BIOINFORMATICS

Vol. 26 ECCB 2010, pages i511–i516
doi:10.1093/bioinformatics/btq380

De-correlating expression in gene-set analysis

Dougu Nam
School of Nano-Biotechnology and Chemical Engineering, Ulsan National Institute of Science and Technology, Republic of Korea

ABSTRACT

Motivation: Group-wise pattern analysis of genes, known as gene-set analysis (GSA), addresses the differential expression pattern of biologically pre-defined gene sets. GSA exhibits high statistical power and has revealed many novel biological processes associated with specific phenotypes. In most cases, however, GSA relies on the invalid assumption that the members of each gene set are sampled independently, which increases false predictions.

Results: We propose an algorithm, termed DECO, to remove (or alleviate) the bias caused by the correlation of the expression data in GSAs. This is accomplished through the eigenvalue-decomposition of covariance matrixes and a series of linear transformations of data. In particular, moderate de-correlation methods that truncate or re-scale eigenvalues were proposed for a more reliable analysis. Tests of simulated and real experimental data show that DECO effectively corrects the correlation structure of gene expression and improves the prediction accuracy (specificity and sensitivity) for both gene- and sample-randomizing GSA methods.

Availability: The MATLAB codes and the tested data sets are available at ftp://deco.nims.re.kr/pub or from the author.

Contact: dougnam@unist.ac.kr

1 INTRODUCTION

The basic goal of high-throughput gene expression profiling is to identify genes or groups of genes that are responsible for a phenotype of interest and elucidate their functional networks. Typical individual-gene analyses (Khatri and Draghici, 2005; Rivals *et al*., 2007) employ a cutoff threshold to define differentially expressed genes (DEGs) and then search for the biologically pre-defined gene sets that are enriched with the DEGs. However, the use of a threshold value causes a significant loss of information and is far from sufficient for describing the functionality of genes and their modular expression patterns.

On the other hand, the gene set analysis (GSA) methods assess the group-wise pattern of each pre-defined gene set by integrating the signals of all of its members, either strong or weak, without applying a cutoff threshold to genes (Ackermann and Strimmer, 2009; Kim and Volsky, 2005; Mootha *et al*., 2003; Nam and Kim, 2008; Tian *et al*., 2005). Such a group-wise approach of GSA covers a much larger spectrum of expression patterns and hence exhibits higher statistical power than an individual-gene analysis: many genes with relatively weak signals as well as small number of genes with strong signals in a gene set could be significant. Moreover, predictions by GSA are highly reproducible among data sets from independent experiments. Due to such advantages, GSA is becoming a powerful alternative to individual-gene analysis.

According to Tian *et al*. (2005), GSA methods can be classified into two categories depending on the null hypothesis tested as follows:

(1) Q1: The genes in a gene set have the same level of association with the phenotype compared with the rest of the gene set

(2) Q2: None of the genes in a gene set is associated with the phenotype of interest

These two categories can be characterized by how the significance of each gene set is assessed. After summarizing the signals in a gene set (gene set score), the significance of the summary statistic is assessed by randomizing gene labels (Q1) or sample labels (Q2), respectively. As the methods in the Q2 category hypothesize that no gene in a gene set is associated with the phenotype, they test the ‘existence’ of association signal in a gene set. For this reason, they can be termed *association analyses*. On the other hand, Q1 methods test the relative ‘enrichment’ of such association signal compared with the background genes; hence, they can be termed *enrichment analyses*. There are also hybrid-type methods that take into account both the gene and sample randomization of the gene set scores (Efron and Tibshirani, 2007; Mootha *et al*., 2003; Subramanian *et al*., 2005).

The main problem with most of the Q1-based GSA methods is that each gene set is assumed to be a collection of independent samples (from the entire list of genes). Because most biologically defined gene sets have some correlation structures in their expression profiles, this assumption mostly boosts some of the gene set scores and increases false positive predictions (Dinu *et al*., 2009; Goeman and Bühlmann, 2007; Newton *et al*., 2007). The recently developed restandardization method (Efron and Tibshirani, 2007) is basically built on sample randomization, but induces the independence of gene statistics by incorporating the gene-randomized gene set scores. Newton *et al*. (2007) suggested a simple method that adjusts for the different gene set sizes, which still maintains the independence of genes in the random-set model.

