CRRAO Advanced Institute of Mathematics,
Statistics and Computer Science (AIMSCS)

Research Report

Author (s): **Xiaoming Wang, Saumyadipta Pyne,
Irina Dinu**

Title of the Report: **Gene Set Enrichment Analysis for
Multiple Continuous Phenotypes**

Research Report No.: **RR2014-05**

Date: **February 28, 2014**

Prof. C R Rao Road, University of Hyderabad Campus,
Gachibowli, Hyderabad-500046, INDIA.
www.crraoaimscs.org

Gene Set Enrichment Analysis for Multiple Continuous Phenotypes

Xiaoming Wang[1,2], Saumyadipta Pyne[3], Irina Dinu[2]

[1] Department of Medicine, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1
[2] School of Public Health, University of Alberta, Edmonton, Alberta, Canada, T6G 1C9
[3] CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad, India; Public Health Foundation of India, Delhi, India

ABSTRACT

Motivation: Gene set analysis (GSA) methods test the association of sets of genes with phenotypes in gene expression microarray studies. While GSA methods on a single binary or categorical phenotype abound, little attention has been paid to the case of a continuous phenotype, and there are no methods to accommodate multiple and correlated continuous phenotypes.

Results: We propose here an extension of the Linear Combination Test (LCT) to multiple continuous phenotypes, incorporating correlations among gene expressions of functionally related gene sets, as well as correlations among multiple phenotypes. Further, we extend our new method to its non-parametric version, referred as NLCT, to test potential non-linear associations between gene sets and multiple phenotypes, especially recommended for relatively larger microarray studies. Simulation studies and a real microarray example demonstrate the practical aspects of the proposed LCT and NLCT methods for multiple continuous phenotypes.

Availability: Free R-codes to perform LCT for binary and multiple continuous phenotypes are available at http://www.ualberta.ca/~yyasui/homepage.html.

1. INTRODUCTION

Analyzing microarray data at an individual gene level usually leads to a large list of significant genes, even after multiple comparison adjustments have been made. The process of trying to interpret such a large list of genes is difficult. Molecular biologists have put together lists of genes grouped by function, such as biological *pathways*, or sets of genes. Various pathways and gene sets databases have been compiled, for example, Kyoto Encyclopedia of Genes and Genomes (KEGG) [1,2], Gene Ontology [3], Biocarta [4] and Molecular Signature Data Base [5]. There has been a shift in focus from gene level analysis to pathway level, or gene set level, with many Gene Set Analysis (GSA) methods being proposed in the past decade. The most popular one is Gene Set Enrichment Analysis (GSEA) [6]. Extensive reviews and methodological discussions were given by Goeman and Buhlmann [7] and Nam and Kim [8].

While GSA methods on a single binary or categorical phenotype abounds, little attention has been paid to the case of a continuous phenotype, and there are no methods to accommodate multiple and correlated continuous phenotypes. Such correlated continuous variables are measured routinely and many important clinic-pathological observations such as lung functions, tumor size or measurements of marker proteins are continuous. A naïve approach to analyzing such data with existing GSA methods would be to categorize the continuous phenotypes into two or more discrete classes, as well as analyze the multiple correlated phenotypes univariately, i.e., one at a time. Such artificial categorization and univariate analyses may lead to less efficiency in gene-set analysis and even cause inaccurate identification of significant gene sets, especially if the multiple phenotypes exhibit moderate to large correlations.

There is an important methodological distinction between the *competitive* and *self-contained* GSA approaches [6, 7]. For a binary phenotype, e.g., competitive methods use gene permutation to test whether or not the association of the phenotype with a gene set is similar to its association with the other gene sets (the "Q1 hypothesis"), while self-contained methods employ sample permutation to test the equality of the two mean vectors of gene-set expressions which correspond to the two phenotype groups (the "Q2 hypothesis"). Here, we focused on the self-contained methods because, unlike the gene permutation approaches, sample permutation preserves correlations within gene sets -- a property that we have used to design the proposed method for continuous phenotypes.

Although correlations among gene expression measurements of biological pathways or functionally related gene sets have long been observed, to the best of our knowledge, only the modified Hotelling's T^2 test for categorical phenotype [9] and the Linear Combination Test (LCT) for binary phenotype [10], and for continuous phenotype [11] have used a covariance matrix estimator of gene expressions to compute the enrichment test statistic. It has been realized that incorporation of correlations among gene expressions in a GSA approach can significantly improve efficiency of the analysis [9]; however, it could also bring heavy computational burden as well. The Linear Combination Test was designed to incorporate correlations among gene expressions while overcome the computational burden. In the case of binary phenotype, it has been showed that LCT was much more computationally efficient than the modified Hotelling's T^2 test and approximated its superior power very well [10]; in the case of continuous phenotype, it has been showed that LCT was superior to other GSA methods under compare [11].

