BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 10 2007, pages 1225–1234
doi:10.1093/bioinformatics/btm092

Gene expression

Domain-enhanced analysis of microarray data using GO annotations

Jiajun Liu[1,*], Jacqueline M. Hughes-Oliver[1] and J. Alan Menius, Jr[2]

[1]Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA and
[2]GlaxoSmithKline Research and Development, Research Triangle Park, NC 27709-3398, USA

Received on October 10, 2006; revised on January 10, 2007; accepted on March 5, 2007

Advance Access publication March 22, 2007

Associate Editor: Trey Ideker

ABSTRACT

Motivation: New biological systems technologies give scientists the ability to measure thousands of bio-molecules including genes, proteins, lipids and metabolites. We use domain knowledge, e.g. the Gene Ontology, to guide analysis of such data. By focusing on domain-aggregated results at, say the molecular function level, increased interpretability is available to biological scientists beyond what is possible if results are presented at the gene level.

Results: We use a 'top–down' approach to perform domain aggregation by first combining gene expressions before testing for differentially expressed patterns. This is in contrast to the more standard 'bottom–up' approach, where genes are first tested individually then aggregated by domain knowledge. The benefits are greater sensitivity for detecting signals. Our method, domain-enhanced analysis (DEA) is assessed and compared to other methods using simulation studies and analysis of two publicly available leukemia data sets.

Availability: Our DEA method uses functions available in R (http://www.r-project.org/) and SAS (http://www.sas.com/). The two experimental data sets used in our analysis are available in R as Bioconductor packages, 'ALL' and 'golubEsets' (http://www.bioconductor.org/).

Contact: jliu6@stat.ncsu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Advances in microarray technology have greatly enhanced gene expression studies. In these studies, thousands of genes are monitored and probed on only a few to tens of samples. Gene expression data present both great opportunities and challenges. Patterns of gene expression can be used to determine genes with similar behavior, suggest biomarkers for a specific disease and propose targets for drug intervention. However, several aspects of gene expression data analysis make it difficult to apply classical statistical methods:

- All gene expression data share the common problem of $p \gg n$, where p is the number of genes and n is the number of samples. A typical microarray data set has thousands of genes, but often less than 100 samples.
- Many of the genes are involved in multiple biological pathways and are therefore highly correlated.
- Interpretation of analysis output is problematic when hundreds of genes are identified as important.

Various dimension reduction approaches have been developed to alleviate the problem of $p \gg n$. For example, Li and Li (2004) proposed a dimension reduction method by combining principal components analysis (PCA) and sliced inverse regression (SIR) to produce linear combinations of genes that capture both the underlying variation of gene expressions and the phenotypic information, and then use the extracted combination of genes in the subsequent survival model formulation.

Our goal, however, is not only to solve the common $p \gg n$ problem of high-dimensional data, but also to enhance the interpretability of the models within the context of biological knowledge. While Li and Li (2004) offer a reduced-dimension model with some added interpretability, the interpretability is driven by statistical learning, not biological context.

Others have tried to analyze gene expression data focusing on domain-aggregated results to increase interpretability. The most common methods use a standard 'bottom–up' approach, where genes are first tested individually and then aggregated according to biological functions or pathways by domain knowledge such as the Gene Ontology (GO) by Ashburner *et al.* (2000). This type of knowledge structure provides great utility in the annotation and biological interpretation of gene sets obtained in microarray experiments. Ashburner *et al.* (2000) enable functional annotations of a given gene set by clustering genes according to their biological characteristics. Furthermore, through the GO hierarchical structure, it is also possible to represent biological concepts with different conceptual levels, from very general to very precise.

Various software tools are currently available for the ontological analysis of high-throughput gene expression experiments. GenMapp (Dahlquist *et al.*, 2002), ChipInfo (Zhong *et al.*, 2003), GoMiner (Zeeberg *et al.*, 2003), GeneMerge (Castillo-Davis and Hartl, 2003), FatiGO (Al-Shahrour *et al.*, 2003), OntoTools (Draghici *et al.*, 2003b), FuncAssociate (Berriz *et al.*, 2003), GOstat (Beissbarth and Speed, 2004) and

[*]To whom correspondence should be addressed.

Table 1. Contingency table classification of genes according to whether they are identified as differentially expressed (flagged) and whether they are part of the GO term

	Flagged genes	Non-flagged genes
In GO term	n_{11}	n_{12}
Not in GO term	n_{21}	n_{22}

ErmineJ (Lee *et al.*, 2005) are some of the popular ones. Khatri and Draghici (2005) did a comparative study of some commonly used tools for analyzing gene expressions in light of GO enrichment; their study was from a software point of view, and as such focused on aspects such as scope of the reported analysis, user interface, platform, type of application (web based or not), etc. Though most of these tools have different implementation, they use the same basic procedure. All tools start by identifying differentially expressed genes via some ordering metric. Statistical hypotheses are then created to test whether each GO term contains an unusually large number of differentially expressed genes, in which case that GO term is identified as important.

Several statistical approaches have been used to calculate *P*-values for each GO term. Draghici *et al.* (2003a) showed that the number of significant genes in a given term can be modeled by a hypergeometric distribution, and hence a test may be carried out with the distributional information. More popular approaches are the χ^2 test for equality of proportions and Fisher's exact test. To apply these two tests, a contingency table is created as shown in Table 1. Man *et al.* (2000) performed extensive simulations to show that the χ^2 test has better power and robustness than Fisher's exact test. Fisher's exact test is usually applied only when at least one of the expected values in a contingency table is smaller than five because in this case the χ^2 test is no longer appropriate.

A different approach is proposed by Mootha *et al.* (2003), and it is called gene set enrichment analysis (GSEA); see also Subramanian *et al.* (2005). Rather than performing a test based on the number of differentially expressed genes in a GO term, they first create a score for each GO term based on the rank of significance for all individual genes contained in that GO term. The resulting *P*-value for the score of each GO term is found by comparison to the null distribution of the scores as obtained using permutation. Generally speaking, these 'bottom–up' gene set enrichment methods improve the ability to interpret results from gene-level analysis. The disadvantage is that they all require gene-level analysis prior to conducting a GO-level analysis. Results from the gene-level analysis are directly related to the effectiveness of these methods. For some of these pre-processing procedures, we need to identify flagged or non-flagged genes, meaning that a statistical threshold has to be set. This threshold can have a major impact on the analysis results, but is not easy to define. Several other approaches have been proposed, including the random effects model of Goeman *et al.* (2004), the PAGE approach of Kim and Volsky (2005), the approach of Tian *et al.* (2005), the SAFE approach of Barry *et al.* (2005) and the composite GO annotation approach of ADGO by Nam *et al.* (2006).