In this article, an algorithm dubbed *DECO* is proposed that removes the correlation bias in the expression of each gene set in GSA. The method is based on the eigenvalue-decomposition of the covariance matrix of each gene set and a series of linear transformations of data. This approach adjusts for the gene-set specific correlation structures to improve the power of many gene- or sample-randomizing GSA methods.

Eigenvalue-decomposition of covariance matrix has been widely applied to gene expression data analyses through principal component analysis (PCA) and its generalization, singular value decomposition (SVD) (Alter *et al*., 2000; Raychaudhuri *et al*., 2000; Yeung and Ruzzo, 2001). PCA and SVD are standard dimension reduction techniques and are mostly used for capturing global expression patterns of genes or arrays. Recently, PCA was combined with L_1 penalty to reduce noise effectively in identifying differentially expressed genes (Witten and Tibshirani, 2008). We present how to remove the correlation bias in GSA by rescaling the principal components of each gene set along the eigen-axes.

Because the accurate estimation of the covariance matrix of each gene set is critical in this method, we suggested using the shrinkage covariance estimator (Schafer and Strimmer, 2005) instead of the unstable sample covariance. We might also consider

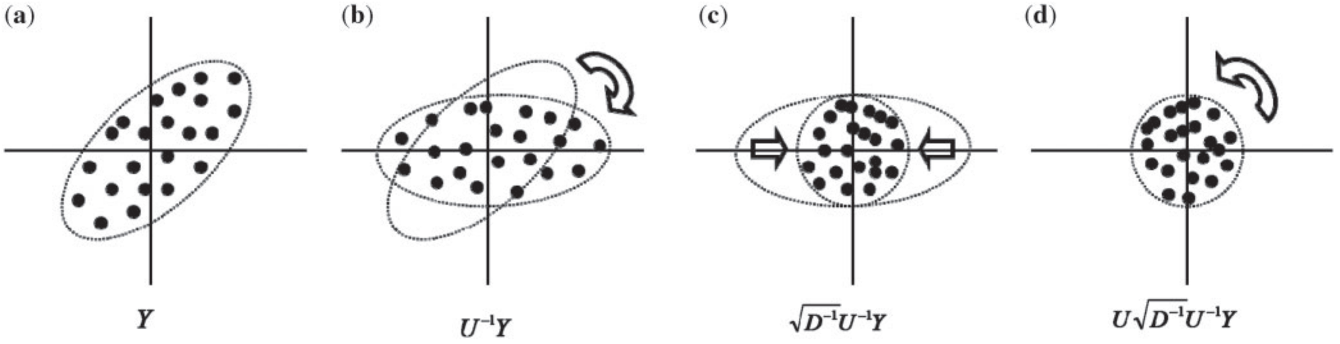

Fig. 1. De-correlation process for gene set expression data Y and covariance matrix C. Let $C = UDU^{-1}$ be the eigen-decomposition of C; each step then represents (**a**) normalized correlated data, (**b**) rotating the data to the eigen-axes, (**c**) shrinking the data to a unit ball and (**d**) returning the data to the original axes.

that there may still be errors in estimating covariance matrixes especially for large gene sets, and that de-correlating itself does not transform data into independent samples except for multivariate normal distributions. For these reasons, we suggest applying some moderate de-correlations for analyzing real experimental data to avoid overfitting. DECO is tested for two simulated data sets with different null hypotheses as well as two real expression data sets. The findings demonstrate that DECO overall improves the prediction of key pathways for both the gene- and sample-randomizing GSA methods.

2 METHODS

The main purpose of this study is to demonstrate the effects of de-correlation in GSA; hence, we adopt the simple average absolute t-statistic for the gene-set summary score and the random permutation of gene or sample labels to assess the significance of each summary score. The absolute values of each t-statistic were used to account for the possible bi-directional expression changes in each gene set.

2.1 DECO: removing correlation from data

The de-correlation process is illustrated in Figure 1 and described as follows:

Let X represent $n \times m$ expression profiles of a gene set with n genes and m samples.

(1) Normalize the profiles of each gene by taking log and Z-transformation, and let Y represent the transformed data.

(2) Estimate the covariance matrix (C) of n genes from Y.