We propose here an extension of LCT to multiple continuous phenotypes, incorporating correlations among gene expressions of functionally related gene sets, as well as correlations among multiple phenotypes. Further, we extend our new method to its non-parametric version, referred as NLCT, to test potential non-linear associations between gene sets and multiple phenotypes, especially recommended for relatively larger microarray studies. The rest of the article is organized as follows. In section 2 we give detailed derivations of the two proposed methods. In section 3, we used simulation studies to compare performances of these two methods using variety of designs on different sample size, gene set size and settings on correlation among genes and correlation among phenotypes. Section 4 presents the performances of the proposed GSA methods using real microarray gene expression data from prostate tumor samples of African-American prostate cancer patients [13].

2. METHODS

2.1 Linear Combination Test for Multiple Continuous Phenotypes

Consider data from a microarray study measuring p gene expressions on a total of n subjects, a predefined gene set $X = (x_1, \ldots, x_p)^T$ and multiple continuous phenotypes consisting of q outcome variables $Y = (y_1, \ldots, y_q)^T$. Suppose columns in both X and Y are centered and scaled across the subjects. We are interested in testing whether there is a significant relationship between the gene set X and the multiple phenotypes Y. The null hypothesis to be tested is that expressions of the genes in the predefined gene set X are not associated with the phenotypes Y. One way of expressing this multivariate hypothesis linearly and univariately as a null hypothesis is:

H_0: There is no association between any of the linear combinations of $x_1, \ldots, x_p$ *and any of the linear combinations of* $y_1, \ldots, y_q$.

To address the linear relationship test, let $Z(X, A) = a_1 x_1 + \cdots + a_p x_p$ be a linear combination of $x_1, \ldots, x_p$, and $Z(Y, B) = b_1 y_1 + \cdots + b_q y_q$ a linear combination of $y_1, \ldots, y_q$, where $A \in R^p$ and $B \in R^q$ represent the coefficient vectors of a_i 's and b_j 's, respectively. For given coefficient vectors A and B of the combination coefficients, we can focus on testing whether the combination $Z(X, A)$ is associated with the combination $Z(Y, B)$. This is a classical correlation test and a commonly used test statistic is based on measuring the Pearson correlation between $Z(X, A)$ and $Z(Y, B)$, i.e. $\rho = \rho(Z(X, A), Z(Y, B))$. If both X and Y are normally distributed, then the variable $t = \rho\sqrt{\frac{n-2}{1-\rho^2}}$ follows a Student's t-distribution with degrees of freedom $n-2$ under the null hypothesis [19]. This also holds approximately if the observed values are non-normal, provided sample size n is large enough [20].

For testing the null hypothesis, we consider the linear combinations of $x_1, \ldots, x_p$ and $y_1, \ldots, y_q$, exhibiting the highest correlation, i.e. choosing coefficient vectors A and B to maximize the Pearson correlation between $Z(X, A)$ and $Z(Y, B)$. This leads to the proposed multiple version of the linear combination test (LCT) for testing the null hypothesis:

$$T^2 = \max_{A,B} \rho(Z(X, A), Z(Y, B)) \tag{1}$$

Here, we built upon the LCT method derivations for binary and continuous phenotype [10, 11], both of which are special cases of the multiple version of LCT.

Let $\Sigma_{XX} = \mathrm{cov}(X, X)$ be the covariance matrix of X whose (i, j) entry is $\sigma_{ij} = \mathrm{cov}(x_i, x_j)$; and similarly, let $\Sigma_{YY} = \mathrm{cov}(Y, Y)$ and $\Sigma_{XY} = \mathrm{cov}(X, Y)$ be the covariance matrix of Y and the covariance matrix between X and Y. The above statistic can be written as

$$T^2 = \max_{A,B} \frac{(A^T \Sigma_{XY} B)^2}{A^T \Sigma_{XX} A \bullet B^T \Sigma_{YY} B} \tag{2}$$