We propose a 'top–down' approach, in which we find a summary statistic for each GO term; this summary statistic can be viewed as a new latent variable created from the individual genes in that GO term. These sets of biologically related genes can create a smaller number of new variables with informative descriptions of the biology. If the ontology is accurate, we shall see most of the highly correlated genes being grouped in one set. On one hand, we can reduce the number of variables and alleviate the problem of $p \gg n$. On the other hand, we are able to increase interpretability of our findings like the other enrichment methods. These new meaningful variables can be used to either test for significance of each GO term or make predictive models. In this article, we focus on the testing aspect of our methods. We introduce and explain our procedure in the Methods section. In the Results section, we use both simulated data and results from actual experiments to demonstrate our procedure's ability to identify significant GO terms. Our method is also compared to other approaches. We discuss advantages and potential drawbacks of our method in the last section.

2 METHODS

Our proposed method is called domain-enhanced analysis (DEA). It is a 'top–down' approach that aggregates the genes before we proceed to the analysis. The procedure is described below:

(1) Group genes into each GO term. (Some genes will exist in multiple terms.)

(2) Create a new latent variable for each GO term by combining the individual genes within that GO term.

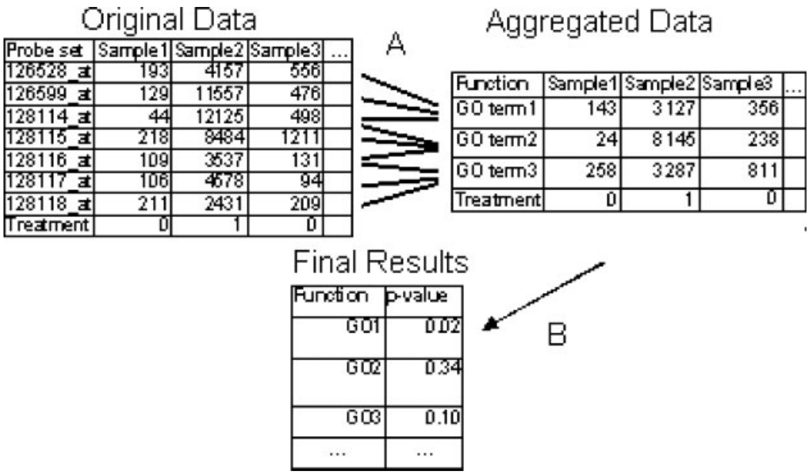
(3) Use the latent variable to test the predictive ability of each GO term.

(4) Adjust *P*-values accordingly for multiple testing.

Figure 1 is a flow chart to illustrate our method. The procedure is straightforward but has much room for expansion according to choices made for steps 2–4. The simplest choice for step 2 is to use an unweighted average as the latent variable; we call this approach DEA-Mean. We also investigated aggregation based on the first principal component (DEA-PCA) and on the first latent variable (score) from partial least squares (DEA-PLS) by Hoskuldson (1988) or de Jong (1993). We use the 'PLS' procedure in SAS to find PLS latent variables and the 'PRINCOMP' procedure in SAS to find PCA latent variables.

Consider the set of genes for a GO term as forming a predictor matrix X of n rows (samples) and p columns (number of genes in the set). A corresponding $n \times q$ matrix Y may represent q responses or q treatment factors for each sample; in either case we refer to Y as the response. PLS finds a linear combination of the columns of X that has maximum covariance with Y. PCA finds a linear combination of the columns of X to maximize the variance of X without regard to Y, and the mean is not designed to maximize any particular variation. Based on preliminary simulation results (not presented here), PLS is expected to outperform the other two options.

For the applications in this article, the 'response' is actually univariate and indicates membership in one of two classes. As such, y represents the n-vector of responses and we choose to use a 0/1 coding for these responses. Preliminary theoretical derivations show that both the 0/1 and $-1/1$ codings for y magnify effects of differentially



Probe set	Sample1	Sample2	Sample3	...
126528_at	193	4157	556	
126599_at	129	11557	476	
128114_at	44	12125	498	
128115_at	218	8484	1211	
128116_at	109	3537	131	
128117_at	106	4678	94	
128118_at	211	2431	209	
Treatment	0	1	0	

Function	Sample1	Sample2	Sample3	...
GO term1	143	3127	356	
GO term2	24	8145	238	
GO term3	258	3287	811	
Treatment	0	1	0	

Function	p-value
GO1	0.02
GO2	0.34
GO3	0.10
...	...

Fig. 1. A flow chart illustrates our procedure. (**A**) Steps 1 and 2 aggregate genes and create latent variables. For PLS procedure, treatments are considered as response variable. (**B**) Steps 3 and 4 test for significance of each latent variable with respect to treatment and adjust *P*-values for multiple testing.

expressed genes, but these codings do not result in equivalent test statistics or properties. This work is forthcoming in a separate article. Furthermore, our DEA procedure can be expanded to three or more classes. With either two binary vectors or one multi-class vector, the PLS procedure can be applied to find the first latent variable of the gene expressions. An ANOVA or F test can be applied instead of the *t*-test to test for predictive ability of each latent variable with respect to a multi-class response.

The PLS algorithm used in the article is by de Jong (1993). It is designed for continuous response variables, and we use it without modification even though in our case the response is a dichotomous quantity. The problem of using a categorical response in PLS has been considered by Bastien *et al.* (2005), Ding and Gentleman (2005), Fort and Lambert-Lacroix (2005), Huang and Pan (2003), Marx (1996), and Nguyen and Rocke (2002a, b, 2004). These alternative algorithms can potentially improve the summary latent variables for DEA-PLS. Other extension of PLS, such as the nonlinear PLS (Malthouse *et al.*, 1997), can also be worthy of consideration for special cases.

Because PLS finds a linear combination of X by looking at the relationship between X and y, it is subject to random variation. To overcome this, cross-validation is used to determine whether the first PLS latent variable significantly contributes to the prediction of y. After evaluating predicted residual sum of squares (PRESS) for a model where no PLS latent variables are retained (so that prediction occurs using the mean response) and a model that performs PLS regression using only one retained PLS latent variable, the test procedure by van der Voet (1994) is applied to detect significant improvement from the one-latent-variable model. If the first PLS latent variable for a GO term overfits y, we consider that gene set as non-important and simply omit it from steps 3 and 4 above. An alternative to completely dropping this gene set would be to assign it a very large *P*-value so that it gets very little attention in steps 3 and 4. By filtering gene sets in this manner, we limit the chances of retaining irrelevant latent summaries. As expected, we find that filtering is most intense for GO terms that contain large numbers of genes. It is likely that these cases require more than one PLS summary variable, but for now we limit ourselves to retaining at most one PLS latent variable; extension to retaining multiple PLS summaries is certainly possible. Cross-validation and van der Voet's test are implemented in SAS-PLS using options 'CVTEST' and 'CV=SPLIT' with a default significance level of 0.10.