(3) Apply the eigenvalue-decomposition to the positive-definite symmetric matrix C as follows: $C = UDU^{-1}$, where U is the $n \times n$ eigenvector matrix and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix with positive eigenvalues $\lambda_1, \ldots, \lambda_n$.

(4) Apply the linear transformations to Y. The meaning of each transformation is shown in Figure 1: $\overline{Y} = U \cdot \sqrt{D^{-1}} \cdot U^{-1} \cdot Y$, where $\sqrt{D^{-1}} = \mathrm{diag}(\lambda_1^{-1/2}, \ldots, \lambda_n^{-1/2})$. This gives $\bar{Y}$ an identity covariance matrix.

For Affymetrix data, both log and Z-transformation are required in *Step 1*: The log-transformation gives data a more symmetric distribution, and Z-transformation itself does not affect the individual t-statistic for each gene and is necessary for transforming the coordinate axes. In *Step 2*, we employed the shrinkage covariance estimator (Schafer and Strimmer, 2005) to estimate the covariance matrixes instead of the sample covariance. The shrinkage estimator provides a more accurate estimate which is always positive-definite such that all of the eigenvalues in *Step 3* become positive. See the next section for the shrinkage estimator. *Step 4* actually de-correlates the data by giving them a near spherical shape (Fig. 1). After reading the data in terms of the axes of eigenvectors (U^{-1}), each instance of the directionality of the data is restricted to a unit sphere ($\sqrt{D^{-1}}$). Through this transformation, large eigenvalues (>1) are shrunk to 1 while small eigenvalues (<1) are amplified to 1. However, some eigenvalues of large gene sets can be extremely small and amplifying them may yield unstable predictions. Moreover, large eigenvalues are most responsible for the coordinated patterns of gene expression. Therefore, we may focus on reducing large eigenvalues by using the following truncated diagonal matrix instead of $\sqrt{D^{-1}}$:

$$\sqrt{\overline{D}^{-1}} = \mathrm{diag}(\beta_1, \ldots, \beta_n), \quad \beta_i = \begin{cases} \lambda_i^{-1/2}, & \lambda_{\mathrm{i}} > T \\ T^{-1/2}, & \text{otherwise} \end{cases} \quad 0 < T \leq 1.$$

Here, $T = 1$ is assumed in our tests, which means that small eigenvalues are used as they are. We may choose other smaller values, but the results were not sensitive on this parameter and avoiding the amplification of very small eigenvalues matters. When we analyze real expression data, the normality of the data distribution is not guaranteed, and the number of samples may not be sufficiently large to estimate the covariance matrices of gene sets accurately. Therefore, we recommend using a square-root de-correlation that employs $\sqrt{\tilde{D}^{-1}} = \sqrt[\gamma]{D^{-1}}, \gamma = 4$ instead of $\sqrt{D^{-1}}$. This corresponds to taking one more square-root on $\sqrt{D^{-1}}$. We call these two methods moderate de-correlations because they less amplifies small eigenvalues or (and) less shrinks large eigenvalues than the original version. We denote the methods that use the truncated matrix ($\sqrt{\overline{D}^{-1}}$) as DECO-t, and the square root matrix ($\sqrt{\tilde{D}^{-1}}$) as DECO-sqrt. In the last step, the data are transformed into the original axes (U) to obtain the desired data. We note that even without this step, the transformed data $\sqrt{D^{-1}} \cdot U^{-1} \cdot Y$ have the identity covariance matrix. However, this additional step can remove the estimation error involved in the eigenvectors (U) by returning the data to the original axes. Moreover, it was possible to suggest the moderate versions of DECO because of this step.

2.2 Estimating a large covariance matrix with a small number of data samples

If the number of data samples is not sufficiently large compared to the dimension of the covariance matrix to be estimated, the standard maximum likelihood estimator S^{ML} or the sample covariance $S = \frac{n}{n-1} S^{ML}$ no longer provide a good approximation of the true covariance matrix. Moreover, S^{ML} and S often become ill-conditioned because some of the eigenvalues can become zero for large gene sets. Hence, they will be no longer positive-definite. Therefore, we adopt a recently developed shrinkage covariance estimator (Schafer and Strimmer, 2005) for estimating the covariance matrices of gene sets. Shrinkage covariance estimation compromises between the unbiasedness and small variance of the estimator and outperforms its previous methods. Moreover, it is well-conditioned and always positive-definite even with small number of samples. The shrinkage estimator $S^* = [s^*_{ij}]$ is defined as follows:

$$s^*_{ij} = \begin{cases} s_{ii}, & \text{if } i=j \\ r^*_{ij}\sqrt{s_{ii}s_{jj}}, & \text{otherwise} \end{cases} \quad \text{and}$$

$$r^*_{ij} = \begin{cases} 1, & \text{if } i=j \\ r_{ij}\min(1,\max(0,1-\bar{\lambda}*)), & \text{otherwise} \end{cases} \quad \text{with}$$

$$\bar{\lambda}* = \frac{\sum_{i\neq j}\overline{Var}(r_{ij})}{\sum_{i\neq j} r^2_{ij}}$$

where s_{ii} and r_{ij} denote the empirical variance and correlation. Let x_{ki} be the standardized k-th observation of the i-th variable (gene). Let $w_{kij} = x_{ki}x_{kj}$ and $\overline{w}_{ij} = \frac{1}{n}\sum_{k=1}^{n} w_{kij}$; this gives $\overline{Var}(r_{ij}) = \frac{n}{(n-1)^3}(w_{kij} - \overline{w}_{ij})^2$. See Schafer and Strimmer (2005) for a detailed explanation.

2.3 Gene set analysis with DECO

In each gene set with n members, we evaluate the summary score using the transformed data $\bar{Y}$. In an association analysis, we simply randomize the sample labels of $\bar{Y}$ to assess the significance of the summary score. In an enrichment analysis, we randomly choose n genes to compute the randomized summary scores. In this step, each randomized gene set should also be de-correlated for a fair comparison because a random gene set can have some level of correlations. Therefore, an enrichment analysis is more time-consuming compared to an association analysis when DECO is applied.

3 EVALUATION

3.1 A simulation test for correlated data without DEG

We begin by revisiting the simulation test for correlated gene sets conducted by Dinu *et al.* (2009). They generated correlated gene set expression profiles from multivariate normal distributions with no DEG between two sample groups. This induces the Q2 hypothesis. We repeated their experiment and examined the P-value distributions of gene sets for enrichment and association analyses. We generated expression profiles of 100 gene sets each having 20 dimensions (genes). The profiles of each gene set were generated using multivariate normal distributions with zero mean vectors and covariance matrixes with constant off-diagonal entries that were sampled from Unif (0.3, 0.8). Profiles of additional 2000 genes were also sampled independently from a standard normal distribution as a background distribution. We generated 20 samples in each of the two groups compared. The one hundred P-values sorted in the ascending order for the enrichment and association analyses are shown in Figure 2a. As no DEG was involved in this test, uniform distributions of P-values were desirable for bothtypes of analyses. However, the correlation structure increased the number of false positive predictions in the enrichment analysis such that the corresponding P-values exhibited an S-shaped distribution. DECO-t clearly corrected the bias of the P-value distribution in the enrichment analysis such that the corrected P-values also exhibited a uniform distribution (Fig. 2b). On the other hand, the association analysis was not affected by the correlations and exhibited a uniform P-value distribution in this test. However, we may not conclude at this stage that correlations do not disturb association analysis.

For a comparison, we also implemented the restandardized GSA (Efron and Tibshirani, 2007) for the average absolute t-score. The basic purpose of the restandardized GSA is incorporating both gene and sample randomizing effects in gene set analysis, and exhibited some advantages over existing methods. Its P-values also showed a uniform distribution (data not shown, codes provided) which implies the restandardized method does not yield additional false positives unlike enrichment analysis.