When the dimension of X and/or dimension of Y are high, singularity of Σ_{XX} and Σ_{YY} have to be taken care of very carefully, especially when the size of the gene set is larger than the sample size, i.e., $p > n$. A possible remedy for the singularity problem is to employ the shrinkage technique proposed by Schafer and Strimmer [12], and replace Σ_{XX} and Σ_{YY} with their shrinkage versions, namely, Σ^*_{XX} and Σ^*_{YY}. More specifically, the (i, j) entry of the shrinkage covariance matrix Σ^*_{XX} is given by $\sigma^*_{ij} = \gamma_{ij} \sqrt{\sigma_{ii}\sigma_{jj}}$, with shrinkage coefficients $\gamma_{ij} = 1$, if $i = j$, and $\gamma_{ij} = \rho_{ij} \min(1, \max(0, 1-\lambda^*))$, if $i \neq j$, where ρ_{ij} is the sample correlation between x_i and x_j, and the optimal shrinkage intensity can be estimated by $\lambda^* = \sum_{i \neq j} \mathrm{var}(\rho_{ij}) / \sum_{i \neq j} \rho^2_{ij}$. Based on this shrinkage strategy, we get the shrinkage version of the test statistic

$$T^{2*} = \max_{A,B} \frac{(A^T \Sigma_{XY} B)^2}{A^T \Sigma^*_{XX} A \bullet B^T \Sigma^*_{YY} B} \tag{3}$$

The computational cost of calculating (3) has to be taken into consideration, since the right hand side is a nonlinear programming problem involving $p + q$ parameters. The computational price for incorporating the shrinkage covariance matrices into the test statistic and maximizing directly the right hand side of (3) can be very high, especially when permutation is used for calculating p-value of the test. To address the computational efficiency problem, we adopt a strategy of using two groups of normalized orthogonal bases, instead of using the original observation vectors of

X and Y. We perform eigenvalue decompositions for the two shrinkage covariance matrices, $\Sigma_{XX}^{*} = UD_XU^T$ and $\Sigma_{YY}^{*} = VD_YV^T$, and obtain two groups of orthogonal basis vectors, $\tilde{X} = (\tilde{x}_1,\ldots,\tilde{x}_p) = (x_1 - \bar{x}_1,\ldots,x_p - \bar{x}_p)UD_x^{-1/2}$ and $\tilde{Y} = (\tilde{y}_1,\ldots,\tilde{y}_q) = (y_1 - \bar{y}_1,\ldots,y_q - \bar{y}_q)\ VD_Y^{-1/2}$. The test statistic in (3) can further be rewritten as

$$T^{2*} = \max_{\alpha,\beta}\frac{(\alpha^T\Sigma_{\tilde{X}\tilde{Y}}\beta)^2}{\|\alpha\|_2^2 \bullet \|\beta\|_2^2}. \qquad (4)$$

Where $\alpha = D_X^{1/2}U^TA$, $\beta = D_Y^{1/2}V^TB$ and $\Sigma_{\tilde{X}\tilde{Y}}$ is the covariance matrix between $\tilde{X}$ and $\tilde{Y}$, with its (i, j) entry being $\mathrm{cov}(\tilde{x}_i,\tilde{y}_j)$.

The optimization problem in (4) can be solved in two steps. Firstly, for a given β, find the optimal α, which is proportional to $\Sigma_{\tilde{X}\tilde{Y}}\beta$; secondly, substitute the optimal α into (4), and find the global optimal β, which is proportional to the first eigenvector of the matrix $\Sigma_{\tilde{X}\tilde{Y}}^T\Sigma_{\tilde{X}\tilde{Y}}$ corresponding to the largest eigenvalue. We note that the value of T^{2*} equals to the largest eigenvalue of $\Sigma_{\tilde{X}\tilde{Y}}^T\Sigma_{\tilde{X}\tilde{Y}}$.

The computation advantage is obvious when permutations are used to calculate p-value of the test, mainly because eigenvalue decompositions of the two shrinkage covariance matrices are not needed for the permuted versions of the data, but only for the original data.

2.2 *Nonlinear Combination Test for Multiple Continuous Phenotypes*

The proposed LCT method assumes a linear relationship between the genes in a gene set and the phenotypes. So do almost all the *self-contained* GSA approaches that have been proposed in the literature. The reason for us to focus on testing linear relationship is mainly for simplicity of the method. When we have access to limited data points, a simpler approach is more reliable than a complex/flexible one. If a larger sample size is available or if there is evidence that the relationships between gene sets and phenotypes could be non-linear/non-monotone, we may consider relaxing the linearity assumption, and testing more general null hypnoses, i.e., H_0: there is no relationship between genes in the gene set and the phenotypes.

The linear combination test proposed can be easily adapted to test both linear and nonlinear relationships between genes in a gene set and phenotypes, by using nonparametric techniques. The main idea here is to apply a non-linear transformation to the vector of genes X , then use linear test methods to check if there is a significant linear relationship between the non-linear transformation of X and the phenotypes Y. This strategy is similar to that of 'basis expansion' which is widely adopted in regression/discrimination analyses [21]. Some widely used non-linear transformations are polynomial transformations of single or multiple genes to achieve higher-order Taylor expansions; cubic splines or wavelets transformations of single genes. We note that the same transformation strategy can be applied to the phenotypes Y. We prefer to leave Y untransformed and keep the method not too high in flexibility, which requires larger sample size as well as higher computational cost.