For the testing procedure in step 3, we use the equal variance two sample *t*-test throughout the article. Other choices of tests are mentioned in the Discussion section. Adjustment for multiplicity in step 4 is done using the approach of Benjamini and Hochberg (1995) to control the false discovery rate (FDR); we will refer to this as the BH adjustment.

3 RESULTS

3.1 Simulation study

To investigate the properties of our proposed DEA method, we carry out several simulated studies. The focus of our simulation is to verify the effectiveness of our PLS summary measurement. We also compare DEA-PLS to DEA-Mean, the Fisher's exact approach and GSEA by Mootha *et al.* (2003).

For the simulated data set, we inherit the mapping structure between GO and genes from a real data set with 3666 genes. These genes are mapped to 556 GO terms from the molecular function hierarchy. Instead of using the expression levels from the experimental data, we simulate them so that we know what GO terms are supposed to be important. After selecting our binary response variable y, we generate differentially and non-differentially expressed genes from a conditional normal distribution to form matrix X.

We start by setting y to be either 1 or 0 by using a Bernoulli (0.5) distribution. The response y is set to be 1s and 0s throughout the article. Other choices of response variable y are mentioned in the Methods section. The sample size is set to 100. Each gene expression x is generated as: $x|y = 0 \sim N(-\mu, 1)$, or $x|y = 1 \sim N(\mu, 1)$. For non-important genes, μ is set to zero, while for important genes, we set μ to be some value depending on the extent of significance of each particular gene. In other words, differentially expressed genes are distributed as a mixture of two normals with common variance.

For this simulated study, there are 16 differentially expressed genes out of a total of 3666 genes: $\mu = \delta$ for genes 1514 to 1522 of X; $\mu = 1.2\delta$ for gene 2002; $\mu = 1.3\delta$ for genes 2872, 2874

Table 2. Gene counts within fully or partially important molecular functions

MFs	Non-important genes counts	Important gene counts
MF139	0	10
MF384	0	1
MF415	0	1
MF601	0	5
MF725	0	2
MF142	5	6
MF622	415	1
MF763	1	1
MF931	14	1

Table 3. BH-adjusted *P*-values of important GO terms from different methods

MFs	*P*-values for DEA-PLS	DEA-Mean	GSEA
MF139	9.92×10^{-19}	1.43×10^{-16}	<0.0001
MF415	8.65×10^{-9}	8.65×10^{-9}	6.24×10^{-1}
MF601	2.58×10^{-15}	4.81×10^{-14}	<0.0001
MF142	2.38×10^{-13}	4.43×10^{-4}	<0.0001
MF763	6.19×10^{-9}	4.43×10^{-4}	8.51×10^{-1}
MF725	1.03×10^{-8}	2.07×10^{-8}	7.37×10^{-1}
MF384	*Non-significant	3.87×10^{-1}	8.51×10^{-1}
MF931	*Non-significant	8.30×10^{-1}	8.51×10^{-1}
MF622	*Non-significant	8.30×10^{-1}	9.62×10^{-1}

*Cross-validated PLS resulted in no GO-term summary variable.
These results are from a single simulation replicate where $\delta = 0.3$. *P*-values for Fisher's exact test equal one in every case.

and 2887 to 2889; $\mu = 1.8\delta$ for gene 1283 and δ takes values $0, 0.1, 0.3, 0.5, 0.7, 0.9$. These differentially expressed genes are associated with nine molecular functions, as detailed in Table 2. Molecular functions 139, 384, 415, 601 and 725 are annotated only to differentially expressed genes, while molecular functions 142, 622, 763 and 931 are also annotated to some genes that are not simulated to be differentially expressed. We refer to the former as fully important molecular functions and to the latter as partially important molecular functions. The degree of partial importance for the latter group of molecular functions can be quite small as seen in Table 2. For example, molecular function 622 has only one differentially expressed gene but 415 non-differentially expressed genes.

We first create a single simulation replicate generated using $\delta = 0.3$. An equal-variance two-sample *t*-test is conducted for each gene, where samples are defined according to $y = 0$ or 1. Adjustment for multiplicity is done using the BH procedure. After adjustment, none of the genes are detected to be important using $\alpha_{\mathrm{I}} = 0.05$. Hence, if Fisher's exact test were used to test for significant GO terms, none would be found simply because no signal would be detected at the gene level. In other words, Fisher's exact test approach fails to identify any molecular function as being important since all Fisher's exact *P*-values equal one. It is possible to get around this by changing the gene-level threshold α_{I}, but finding a justifiable threshold can be problematic. Another option is GSEA. Since GSEA creates scores for each GO term using the gene ranking of importance, it avoids the problem of specifying α_{I}.

Table 3 provides results from GO-level analyses for the single simulation replicate we generated using $\delta = 0.3$. After obtaining either the PLS or mean summary for a GO term, an equal-variance two-sample *t*-test is conducted, where samples are again defined according to $y = 0$ or 1. For GSEA, we create scores for each GO term and test each score using a null distribution generated by permutation described by Mootha *et al.* (2003). The size of the permutation is 10 000. BH-adjusted *P*-values for these methods are presented in Table 3 for all nine important molecular functions. Using any reasonable α_{G} GO-level threshold for these multiplicity-adjusted *P*-values, it is clear that DEA-PLS and DEA-Mean outperform GSEA. By only looking at the rank ordering of the genes, GSEA loses power by ignoring the actual gene expression level.

DEA-PLS has the smallest *P*-values and finds six of the nine important GO terms with no false discovery. Two of the missed GO terms have marginal signals; both 'MF622' and 'MF931' only relate to one important gene and many unimportant genes. It is more likely that we should regard them as non-important GO terms. It is interesting to note that 'MF384' and 'MF415' are listed as important and non-important, respectively. Both have only one gene mapped to them, and they are both simulated to be important genes. At gene-level analysis, both genes are listed as non-significant, while at GO-level analysis, one is significant and the other is not. It is because at the gene level we are doing 3666 *t*-tests, but at the GO level, we are only doing 556 *t*-tests. With BH adjustment, it turns out that the threshold of significance is just between the two GO terms. In other words, by decreasing the number of tests, we are able to increase power even for those GO terms mapped to only one gene.