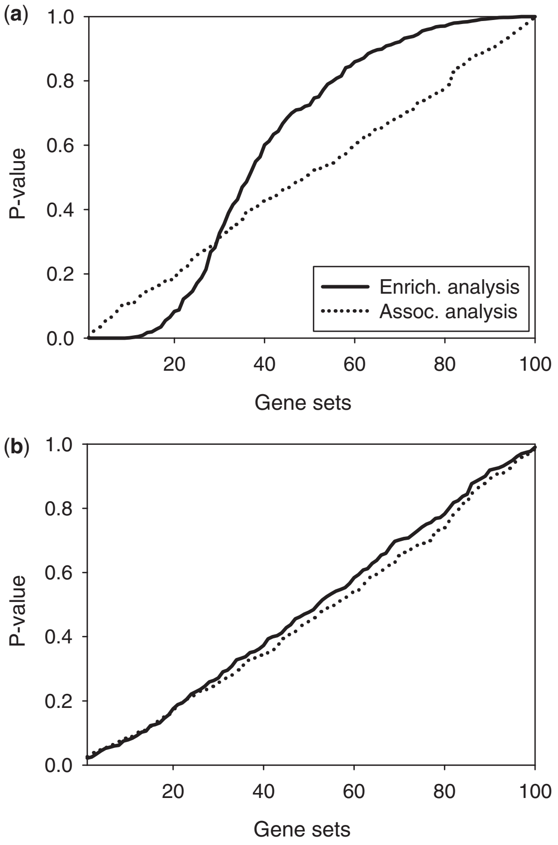

Fig. 2. Analysis results for (**a**) correlated and (**b**) de-correlated gene sets with no DEGs. Sorted P-values of a hundred gene sets are shown. Average t-statistic is used for gene set score and 1000 permutations were performed to evaluate P-values.

3.2 A simulation test for correlated data with DEG sets

This section considers a more general case that involves DEG sets and different gene set sizes. In this test, we let 70 gene sets have 20 members and 30 gene sets, 100 members. Among them, we generated 30 DEG sets (20 gene sets with 20-dimension and 10 gene sets with 100-dimension) by adding $\delta = 0.5$ in the second group on half of the members of the gene sets. Figure 3 shows (i) the ROC curves for the enrichment and association analyses, and (ii) the P-value distributions of each method, before and after applying DECO-t. In both types of analyses, DECO-t improved the accuracy considerably (Fig. 3a). Such improvements were also observed for different δ values and different portions of DEGs in each gene set. Besides, DECO was also tested for average squared t-score, and similar improvements were observed.

An interesting observation was that the improvement in the association analysis by de-correlation was not less than its counterpart, while the correlations did not affect the association

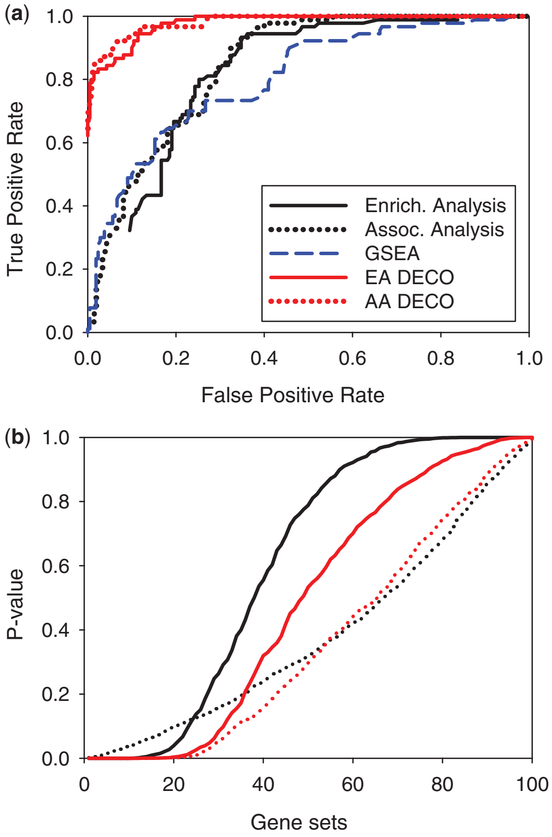

Fig. 3. (**a**) The ROC curves for correlated (black lines) and de-correlated (red lines) gene sets with 30 DEG sets. EA stands for enrichment analysis, and AA for association analysis. GSEA result is also shown. (**b**) The sorted *P*-values of the hundred gene sets are shown. DECO reduced the *P*-values for both enrichment and association analyses.

analysis for the data without DEGs in the preceding simulation test. This was consistently observed for different levels of correlations and other parameters. We infer that such an improvement by DECO in the association analysis is partially attributed to the distortion of the correlation structure by the DEGs: typically, a portion of genes in a gene set alter their expressions, and the gaps in the expression of these genes may be relatively amplified by the de-correlation process.