3. SIMULATION STUDY

Our simulation study is designed to check performance of both LCT and NLCT methods. More specifically, we focus on type I error and power of the proposed tests, by varying gene-set size, sample size, and magnitude of correlations among genes.

We describe below our simulation study design. For each gene-set of size p, a gene expression matrix $\boldsymbol{X}_{n\times p}$ was generated from a multivariate normal distribution. The correlation between each pair of genes was set at ρ, with values of 0.0, 0.3, 0.6, or 0.9. For each gene set, a group of continuous phenotypes of size q were generated from the following multivariate linear model,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ , \tag{5}$$

where $\boldsymbol{\beta}_{p\times q}$ is a coefficient matrix, and $\boldsymbol{\varepsilon}_{n\times q}$ the error matrix generated from a multivariate normal distribution. The correlation between each pair of errors was set at ρ. In the null model, used to check the size of the test, we set all entries of $\boldsymbol{\beta}$ to 0, so that columns in $\boldsymbol{X}$ are not correlated with columns in $\boldsymbol{Y}$. In the alternative model, used to check the power of the test, we randomly selected five rows and three columns of the coefficient matrix, and set the corresponding fifteen entries to a common value μ, ranging from 0 to 5, with an increment of 0.25. The rest of the entries in the coefficient matrix were set at 0. We note that the five selected columns of $\boldsymbol{X}$ are correlated with the three selected columns of $\boldsymbol{Y}$, and that the correlation increases with μ. We used various sample sizes and gene-set sizes, including large p, and small n, a scenario common to gene-set analysis. The simulation data were replicated 1,000 times in each model. The p-values were based on 1,000 permutations.

The type I errors are similar across the LCT and NLCT methods (Table 1), with those of LCT closer to nominal level of 0.05, indicating lower sample size could lead to relatively higher type I errors of NLCT compare to LCT. As the sample size increases (n = 50) the type I error becomes closer to the nominal level 0.05, as expected from their use of permutation of phenotype labels.

Figure 1 illustrates the empirical power of both the LCT and NLCT methods using the nominal level of 0.05. The top left panel (n=20, p=20, q=10) shows power change of LCT with correlation among genes. At low correlation values, LCT appears to be conservative and less powerful, which may be explained by the fact that LCT is a test based on linear combination using shrinkage approach. Intuitively, higher magnitudes of correlations between genes imply lower level of variability of the linear combination of those genes. Similar phenomenon can be found for NLCT method on the bottom left panel (n=50, p=50, q=10). The top right panel (n=50, q=10 and ρ =0.6) shows power change of LCT with size of gene set. It implies that, with large gene sets, the efficiency of LCT test drops down significantly, i.e. larger sample size is required to test large gene sets. The bottom right panel (p=50, q=10 and ρ =0.3) shows the power change of NLCT with sample size, indicating low sample size could lead to very low power of test. Also comparing the two red lines in the bottom panels, we can see that NLCT is obviously less efficient than LCT when testing the linear association between genes and phenotypes.

4. APPLICATION

Leptin is a 16-kDa protein hormone that plays a key role in regulating energy intake and expenditure, including appetite and hunger, metabolism and behavior. It is one of the most important adipose-derived hormones [25]. Adiponectin (also refer to as GBP-28, apMI, AdipoQ and Acrp30) is a protein which in humans is encoded by ADIPOQ gene [26]. It is involved in regulating glucose level and fatty acid oxidation. Both leptin and adiponectin are well-known markers of human adiposity and obesity. These are hormones associated with various metabolic and inflammatory conditions. Interestingly, while leptin is found to be over-expressed in obese subjects, adiponectin is generally under-expressed, not just in adipose but also in other tissues. As the role of adipokines in prostate cancer is an important issue, we therefore considered using the continuous transcript levels of these dual markers, leptin and adiponectin, together as a multi-phenotype for application of LCT for geneset enrichment analysis of prostate cancer gene expression data.

We applied LCT to analyze a real Affymetrix microarray dataset consisting of genome-wide transcriptomic measurements of prostate tumor samples from African-American prostate cancer patients [13] against the continuous phenotype of the human leptin gene (*LEP*) and adiponection gene (ADIPOQ) transcript expression levels. The publicly available data were downloaded from Gene Expression Omnibus [22] [GEO: GSE6956]. The data that we used in the present study are part of a larger microarray study into immunobiological differences in prostate cancer tumors between African-American and European-American men. Because the *LEP* and ADIPOQ expression levels may be different between the two groups, we used only the data from the