Previous results were limited to a single simulation replicate to allow the reader a detailed comparison of the Fisher's exact approach, GSEA, DEA-PLS and DEA-Mean. We now move to 50 simulation replicates to provide a more comprehensive comparison. Sensitivities and specificities of the four methods are shown in Figure 2. For these results, $\alpha_{\mathrm{I}} = 0.05$ and $\alpha_{\mathrm{G}} = 0.05$.

By looking at multiple runs of the simulated data sets, it is interesting to note that Fisher's exact test performs reasonably well for δ sufficiently large. Fisher's exact test has an average sensitivity of 54% when $\delta = 0.3$ and 77% when $\delta = 0.9$, which is much better than GSEA's 34% when $\delta = 0.9$. It is quite clear, however, that DEA-PLS is able to detect more true findings than the other three methods. It detects 20% more signals than Fisher's exact test and 60% more than GSEA. On the other hand, DEA-Mean only has marginally better sensitivity than Fisher's exact test. The improved sensitivity of DEA-PLS does not come at the cost of relevant loss of specificity. As seen in Figure 2, all methods have approximately equal levels of specificity. Hence, we can say that DEA-PLS is able to increase

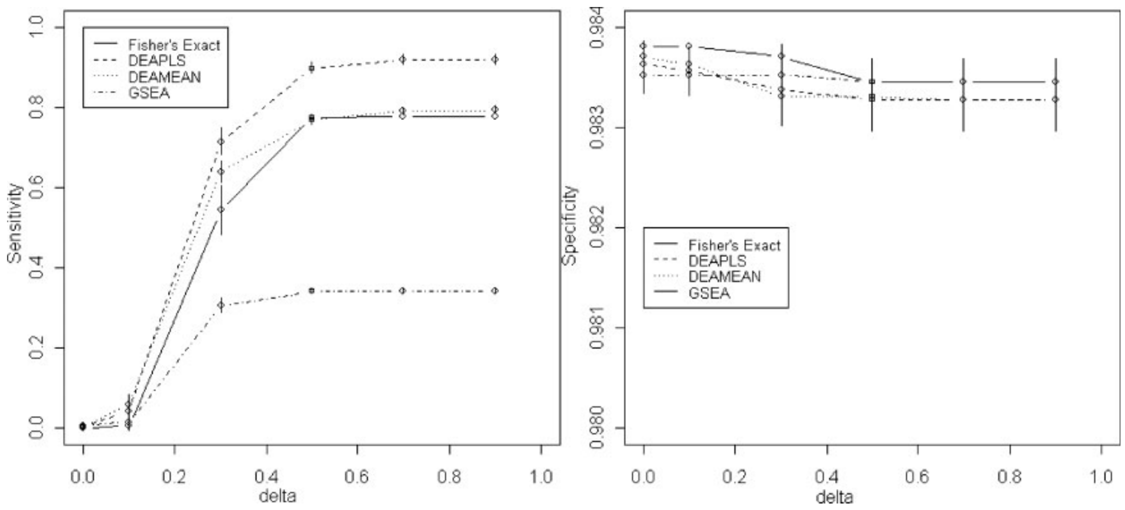


Fig. 2. Sensitivity and specificity of four methods as a function of δ in the simulation study. Sensitivities are based on averages across all simulation replicates, with pointwise approximate 95% confidence intervals shown as vertical bars. Fisher's exact test is solid line, DEA-PLS is dashed line, DEA-Mean is dotted line and GSEA is dashed-dotted line. $\alpha_I = 0.05$ for Fisher's exact test. All GO-level P-values were BH adjusted using $\alpha_G = 0.05$.

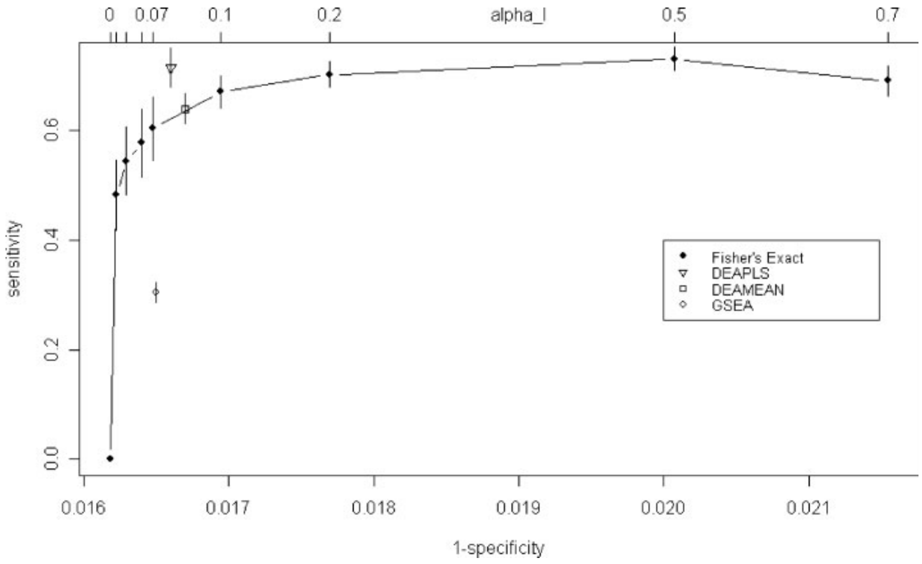


Fig. 3. ROC curve for Fisher's exact test by changing the threshold α_I for gene-level analysis. Data was generated following the simulation design using $\delta = 0.3$. Estimates are based on 50 simulation replicates. Vertical bars are pointwise approximate 95% confidence interval. Points for DEA-PLS, DEA-Mean and GSEA are also shown. These are single points because they do not depend on α_I.

the sensitivity of the analysis without having too many false discoveries.

We also use our simulation study to investigate the effect of α_I on Fisher's exact test. For 50 simulation replicates with $\delta = 0.3$, we apply Fisher's exact test repeatedly for $\alpha_I = 0$, 0.01, 0.03, 0.05, 0.07, 0.10, 0.20, 0.50, 0.70. The resulting ROC curve is presented in Figure 3. Clearly, sensitivity can change quite drastically as α_I changes. This shows that the effectiveness of Fisher's exact test is directly related to whether we choose the ideal α_I threshold for gene-level analysis. We also plot (sensitivity, 1-specificity) pairs for DEA-PLS, DEA-Mean and GSEA in the graph. DEA-PLS has a better sensitivity overall than Fisher's exact test at its best, while DEA-Mean performs almost equivalent to Fisher's exact test for one value of α_I. In our studies, GSEA has the worst sensitivity.