For a comparison, we implemented GSEA, one of the most widely used GSA methods (Mootha *et al.*, 2003; Subramanian *et al.*, 2005). It includes a step to adjust for the different correlations in each gene set by normalizing each set score from randomized set scores. We employed the maxmean set statistic which is simple but exhibits favorable properties over the original Kolmogorov–Smirnov statistic (Efron and Tibshirani, 2007), and computed the false discovery rates (FDR) of each gene set. The ROC was plotted based on the FDR values of each gene set. However, it was only comparable to the uncorrected average *t*-methods for low false positive rates.

We also implemented the restandardized GSA (Efron and Tibshirani, 2007) to predict the DEG sets in this test. Even though the restandardized GSA showed desirable properties in the preceding test, its ROC trajectory was overlapped with those of the uncorrected GSA methods (data not shown). This implies that the restandardization method does not improve the power of GSA for correlated data. This may be an expected result because the correlations affect both the gene- and sample-randomizing analyses to a similar degree and will also affect their combination (restandardized GSA) similarly. The MATLAB codes for the two competitive methods are also available from the author.

3.3 Analysis of p53-perturbed expression data

We tested the algorithms in an analysis of the p53-perturbed expression profiles (Subramanian *et al.*, 2005). The data had 33 and 17 samples in p53-mutant and wild-type groups, respectively. We used the 1892 curated functional gene sets denoted by C_2 in MSigDB (http://www.broadinstitute.org/gsea/msigdb). The expression data set is also available from MSigDB. We analyzed 1415 gene sets that had from ten to 100 members. Among them, we found 111 gene sets that contained the word ‘p53’ in their names or in their full descriptions. Given that only a portion of the p53-related pathways will actually alter their transcription patterns, we regarded gene sets with the median $P < 0.1$ as ‘responsive’ gene sets: The median *P*-values were determined among the three *P*-values in the uncorrected and the two moderate de-correlation analyses for each gene set. We drew 20 and 25 responsive gene sets from the 111 p53-related gene sets for enrichment and association analyses, respectively. Using them as the true positives, we compared the accuracy of the four methods: the uncorrected, DECO-t, DECO-sqrt and GSEA. The ROC curves are shown in Figure 4a and b. The DECO-sqrt showed most favorable results and improved the performance of GSA in both enrichment and association analyses. On the other hand, the performances of GSEA deteriorated in both cases. They were comparable to other methods for only small false positive rates. Table 1 shows how each prediction of the 25 key pathways in the association analysis was improved by DECO-sqrt. While the truncated de-correlation (DECO-t) improved the performance in the enrichment analysis, its superiority was not clear in the association analysis. This implies that the square-root de-correlation provides a good compromise for analyzing real data by reducing the correlation effects, while ameliorating the non-normality of the data and the estimation errors of the covariance matrices.

3.4 Analysis of normal/cancer prostate expression data

We conducted an additional test for normal/cancer prostate expression data (Chandran *et al.*, 2007; Yu *et al.*, 2004). The data are available from the GEO database (http://www.ncbi.nlm.nih.gov/projects/geo) with the series number GSE6919. We used the 171 samples with the platform of HG-U95A. The data are composed of 18 normal samples without any pathogenetic alterations, 63 normal samples from cells adjacent to prostate tumors, 65 prostate tumor samples and 25 metastatic tumor samples. Regarding the first two groups as normal, and the last two groups as cancer samples, we conducted enrichment and association analyses. We used the same gene sets C_2 from MSigDB. We used the keywords ‘cancer’, ‘tumor’, ‘oncogen’ and ‘carcinoma’ to find relevant gene sets. We first chose the gene sets that contained the keywords at least once in their names or at least twice in their full descriptions. Among them, we drew 201 gene sets with a median $P < 0.1$ as true positives in the enrichment analysis. DECO-sqrt still performed best among the three methods, and DECO-t performed only slightly better than the