African-American group to test the LCT methods. For our analysis, we used the expression values of 13,233 genes measured in tumor samples from 33 patients. The tumor samples were resected adenocarcinomas from patients who had not received any therapy before prostatectomy and were obtained from the National Cancer Institute Cooperative Prostate Cancer Tissue Resource (CPCTR) and the Department of Pathology at the University of Maryland. According to Wallace *et al*. [13], the macrodissected CPCTR tumor specimens were reviewed by a CPCTR-associated pathologist who confirmed the presence of tumors in the specimens. The tissues were collected between 2002 and 2004 at four different sites. The median age of prostatectomy was 61 and the median prostate-specific antigen (PSA) at diagnosis was 6.1 ng/mL. Fifty-five percent of the tumors were stage pT2, and 45% were stage pT3 or more. Detailed RNA extraction, labeling and hybridization protocols were as described previously [13].

For comprehensive gene-set analysis, the C2 catalog from MsigDB [8] consisting of 1,892 gene sets were used, including metabolic and signaling pathways from major pathway databases, gene signatures from biomedical literature including 340 PubMed articles, as well as other gene sets compiled from published mammalian microarray studies. Following Subramanian *et. al*. 2005 [8], 1,403 gene sets with size between 15 and 500 were used in our analysis. Each gene set was tested, using both LCT and NLCT approach, for its association with the *LEP* and *ADIPOQ* expression measurements.

We run firstly the univariate versions of LCT and NLCT for each of *LEP* and *ADIPOQ* expressions, followed by the multiple versions of LCT and NLCT using *(LEP, ADIPOQ)* expressions as a multiple phenotype. Table 2 shows percentages of gene sets with p-values less than 0.005, 0.01, 0.05, and 0.10. We expect LCT to be more suitable than NLCT for small to moderate sample sizes. Indeed, for our application, LCT is more efficient than NLCT. For large sample size and when non-linear relationship does exist, we expect NLCT to be more efficient. A larger percent of sets are associated with *LEP* than *ADIPOQ*. For some of the sets, the association with *LEP* is diluted by *ADIPOQ* in the multiple phenotypes analysis. However, 33 sets show a p-value smaller than 0.05 in the multiple phenotypes *(LEP, ADIPOQ)* analysis, although their univariate analysis indicated a p-value larger than 0.05 for each of *LEP* and *ADIPOQ* phenotype (Table 3).

5. DISCUSSION

We focused here on *self-contained* GSA methods. We note that *competitive* and *self-contained* methods test different hypotheses, and therefore it is important for the user to make an informed choice based on the hypothesis of interest and their understanding of the limitations of the two approaches (see reviews by Nam and Kim [7] and Dinu *et. al.* [24]). An important limitation of the self-contained approaches is that only a few genes can drive the association between the gene set and the phenotype. In such cases, post-hoc analysis can be used to reduce the gene set to a core sub-set associated with the phenotype. Such an analysis has been reported in simulations and in a real example for a binary phenotype [24].

Our proposed method is useful for testing associations between sets of genes or pathways and multiple and correlated continuous phenotypes. These are often measured in molecular epidemiology studies, and include clinico-pathological measurements such a tissue features, for example, tumor size, staining based readouts; cellular characteristics such as the amount of lymphocytic infiltration in a tumor environment; and subject-specific measurements such as diagnostic or prognostic marker protein or metabolite concentrations. The LCT algorithm can adjust for continuous or categorical covariates following a regression framework. The *LEP* and *ADIPOQ* levels in the prostate tumor example that we considered may also have been influenced by patient-specific covariates such as body mass index (BMI), age, and/or smoking status. We note that smoking status did not show a significant association with *LEP* expression levels (p-value = 0.36), or *ADIPOQ* expressions levels (p-value = 0.52) in our data, and BMI and age data were not available for our analysis.

The LCT approach can be used for both univariate and multivariate analyses. From the real data analysis, we can see that the univariate LCT for *LEP* is more sensitive/powerful comparing to the multiple LCT. Generally speaking, if we knew previously that more likely a subset of the group of phenotypes is associated with the gene sets than the rest of phenotypes, then focusing on the subset of the phenotypes will gain higher power for the test, for further information is incorporated in the testing. Here, we want to point out that naively (univariately) analyzing a group of multiple correlated phenotypes will lead to problems. In the real data analysis, for controlling type I error (e.g. ≤0.05), it is hard to set a threshold for the two univariate tests, because of correlations between *LEP* and *ADIPOQ*. If we can assume that the two phenotypes are independent, we can set a common threshold (roughly as 0.02532057) for them. We then get 209 (11.32%) significant gene sets tested by the naïve approach, but not including 67 (3.63%) of the 159 (8.61%) significant gene sets tested by the multiple LCT. This indicates that naïve approach can identify only gene sets associated with one of the two multiple phenotypes, instead of their combination.