With the simulation studies, we can see that the drawback of most of the existing methods is that they require gene-level analysis, either finding the important genes or the rank of significance of the individual genes. For situations where gene-level signals are weak, these methods perform poorly, because their effectiveness relies largely on strong signals at the gene level. On the contrary, DEA methods do not require strong gene-level signals or gene-level analysis. Furthermore, they are able to combine the weak signals for genes related to each GO term and make them easier to detect. The simulation results show DEA methods are effective for finding GO-level signals. Fisher's exact test performs reasonably well when gene-level analysis does work, but the α_I threshold for gene-level analysis is critical for the effectiveness of Fisher's exact test. On the other hand, GSEA has the weakest power to detect signals at the GO level. For all simulated data sets, DEA-PLS has the best power to detect significant GO terms.

3.2 Experimental data study

We evaluate DEA-Mean and DEA-PLS against the classic Fisher's exact test on two real gene expression data sets by Chiaretti *et al*. (2004) and Golub *et al*. (1999). Both data were collected to identify genes that distinguish subgroups of leukemia patients. The first data set splits patients into B-cell- and T-cell-type acute lymphoblastic leukemia (ALL), while the second data set consists of patients with ALL and acute myeloid leukemia (AML). Both data sets have been extensively studied in the literature of microarray analysis and are available in R as Bioconductor packages, 'ALL' and 'golubEsets' (http://www.bioconductor.org/). We present the analysis results of Chiaretti's data set in the next subsection and results of Golub's data set as Supplementary Material.

3.3 Leukemia data set by Chiaretti

The data set consists of 128 patients with ALL. It is known that ALL cells are delivered from either B-cell or T-cell precursors. Among these patients, there are 95 with B-cell ALL and 33 with T-cell ALL. The HGU95aV2 Affymetrix chip was used for the experiment. We annotate the data using GO's biological process (BP) ontology. Then, 10 012 probe sets are annotated from a total of 12 625 probe sets in the data. The annotation includes 1955 GO term nodes. These GO terms all are mapped to at least one probe ID in the data.

In some literature, mappings are first done from probe sets to EntrezGene ID and then to the GO terms. In our implementation, we simply use the mapping directly from probe sets to GO terms to be consistent with DEA, which finds a summary for each GO term using all gene expressions in the GO term. We are careful not to add a step of mapping from probe sets to EntrezGene ID and thus weaken the comparison between DEA and Fisher's exact test. To keep our analysis consistent, for our implementation of Fisher's exact test, we also use the mapping from probe sets directly to GO terms.

To apply Fisher's exact test, we first perform a gene-level analysis on each probe to find differentially expressed probe sets. Functions 'lmFit' and 'eBayes' in R package 'limma' are used to fit a linear model for each probe set and produce a *P*-value for each one. The *P*-values are BH adjusted. Using level $\alpha_I = 0.01$, a total of 2016 probe sets are found to be significant. We then proceed to apply one-sided Fisher's exact tests using results from the gene-level analysis. We also apply DEA-PLS and DEA-Mean methods. All three methods find many signals at the GO-term level. We can possibly reduce the number of significant findings by adjusting our *t*-test, but it is not the focus of this article. For now, we suggest focusing on the most significant GO terms.

In Table 4, we present statistics for the 10 most significant GO terms detected by Fisher's exact test. They mostly consist of biological processes related to antigen processing and presentation. These findings are mentioned in previous biological literature including Dalla-Favera (2001), Look (2001) and Novak (2001). DEA-PLS is also able to detect these signals and the *P*-values from DEA-PLS are much smaller. DEA-PLS using all the information from the probe level has greater power than Fisher's exact test because information is lost when the actual expression level of each probe is ignored. Table 4 lists the *P*-values of each GO term

Table 4. GO-level analysis results or 10 most significant GO terms found by Fisher's exact test in Chiaretti's data

	GO term	O	A	Fisher's exact	DEA-PLS (rank)
1	Antigen presentation	40	59	1.30×10^{-12}	2.02×10^{-38}(89)
2	Antigen presentation, exogenous antigen	17	18	4.04×10^{-10}	2.07×10^{-35}(134)
3	Antigen processing	35	55	4.82×10^{-10}	7.20×10^{-37}(110)
4	Antigen presentation, endogenous antigen	30	45	1.62×10^{-9}	1.23×10^{-36}(115)
5	Antigen processing, endo-enous antigen via MHC class I	29	45	9.87×10^{-9}	1.55×10^{-34}(155)
6	Antigen processing, endogenous antigen via MHC class II	17	20	1.53×10^{-8}	1.98×10^{-35}(132)
7	Defense response	283	913	3.16×10^{-7}	1.02×10^{-53}(6)
8	Immune response	259	831	6.46×10^{-7}	1.93×10^{-54}(5)
9	Response to biotic stimulus	292	955	7.79×10^{-7}	2.05×10^{-53}(8)
10	Detection of pest, pathogen or parasite	11	13	7.28×10^{-6}	1.07×10^{-34}(151)

A—the number of probe sets annotated for each GO term.
O—the number of significant probe sets annotated for each GO term.

followed by the rank of that GO term by both methods. Three of the GO terms are consistently being selected by DEA-PLS and Fisher's exact test among their 10 most significant GO terms: immune response, defense response and response to biotic stimulus. Excluding these three GO terms, the most significant GO terms identified by Fisher's exact test are not the most significant ones identified by DEA-PLS.

In Table 5, the 10 most significant GO terms from DEA-PLS are listed. DEA-PLS assigns greater significance to GO terms related to cell activity than to those related to antigen processing and presentation. Specifically, DEA-PLS has smaller *P*-values for T-cell selection, immune cell activation and positive regulation of T-cell receptor signaling pathway. From the literature, distinct gene expression signatures related to cells functions were recognized by Dalla-Favera (2001), Look (2001) and Novak (2001). It also makes perfect sense that the best GO terms to differentiate T-cell and B-cell ALL would be those probes related to cell functions. From the results, we can see that Fisher's exact test detected some of these signals in a less significant role, for example, immune cell activation, cell activation and lymphocyte activation. For some other GO terms like T-cell selection and positive regulation of T-cell receptor signaling pathway, Fisher's exact test cannot effectively identify signals. At this point, we would like to know whether it is justifiable to claim that GO terms found by DEA-PLS are more relevant than those found by Fisher's exact test.