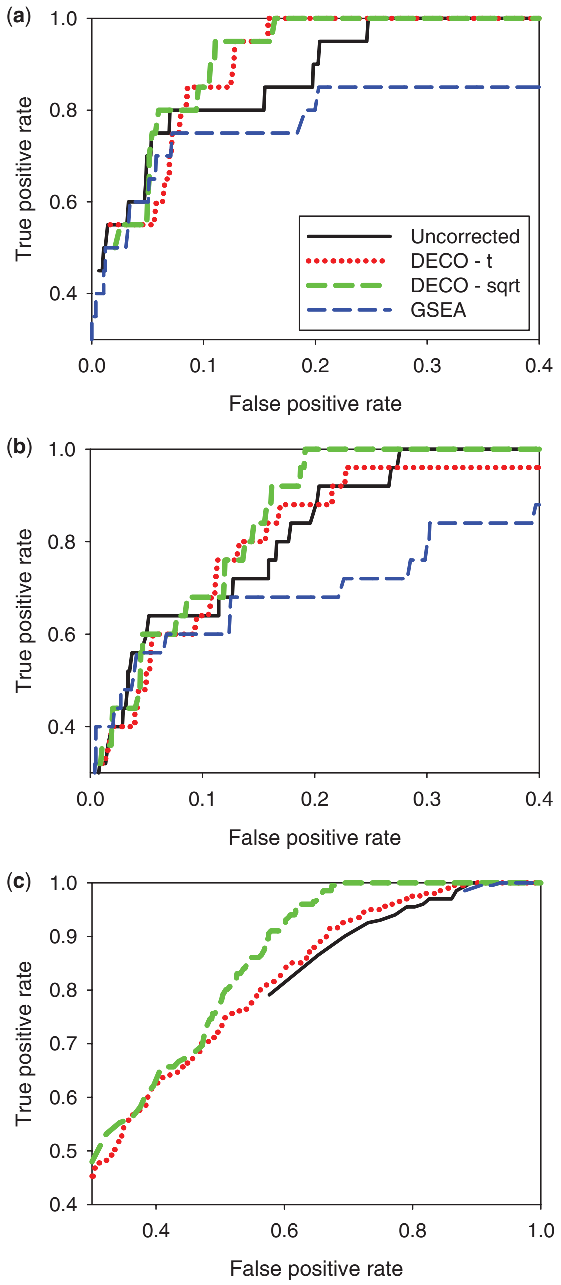

Fig. 4. The ROC curves for p53 perturbed data for (**a**) enrichment and (**b**) association analyses as well as GSEA and (**c**) the ROC curves for normal/cancer prostate data for enrichment analysis.

uncorrected method (Fig. 4c). In the uncorrected analysis, a majority of the gene sets (859) had an empirical P-value of zero, which caused the corresponding ROC to be very short. We also implemented GSEA for this data set, but its ROC was even much shorter than that of uncorrected analysis.

In the association analysis of this data set, all three methods detected too many gene sets with an empirical P-value of zero (over 1360 gene sets among 1415 sets: >96%); therefore, we could not conduct a proper comparison for this case. Since a considerable number of genes alter their expression levels between cancer and normal cells, most gene sets included at least one DEG; hence, the association analysis detected most gene sets as significant.

4 DISCUSSION

In this article, we devised an algorithm to remove (or reduce) the correlations in a gene set which perturb the inference of GSA considerably. By removing the correlations, we can transform the data into nearly independent samples. As uncorrelatedness does not necessarily imply the independence of samples, and the estimation of covariance matrices can be inaccurate, we recommended using a conservative de-correlation (DECO-sqrt) for real expression data to avoid overfitting. Indeed, in the real data analyses, DECO-sqrt showed favorable performances over DECO-t as well as the uncorrected method. When we applied the original DECO to real expression data sets, it showed a rather unstable performance: its ROC for the prostate data was a little better than the curve for DECO-t, but was even worse than the uncorrected method for the p53 data. This instability may be caused by the inaccurate estimation of small eigenvalues.

Ideally, it would be desirable to remove correlations of the full list of genes at once rather than de-correlating individual gene sets, because correlations exist not only within pre-defined gene sets but among genes across different gene sets. However, it is mostly inaccurate and infeasible to estimate and decompose such a high dimensional covariance matrix. In this regard, DECO serves as a practical solution for the correlation issue by locally de-correlating each gene set.

In DECO, reducing the estimation error for the covariance matrix of a gene set was crucial. The use of the well-known sample covariance yielded poor predictions in the analyses of real data sets. Therefore, the use of a shrinkage covariance estimator (Schafer and Strimmer, 2005) is essential when applying DECO to GSA.

For the same reason, DECO may perform better for over dozens of data samples.

