.2)	MP	0.0100	0.006	0.0593	0.010	0.1145	0.022	0.2325	0.042
			LW	0.0180	0.006	0.0563	0.020	0.1058	0.036	0.2249	0.084
		(0.4 0.4)	MP	0.0180	0.008	0.0590	0.020	0.1133	0.032	0.2182	0.054
			LW	0.0100	0.010	0.0503	0.032	0.1087	0.048	0.2387	0.094
	0.5	(0, 0)	MP	0.0120	0.024	0.0650	0.068	0.1138	0.100	0.2016	0.162
			LW	0.0210	0.078	0.0652	0.160	0.1189	0.230	0.2303	0.330
		(0.2 0.2)	MP	0.0300	0.024	0.0720	0.072	0.1269	0.106	0.2608	0.170
			LW	0.0200	0.084	0.0823	0.154	0.1467	0.222	0.2658	0.316
		(0.4 0.4)	MP	0.0290	0.038	0.0920	0.072	0.1359	0.112	0.2161	0.184
			LW	0.0150	0.070	0.0747	0.168	0.1300	0.228	0.2444	0.338
	0.8	(0, 0)	MP	0.0150	0.084	0.0667	0.136	0.1261	0.178	0.2534	0.258
			LW	0.0070	0.240	0.0520	0.404	0.1199	0.508	0.2179	0.618
		(0.2 0.2)	MP	0.0260	0.068	0.0563	0.142	0.1147	0.198	0.2341	0.276
			LW	0.0220	0.234	0.0603	0.420	0.1054	0.516	0.2095	0.622
		(0.4 0.4)	MP	0.0153	0.068	0.0738	0.138	0.1299	0.176	0.2548	0.258
			LW	0.0270	0.274	0.0607	0.434	0.1129	0.512	0.2118	0.636
5	0.2	(0, 0)	MP	0.0080	0.022	0.0503	0.094	0.0924	0.194	0.1817	0.386
			LW	0.0120	0.066	0.0443	0.188	0.0933	0.344	0.1883	0.726
		(0.2 0.2)	MP	0.0180	0.024	0.0683	0.084	0.1030	0.138	0.2140	0.364
			LW	0.0170	0.054	0.0642	0.144	0.1028	0.288	0.2125	0.646
		(0.4 0.4)	MP	0.0060	0.040	0.0350	0.078	0.0848	0.158	0.1969	0.304
			LW	0.0150	0.072	0.0446	0.192	0.0762	0.328	0.1761	0.668
	0.5	(0, 0)	MP	0.0143	0.158	0.0479	0.402	0.0990	0.608	0.2144	1.062
			LW	0.0140	0.442	0.0626	1.148	0.1221	1.734	0.2131	2.586
		(0.2 0.2)	MP	0.0157	0.148	0.0586	0.418	0.0974	0.662	0.2090	1.128
			LW	0.0127	0.452	0.0646	1.162	0.1076	1.796	0.2165	2.644
		(0.4 0.4)	MP	0.0160	0.150	0.0602	0.418	0.0978	0.678	0.2062	1.142
			LW	0.0108	0.498	0.0572	1.154	0.1134	1.680	0.2226	2.548
	0.8	(0, 0)	MP	0.0193	0.468	0.0662	1.006	0.1194	1.452	0.2320	2.054
			LW	0.0216	1.660	0.0609	3.044	0.1220	3.702	0.2385	4.286
		(0.2 0.2)	MP	0.0127	0.478	0.0567	1.036	0.1130	1.440	0.2257	2.156
			LW	0.0145	1.654	0.0506	2.970	0.1157	3.670	0.2269	4.278
		(0.4 0.4)	MP	0.0200	0.510	0.0587	1.020	0.1080	1.468	0.2163	2.164
			LW	0.0201	1.772	0.0729	3.066	0.1303	3.662	0.2350	4.234

Additional Files

Additional file 1

File format: pdf

Title : Results from gene set analyses for the Gender and p53 data sets

Description: This file contains two tables. Tables 1 and 2 summarize gene set analysis results based on three methods for the Gender and p53 data sets respectively.