LCT methods rely on the linearity assumption. To check the linearity assumption, exploratory data analysis should be used prior to running a formal inference. However, a small sample size which is common in microarray studies, would limit a thorough check for non-linearities. NLCT can be used for relatively larger sample sizes, or in the case the linear assumption does not hold. Our simulations and analyses of real microarray studies indicated LCT methods perform very well for small sample sizes. The question of how small is small is debatable and depends largely on the study design. In the case of a binary/categorical phenotype, at least five samples per group are desirable. In the case of a continuous phenotype, assessing significance based on less than 10 samples is dangerous; an alternative would be to rely upon representations that are more descriptive/exploratory in nature. In terms of computation, both LCT and NLCT are highly efficient compared to other GSA methods, especially given the incorporation of the covariance matrix into the estimations. A limitation of our study is that the findings come from a relatively small observational study and therefore cannot be generalized to other populations.

Acknowledgments

SP is supported by Ramalingaswami Fellowship of DBT and grants from MoS&PI, and DST CMS (Project No. SR/SA/MS:516/07 dated 21/04/2008).

REFERENCES

1. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M.; KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res. 40, D109-D114 (2012).

2. Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).

3. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium (2000) Nature Genet. 25: 25-29

4. Nishimura D: BioCarta. Biotech Software & Internet Report 2001, 2(3):117-120.

5. Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo and Jill P. Mesirov (2011). Molecular signature database (MSigDB) 3.0. *Bioinformatics*, 27 (12), 1739–1740.

6. Subramanian A, Tamayo P, Mootha VK et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci U S A;102:15545-15550.

7. Goeman, J. J. and Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics, 23, 980-7.

8. Nam, D. and Kim, S.Y. (2008) Gene-set approach for expression pattern analysis. Briefings in Bioinformatics, 9, 189–197.

9. Tsai C. and Chen J.J. (2009) Multivariate analysis of variance test for gene set analysis. Bioinformatics 25(7): 897-903.

10. Wang X, Dinu I., Liu W., Yasui Y. Linear Combination Test for Hierarchical Gene Set Analysis. *Statistical Applications in Genetics and Molecular Biology*, 2011, 10(1):Article 13.

11. Dinu I., Wang X., Vatanpour S., Kelemen L.E., Pyne S. Linear combination test for gene set analysis of a continuous phenotype. *BMC Bioinformatics. In Press.*

12. Schäfer, J. and K. Strimmer (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist. Appl. Genet. Mol. Biol. 4, 32.

13. Wallace T.A., Prueitt R.L., Yi M.H., Howe T.M., Gillespie J.W., Yfantis H.G., Stephens R.M., Caporaso N.E., Loffredo C.A., Ambs S. (2008). Tumor Immunobiological Differences in Prostate Cancer between African-American and European-American Men. Cancer Research, 68: 927-936.

14. Chang, S., Hursting, S. D., Contois, J. H., Strom, S. S., Yamamura, Y., Babaian, R. J., Troncoso, P., Scardino, P. T., Wheeler, T. M., Amos, C. I. and Spitz, M. R. (2001). Leptin and prostate cancer. Prostate 46: 62–67.

15. Saglam K, Aydur E, Yilmaz M, Goktas S. Leptin influences cellular differentiation and progression in prostate cancer. (2003). J Urol 169: 1308–11.

16. Singh S.K., Grifson J.J., Mavuduru R.S., Agarwal M.M., Mandal A.K., Jha V. (2010). Serum leptin: A marker of prostate cancer irrespective of obesity. Cancer Biomarkers, 7(1): 11-15.

17. Goktas S, Yilmaz, MI, Caglar K, Sonmez A, Kilic S, Bedir S. Prostate cancer and adiponectin. Urology 65(6): 1168 – 1172.

18. Bub JD, Miyazaki T, Iwamoto Y (2006) Adiponectin as a growth inhibitor in prostate cancer cells. Biochemical and Biophysical Research Communications 340: 1158–1166.

19. N.A Rahman (1968). A Course in Theoretical Statistics; Charles Griffin and Company.

20. Kendall, M.G., Stuart, A. (1973). The Advanced Theory of Statistics, Volume 2: Inference and Relationship, Griffin.

21. Hastie T, Tibshirani R, Friedman J (2009). The Elements of Statistical Learning. Springer.

22. Edgar, R., Domrachev, M. & Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30, 207–210.

23. Liu Q, Dinu I, Adewale AJ, Potter JD, and Yasui Y. (2007) Comparative Evaluation of Gene-set Analysis Methods, BMC Bioinformatics 8:431

24. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulsky KS, Halloran PF, Yasui Y. (2008) Gene Set Analysis and Reduction. Briefings in Bioinformatics, 10(1): 24-34.