Without enough comparative studies in the literature on the extent of importance for these GO terms differentiating leukemia subtypes, we try to find the answer by drilling down to the probe level. Searching through probe-level analysis, we find that four of the five most important probes are clearly related to the most significant GO terms found by DEA-PLS. More specifically, lymphocyte differentiation, lymphocyte activation, immune cell activation and cell activation are related to the first, third, fourth, fifth and ninth of the 10 most significant probe sets. T-cell selection is one of the more precise terms. With only eight probe sets mapped to it, first and fourth most significant probe sets are among them. The most significant GO terms found by DEA-PLS are mapped to some of the most significant probe sets. Likewise, these probes were also significant using Fisher's exact test. The sixth, seventh, eighth and tenth most significant probe set are mostly related to antigen processing and presentation. DEA-PLS also finds these GO terms to be significant, but only in a slightly less significant role.

Some of the advantages of using GO to annotate genes are its hierarchical structure and many available tools for displaying GO terms. The GO web site http://www.geneontology.org/ lists many useful tools for searching and browsing GO terms. We use a web-based tool 'AmiGO' available at http://www.godatabase.org/cgi-bin/amigo/go.cgi to display our results from DEA-PLS and Fisher's exact test. The output graphic for Fisher's exact test is presented in Figure 4 and for DEA-PLS is displayed in Figure 5. In each figure, the 10 most significant GO terms are underlined. Their rank of importance for each respective method is at the upper left corner of each node, while the rank of importance of the most significant probe sets contained in the GO term are shown in the curly bracket. By displaying the hierarchical structures of GO terms, we are able to gain additional insight. If we only look at the 10 most significant GO terms found by both methods, these terms are made up of two distinct subtrees. First of all, these two subtrees are both descendants of the GO terms immune response, defense response and response to biotic stimulus, which are nested by sequence. This explains why these three GO terms are identified as significant by both DEA-PLS and Fisher's exact test. Furthermore, the important GO terms found by both methods are nested nicely to each other, which shows how these biological functions are being affected by different leukemia subtypes from more precise terms to more general terms. It is encouraging to see that the subtree of GO terms found by DEA-PLS is related to more significant probe sets than those terms found by Fisher's exact test, suggesting that DEA-PLS is maintaining the right order of importance for the GO terms found.

In conclusion, it is clear that both DEA-PLS and Fisher's exact test are able to pick out some useful significant GO terms. DEA-PLS has better power with all the information included, and hence it is able to find more signals at the GO-term level.

Table 5. GO-level analysis results for 10 most significant GO terms found by DEA-PLS in Chiaretti's data

	GO term	O	A	Fisher's exact (rank)	DEA-PLS
1	T-cell selection	5	8	2.31×10^{-2} (122)	5.96×10^{-57}
2	Immune cell activation	47	114	3.47×10^{-5} (12)	9.87×10^{-56}
3	Cell activation	47	115	4.52×10^{-5} (13)	9.87×10^{-56}
4	Transition metal ion homeostasis	9	28	2.13×10^{-1} (463)	1.24×10^{-55}
5	Immune response	259	831	6.46×10^{-7} (8)	1.93×10^{-54}
6	Defense response	283	913	3.16×10^{-7} (7)	1.02×10^{-53}
7	Lymphocyte activation	40	101	3.56×10^{-4} (22)	2.05×10^{-53}
8	Response to biotic stimulus	292	955	7.79×10^{-7} (9)	2.05×10^{-53}
9	Hemopoiesis	28	94	1.19×10^{-1} (314)	2.08×10^{-52}
10	Positive regulation of T-cell receptor signaling pathway	3	4	4.56×10^{-2} (165)	5.60×10^{-52}

A—the number of probe sets annotated for each GO term.
O—the number of significant probe sets annotated for each GO term.

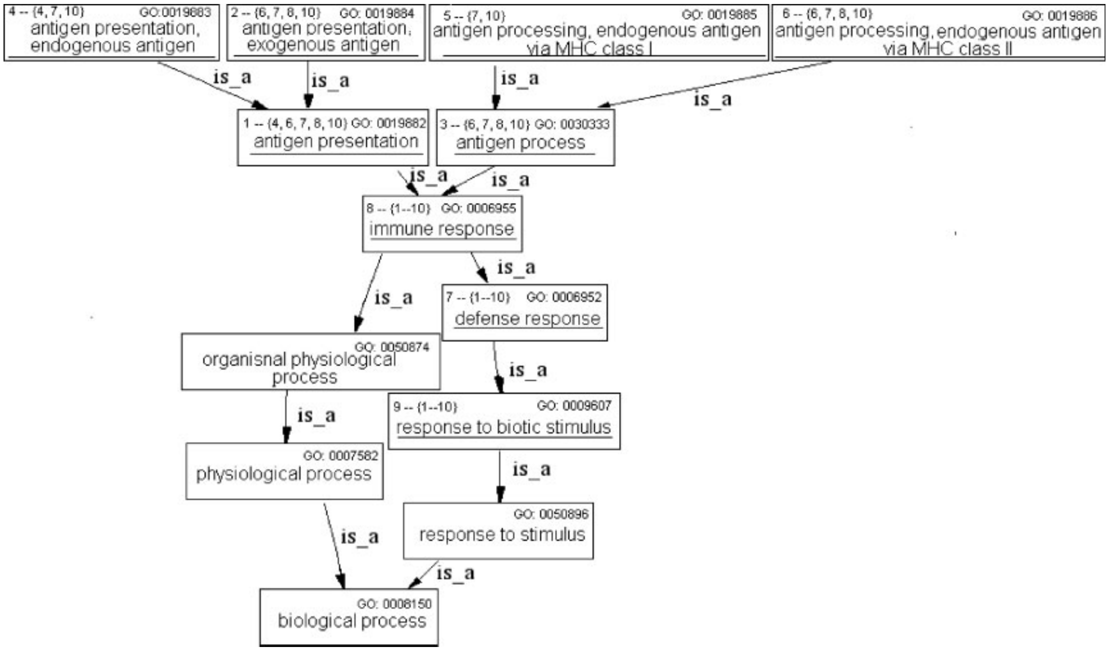

4 -- {4, 7, 10} GO:0019883
antigen presentation, endogenous antigen
2 -- {6, 7, 8, 10} GO: 0019884
antigen presentation, exogenous antigen
5 -- {7, 10} GO: 0019885
antigen processing, endogenous antigen via MHC class I
6 -- {6, 7, 8, 10} GO: 0019886
antigen processing, endogenous antigen via MHC class II
is_a
is_a
is_a
is_a
1 -- {4, 6, 7, 8, 10} GO: 0019882
antigen presentation
3 -- {6, 7, 8, 10} GO: 0030333
antigen process
is_a
is_a
8 -- {1--10} GO: 0006955
immune response
is_a
is_a
7 -- {1--10} GO: 0006952
defense response
GO: 0050874
organisnal physiological process
is_a
is_a
9 -- {1--10} GO: 0009607
response to biotic stimulus
GO: 0007582
physiological process
is_a
GO: 0050896
response to stimulus
is_a
is_a
GO: 0008150
biological process

Fig. 4. GO graph of findings by Fisher's exact test for Chiaretti's data. Specific terms are at the top, while general terms are at the bottom. Top 10 significant GO terms are underlined, with rank shown in upper left of rectangle. Ranks for top 10 probe sets are shown in curly brackets.