Note that in simulation tests, all the 100 gene sets had relatively strong correlations between 0.3 and 0.8. However, in real expression data, only a portion of gene sets have such strong correlations. This may be a reason why the de-correlation effect was clearer in simulation tests.

A possible weakness with de-correlation analysis is that the members of de-correlated gene sets are the linearly transformed genes, which makes it hard to interpret the individual genes' behavior. However, gene sets, the basic analysis units of GSA remain the same and interpretation of gene sets is not interrupted by de-correlation. One may also apply the individual gene analysis on top of the GSA results to complement the analysis.

The correlation bias is a commonly recognized problem in gene set analysis. Correlation in a gene set is known to increase the number of false positive predictions for the gene-randomizing GSA methods; but we also found in this study that correlations can reduce the power of a sample-randomizing GSA method. DECO provides an intuitive and effective solution to these problems.

ACKNOWLEDGEMENTS

The author is grateful to Profs. Taesung Park, Sangsoo Kim and Seon-Young Kim for valuable discussions.

Table 1. The 25 responsive p53-related gene sets and their ranks of *P*-values in association analysis

Responsive p53 gene sets	Rank of gene set (Uncorrected)	Rank of gene set (DECO-sqrt)	Ranks improved
P21_MIDDLE_DN	396	184	212
HBX_HEP_UP	407	248	159
DNMT1_KO_UP	241	135	106
BRCA1_SW480_UP	296	210	86
MMS_HUMAN_LYMPH_HIGH_24HRS_DN	268	184	84
LIZUKA_G2_GR_G3	251	222	29
P53_SIGNALING	60	36	24
OXSTRESS_BREASTCA_UP	56	36	20
HSA04115_P53_SIGNALING_PATHWAY	17	1	16
HBX_NL_UP	305	291	14
BLEO_HUMAN_LYMPH_HIGH_24HRS_UP	28	22	6
ZMPSTE24_KO_DN	78	75	3
KANNAN_P53_UP	1	1	0
P53PATHWAY	1	1	0
P53HYPOXIAPATHWAY	1	1	0
BLEO_HUMAN_LYMPH_HIGH_4HRS_UP	1	1	0
P53GENES_ALL	1	1	0
STRESS_P53_SPECIFIC_UP	1	1	0
MMS_HUMAN_LYMPH_HIGH_24HRS_UP	1	1	0
HASLINGER_B_CLL_11Q23	63	68	−5
SHEPARD_NEG_REG_OF_CELL_PROLIFERATION	51	75	−24
FSH_HUMAN_GRANULOSA_UP	86	121	−35
HASLINGER_B_CLL_12	32	80	−48
SHEPARD_POS_REG_OF_CELL_PROLIFERATION	177	239	−62
SHEPARD_CRASH_AND_BURN_MUT_VS_WT_UP	195	285	−90

Funding: National Research Foundation of Korea grant funded by the Korean government (MEST; 2010-0016668); The Korea Research Council of Fundamental Science & Technology (KRCF; C-RESEARCH-9-10-NIMS); Ulsan National Institute of Science and Technology.

Conflict of Interest: none declared.

REFERENCES

Ackermann,M. and Strimmer,K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.

Alter,O. *et al*. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97,** 10101–10106.

Chandran,U.R. *et al*. (2007) Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer*, **7**, 64.

Dinu, I. *et al*. (2009) Gene-set analysis and reduction. *Brief Bioinform.*, **10**, 24–34.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.

Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Kim,S.Y. and Volsky,D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.

Mootha,V.K. *et al*. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Nam,D. and Kim,S.Y. (2008) Gene-set approach for expression pattern analysis. *Brief Bioinform.*, **9**, 189–197.

Newton,M.A. *et al*. (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat.*, **1**, 85–106.

Raychaudhuri,S. *et al*. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.*, **5,** 455–466.

Rivals,I. *et al*. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

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

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tian,L. *et al*. (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA*, **102**, 13544–13549.

Witten,D.M. and Tibshirani,R. (2008) Testing significance of features by lassoed principal components. *Ann. Appl. Stat.*, **2**, 986–1012.

Yeung,K.Y. and Ruzzo,W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**, 763–774.

Yu,Y.P. *et al*. (2004) Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J. Clin. Oncol.*, **22**, 2790–2799.