25. Brennan AM, Mantzoros CS. (2006) Drug Insight: the role of leptin in human physiology and pathophysiology--emerging clinical applications. Nat Clin Pract Endocrinol Metab 2 (6): 318–327.

26. Maeda K, Okubo K, Shimomura I, Funahashi T, Matsuzawa Y, Matsubara K. (1996) cDNA cloning and expression of a novel adipose specific collagen-like factor, apM1 (AdiPose Most abundant Gene transcript 1). Biochem. Biophys. Res. Commun. 221 (2): 286–9.

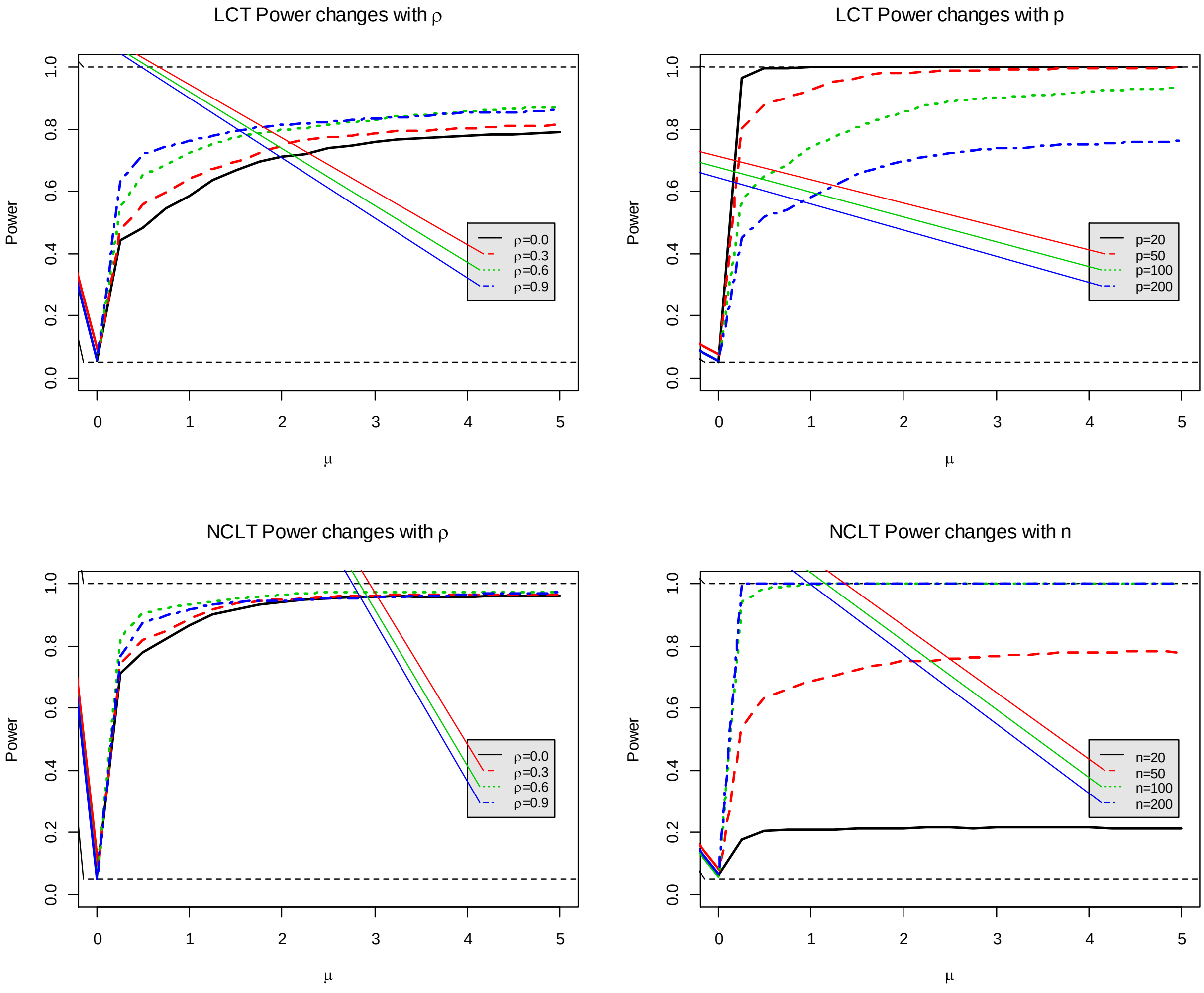

Figure 1. Power changes of the two GSA methods: LCT and NLCT.

Table 1. Type I errors of the multiple phenotype GSA methods LCT and NLCT, with dimension of the multiple phenotype set at q=10.