Furthermore, DEA-PLS is able to find the most significant GO terms that are related to the most significant probe sets. By combining the actual gene expression information for each GO term, DEA-PLS is able to maintain the right order of importance for the GO terms. This simply cannot be achieved by Fisher's exact test, because it treats all the important probe sets with equal weight. With four of the top five most significant probe sets related to biological processes like immune cell activation, lymphocyte activation and lymphocyte differentiation, etc. they certainly should be and are able to stand out from the rest by using DEA-PLS.

We also carried out DEA-Mean in the analysis, though we did not list the results here. DEA-Mean is straightforward and fast to implement, but we find that its behavior can be erratic. One obvious drawback is that it has very weak power when the number of probes related to a GO term becomes large. Hence, DEA-Mean finds immune response, defense response and response to biotic stimulus to be less significant while they are being recognized very significant by both DEA-PLS and Fisher's exact test.

4 CONCLUSION

We focused our work on functional analysis instead of individual gene-level analysis. By using the GO, DEA is able to increase interpretability of the analysis results without loss of sensitivity. As we presented in our analysis on Chiaretti's data, the results can be displayed as a hierarchical structure. Hence, we are able to find clusters of nested GO terms which can be used to differentiate leukemia subtypes.

Additionally, we addressed some of the problems of current pathway analysis methods. We propose to combine information and create a summary statistic for each GO term from the individual gene expressions, and then test for significance using the newly created variables. We considered two such summaries, mean and PLS score. As demonstrated in our simulation and analysis on two publicly available data sets, the new test has greater power to detect signals at the GO level than do methods such as Fisher's exact test or GSEA. These methods lose power because they rely only on the rank of significance for individual genes. Furthermore, in our analysis on Chiaretti's data set, we were able to elaborate that DEA-PLS maintains an appropriate order of importance for identified GO terms according to their actual gene expressions. Instead of testing for over-representation of each GO term only by the number of significant genes as Fisher's exact test does, DEA-PLS considers the extent of importance of each gene related to the GO term by finding a linear combination of gene expressions using PLS. DEA-PLS has another advantage that it does not require gene-level analysis, and thus avoids dependence on the effectiveness of gene-level analysis.

DEA-PLS also can be considered as a type of dimension reduction. While GSEA and Fisher's exact test provide summaries for each GO term, they do so by giving a single number that represents all samples. DEA, on the other hand, provides summaries for each GO term and each sample, thus providing a more specific summarization. By creating summary measurements for each GO term, the new data usually have fewer variables than the original gene-level data. These new variables represent meaningful biological functions instead of many individual genes. It could be advantageous to build a predictive model using these new variables because the predictive models would better reflect biological function and therefore be more interpretable.

In our analysis, we simply use a *t*-test to find significant GO terms after we combine the genes using PLS. It is suggested in

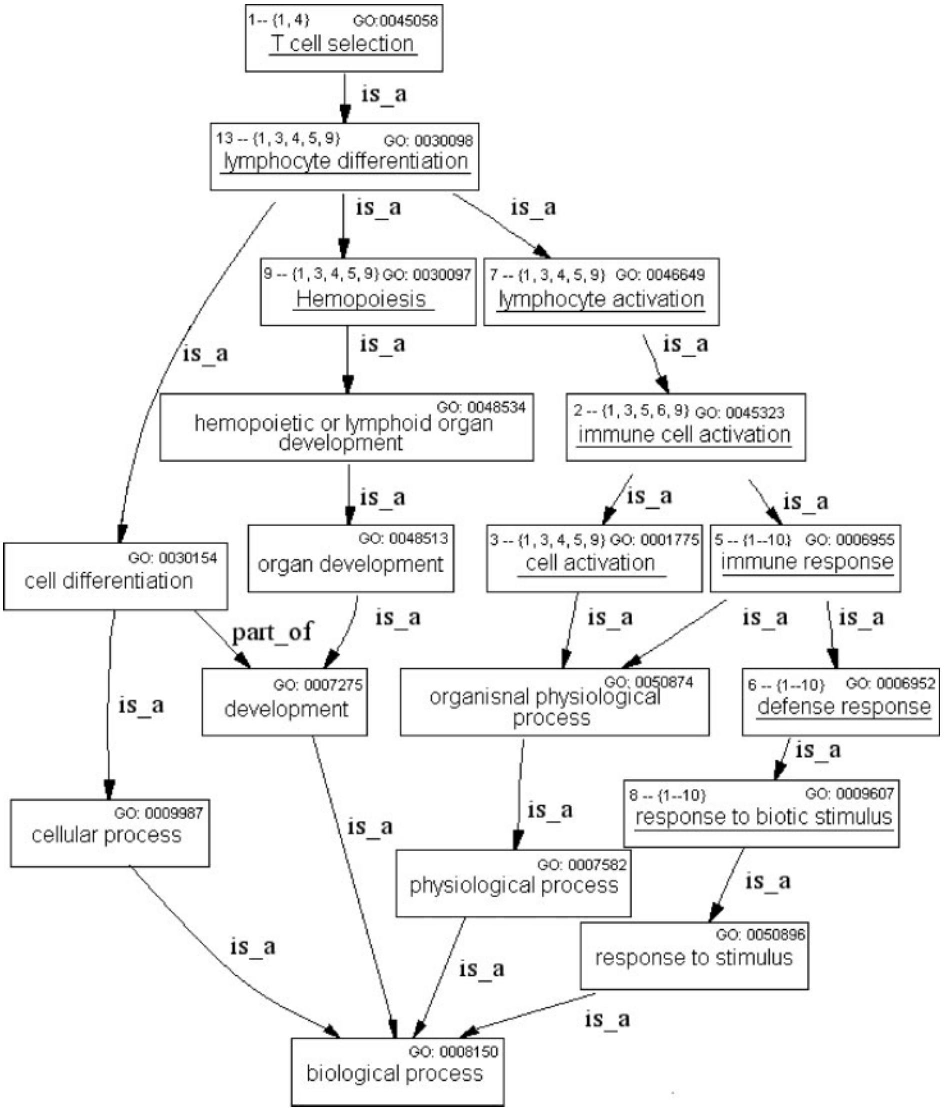


Fig. 5. GO graph of findings by DEA-PLS for Chiaretti's data. Specific terms are at the top, while general terms are at the bottom. Top 10 significant GO terms are underlined, with rank shown in upper left of rectangle. Ranks for top 10 probe sets are shown in curly brackets.

the literature that the *t* distribution is not a valid null distribution for testing gene expression data. Our ongoing research also shows that even when the normality assumption holds for the gene expression data, with a binary response the first score from PLS procedure is not normally distributed. This is partially the reason that we have excessive number of important GO terms with a biased null distribution. In the article, we suggest to only focus on the most significant GO terms. We are currently working on adjusting the test according to the distributional information of the first score from PLS. An alternative approach is to use nonparametric testing instead of the *t*-test. Testing procedures such as the significance analysis of microarray (SAM) method by Tusher *et al.* (2001), empirical Bayes (EB) method of Efron *et al.* (2001) and the mixture model method (MMM) of Pan *et al.* (2001) are possibilities. These methods can potentially improve the performance of the testing part of our procedure. We are currently working on finding the right testing procedure to maximize the effectiveness of DEA methods.

One key aspect of DEA methods is that they are adaptive to many other analysis techniques. As we mentioned, after combining the variables by GO terms using PLS, we basically have a new set of data with fewer but more meaningful variables. We can either build a predictive model with GO terms or improve our testing procedure with many existing methods for microarray data. We can also adapt DEA to other techniques, such as the one proposed by Alexa *et al.* (2006), where they suggest improved scoring of GO terms by decorrelating GO graph structures. With much room to expand and improve, DEA is a promising new method for providing accurate and interpretable analysis of microarray data.

ACKNOWLEDGEMENTS

We gratefully acknowledge many hours of feedback and support offered by Drs Gary Merrill, Guanghui Hu, Lei Zhu and Kwan Lee, all of GlaxoSmithKline. The work of J.H.-O. was supported by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 1 P20 HG003900-01. We also thank the reviewers for their careful evaluation and excellent suggestions. Funding to pay the Open Access publication charges was provided by GlaxoSmithKline, Inc.

Conflict of Interest: none declared.

REFERENCES

Al-Shahrour,F. *et al*. (2003) Fatigo: a web tool for finding significant association of gene ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Alexa,A. *et al*. (2006) Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, **22**, 1600–1607.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Barry,W.T. *et al*. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.

Bastien,P. *et al*. (2005) Pls generalised linear regression. *Comput. Stat. Data Anal*., **48**, 17–46.

Beissbarth,T. and Speed,T.P. (2004) Gostat: findstatistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Benjamini,Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., B*, **57**, 289–300.

Berriz,G.F. *et al*. (2003) Characterizing gene sets with funcassociate. *Bioinformatics*, **19**, 2502–2504.

Castillo-Davis,C. and Hartl,D.L. (2003) Genemerge–post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, **19**, 891–892.

Chiaretti,S. *et al*. (2004) Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, **103**, 2771–2778.

Dahlquist,K.D. *et al*. (2002) Genemapp, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet*., **31**, 19–20.

Dalla-Favera,R. (2001) Microarray analysis of b cell chronic leukemia. *Program and Abstracts of the FASEB 2001 Conference on Hematological Malignancies*.

de Jong,S. (1993) Simpls: an alternative approach to partial least squares regression. *Chemom. Intell. Lab Syst*., 18: 251–263.

Ding,B. and Gentleman,R. (2005) Classification using generalized partial least squares. *J. Comput. Graph. Stat*., **14**, 280–298.

Draghici,S. *et al*. (2003a) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Draghici,S. *et al*. (2003b) Onto-tools, the toolkit of the modern biologist: Onto-express, onto-compare, onto-design and onto-translate. *Nucleic Acids Res*., **31**, 3775–3781.

Efron,B. *et al*. (2001) Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc*., **96**, 1151–1160.

Fort,G. and Lambert-Lacroix,S. (2005) Classification using partial least squares with penalized logistic regression. *Bioinformatics*, **21**, 1104–1111.

Goeman,J. *et al*. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hoskuldson,A. (1988) Pls regression methods. *J. Chemom*., **2**, 211–228.

Huang,X. and Pan,W. (2003) Linear regression and two-class classification with gene expression data. *Bioinformatics*, **19**, 2072–2078.

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.

Lee,H. *et al*. (2005) Erminej: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, **6**, 269.

Li,L. and Li,H. (2004) Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, **20**, 3406–3412.

Look,A. (2001) Molecular pathogenesis of t-cell acute lymphoblastic leukemia. *Program and Abstracts of the FASEB 2001 Conference on Hematological Malignancies*.

Malthouse,E. *et al*. (1997) Nonlinear partial least squares. *Comput. Chem. Eng*., **21**, 875–890.

Man,M.Z. *et al*. (2000) Power sage: comparing statistical tests for sage experiments. *Bioinformatics*, **16**, 953–955.

Marx,B. (1996) Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, **38**, 374–381.

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

Nam,D. *et al*. (2006) Adgo: analysis of differentially expressed gene sets using composite go annotation. *Bioinformatics*, **22**, 2249–2253.

Nguyen,D.V. and Rocke,D.M. (2002a) Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, **18**, 1216–1226.

Nguyen,D.V. and Rocke.,D.M. (2002b) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39–50.

Nguyen,D.V. and Rocke.,D.M. (2004) On partial least squares dimentsion reduction from microarray-based classification: a simulation study. *Comput. Stat. Data Anal*., **46**, 407–425.

Novak,K. (2001) Conference report. *FASEB 2001 Conference on Hemotological Malignancies, Medscape General Medicine*, **3**.

Pan,W. *et al*. (2001) A mixture model approach to detecting differentially expressed genes with microarray data. *Research Report 2001-011*. Division of Biostatistics, University of Minnesota.

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.

Tusher,V.G. *et al*. (2001) Signficance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

van der Voet,H. (1994) Comparing the predictive accuracy of models using a simple randomization test. *Chemom Intell Lab Syst*., **25**, 313–323.

Zeeberg,B.R. *et al*. (2003) Gominer: a resource for biological interpretation of genomic and proteomic data. *Bioinformatics*, **4**, R28.

Zhong,S. *et al*. (2003) Chipinfo: software for extracting gene annotation and gene ontology information for microarray analysis. *Nucleic Acids Res*., **31**, 3483–3486.