Method	**ρ**	***n*=20**			***n*=50**		
		***p*=20**	***p*=50**	***p*=100**	***p*=20**	***p*=50**	***p*=100**
LCT	0.0	0.050	0.047	0.047	0.052	0.043	0.047
	0.3	0.051	0.053	0.050	0.055	0.045	0.043
	0.6	0.050	0.054	0.042	0.051	0.042	0.048
	0.9	0.053	0.052	0.044	0.058	0.050	0.046
NLCT	0.0	0.041	0.039	0.035	0.062	0.049	0.058
	0.3	0.042	0.047	0.051	0.061	0.052	0.043
	0.6	0.044	0.062	0.047	0.052	0.049	0.042
	0.9	0.050	0.060	0.049	0.044	0.053	0.051

Table 2. Percentages of gene sets with p-values less than 0.005, 0.01, 0.05 and 0.10: univariate *LEP* and *ADIPOQ* LCT, and multiple phenotypes (*LEP*, *ADIPOQ*) LCT.

Method	**P-value**			
	≤.005	≤.01	≤.05	≤.10
LCT for *LEP*	2.8	4.5	19.9	36.1
LCT for *ADIPOQ*	0.4	0.9	3.1	6.3
LCT for *(LEP, ADIPOQ)*	0.9	1.5	8.6	18.6
NLCT for *LEP*	0.6	1.4	7.6	16.0
NLCT for *ADIPOQ*	0.3	0.7	3.8	10.1
NLCT for *(LEP, ADIPOQ)*	0.3	0.8	5.1	11.0

Table 3. Gene sets with multiple phenotypes (*LEP*, *ADIPOQ*) LCT p-values less than 0.05, although univariate *LEP* and *ADIPOQ* LCT p-values are larger than 0.05.

Gene-set name	Gene-set size	*LEP* p-value	*ADIPOQ* p-value	(*ADIPOQ*, *LEP*) p-value
YEN_MYC_WT	8	0.061	0.199	0.034
GLUCONEOGENESIS	50	0.195	0.192	0.036
BYSTRYKH_HSC_BRAIN_TRANS_GLOCUS	144	0.118	0.157	0.047
PENG_LEUCINE_DN	135	0.186	0.189	0.035
AMINOSUGARS_METABOLISM	14	0.223	0.341	0.039
PENTOSE_PHOSPHATE_PATHWAY	21	0.192	0.098	0.044
ZELLER_MYC_UP	22	0.09	0.262	0.032
POMEROY_DESMOPLASIC_VS_CLASSIC_MD_DN	38	0.093	0.079	0.011
FBW7PATHWAY	8	0.088	0.314	0.048
GSK3PATHWAY	24	0.18	0.225	0.028
GOLDRATH_CELLCYCLE	31	0.21	0.105	0.048
STREPTOMYCIN_BIOSYNTHESIS	8	0.094	0.072	0.013
FRUCTOSE_AND_MANNOSE_METABOLISM	24	0.168	0.061	0.007
GLYCOLYSISPATHWAY	8	0.137	0.076	0.013
UBIQUINONE_BIOSYNTHESIS	12	0.086	0.051	0.013
HOFMANN_MANTEL_LYMPHOMA_VS_LYMPH_NODES_UP	45	0.053	0.135	0.022
HOGERKORP_CD44_DN	22	0.057	0.504	0.05
CROMER_HYPOPHARYNGEAL_MET_VS_NON_DN	72	0.11	0.178	0.05
RUTELLA_HEPATGFSNDCS_UP	144	0.058	0.136	0.045
METHOTREXATE_PROBCELL_DN	11	0.102	0.132	0.033
GENOTOXINS_4HRS_DISCR	33	0.194	0.121	0.03
HTERT_UP	57	0.083	0.089	0.036
METHOTREXATE_PROBCELL_UP	14	0.147	0.111	0.046
CAMPTOTHECIN_PROBCELL_UP	17	0.085	0.159	0.038
UV_UNIQUE_FIBRO_UP	20	0.058	0.343	0.014
CITED1_KO_HET_DN	29	0.097	0.157	0.012
HEATSHOCK_YOUNG_UP	11	0.113	0.268	0.042
HSA00051_FRUCTOSE_AND_MANNOSE_METABOLISM	35	0.148	0.075	0.044
HSA00052_GALACTOSE_METABOLISM	27	0.053	0.277	0.044
HSA00521_STREPTOMYCIN_BIOSYNTHESIS	10	0.099	0.124	0.026
HSA01030_GLYCAN_STRUCTURES_BIOSYNTHESIS_1	83	0.106	0.25	0.049
HSA04080_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	218	0.103	0.116	0.044
HSA04120_UBIQUITIN_MEDIATED_PROTEOLYSIS	34	0.124	0.082	0.041