Genomics 92 (2008) 373–383

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Direct functional assessment of the composite phenotype through multivariate projection strategies

Ana Conesa [a,*], Rasmus Bro [b], Francisco García-García [a], José Manuel Prats [c], Stefan Götz [a,d], Karin Kjeldahl [b], David Montaner [a], Joaquín Dopazo [a,d,e]

[a] Bioinformatics Department, Centro de Investigacin Principe Felipe, Valencia, Spain
[b] Department of Dairy and Food Science, Faculty of Life Sciences, Copenhagen University, Denmark
[c] Department of Applied Statistics, Technical University of Valencia, Valencia, Spain
[d] Center for Biomedical Research on Rare Diseases (CIBERER)
[e] Functional Genomics Node (National Institute for Bioinformatics, INB), Valencia, Spain

ARTICLE INFO

Article history:
Received 22 February 2008
Accepted 28 May 2008
Available online 13 September 2008

Keywords:
Data integration
Functional genomics
Multivariate regression
Gene ontology
Phenotype
Gene annotation
Partial least squares
Principal component analysis

ABSTRACT

We present a novel approach for the analysis of transcriptomics data that integrates functional annotation of gene sets with expression values in a multivariate fashion, and directly assesses the relation of functional features to a multivariate space of response phenotypical variables. Multivariate projection methods are used to obtain new correlated variables for a set of genes that share a given function. These new functional variables are then related to the response variables of interest. The analysis of the principal directions of the multivariate regression allows for the identification of gene function features correlated with the phenotype. Two different transcriptomics studies are used to illustrate the statistical and interpretative aspects of the methodology. We demonstrate the superiority of the proposed method over equivalent approaches.

Introduction

Gene expression profiling is used to study the gene regulatory basis of phenotypic or developmental characteristics. Statistical analysis of transcriptomics data is normally addressed through a two-step process: first, a statistical test is performed to derive a *P* value for the association of individual gene expression values to the phenotype or experimental condition(s), and a number of "significant genes" are selected on the basis of an arbitrary *P* value threshold. Most commonly used methods apply modifications of the *t* statistics or ANOVA to generate hypothesis testing of differential expression [1–4]. Secondly, selected genes are further analyzed to identify their relevant association to cellular functionalities [5,6]. Fisher's exact test, the Kolmogorov-Smirnov test, or the chi-squared are common statistics to identify functional classes with a significant enrichment within the pool of differentially expressed genes [7]. This widely used approach presents a number of drawbacks. On one hand, the univariate nature of the by-gene statistical assessments implies that any informative correlation pattern within gene expression will be ignored. On the other hand, strong *P* value corrections need to be applied to deal with the concomitant multiple testing scenarios and this can seriously hamper the identification of significant features on large datasets [8]. Furthermore, as functional assessments—which paradoxically have their foundation on the correlated nature of gene activity—are performed after univariate gene selection, results are dependent on the *P* value cutoff of choice, which can be problematic. Thus, too strict *P* value thresholds may lead to univariately nonsignificant genes (that are in fact significant in the multivariate space, but remain undetected) while too permisive cutoffs may result in multivariate important features getting lost among irrelevant information. Finally, when the target phenotype is not composed by a single variable but a space of different measurements (e.g., age, gender, different clinical parameters), the evaluation of differential expression under a univariate strategy can imply multiple and difficult assessments.

Multivariate approaches to gene expression analysis try to overcome these limitations. Principal component analysis (PCA), factor analysis, and multiple correspondence analysis are multivariate space reduction methodologies that exploit the correlation structure in the data to identify relevant patterns of variation [9,10]. These approaches have been applied to the analysis of transcriptomics data and have showed their potential in capturing relevant associations in the multivariate expression space that would escape to univariate analysis [11–13]. Several authors have proposed different strategies for deriving gene-associated

* Corresponding author. Fax: +34 96 328 97 01.
E-mail address: aconesa@cipf.es (A. Conesa).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.05.015

significance values of differential expression in this multivariate analysis context. Lu et al. [13] used the Hotelling T^2—a multivariate extension of the univariate *t* statistics—to select significant genes, and proposed a recursive method to deal with the singular data structures that appear when the number of variables greatly exceeds the number of observations. Landgrebe et al. [14] applied ANOVA on PCA results to identify components with a significant difference between experimental groups and applied a VIP-like statistics to evaluate component significance. Nueda et al. [15] employed the gene leverage along with a permutation test to find significant contributions to the multivariate projection. In all these examples data analysis focuses exclusively on expression values and does not incorporate a priori knowledge. Approaches that consider the functional role of genes while trying to capture the cooperative acting of the set of genes as a whole are, e.g., the so-called gene set methods, such as the GSEA [16] and FatiScan [17]. In these methodologies, genes are ranked according to a measure of differential expression and the enrichment of functional classes towards the extremes of this ranking, rather than a single group of genes, is tested. These methods have proven

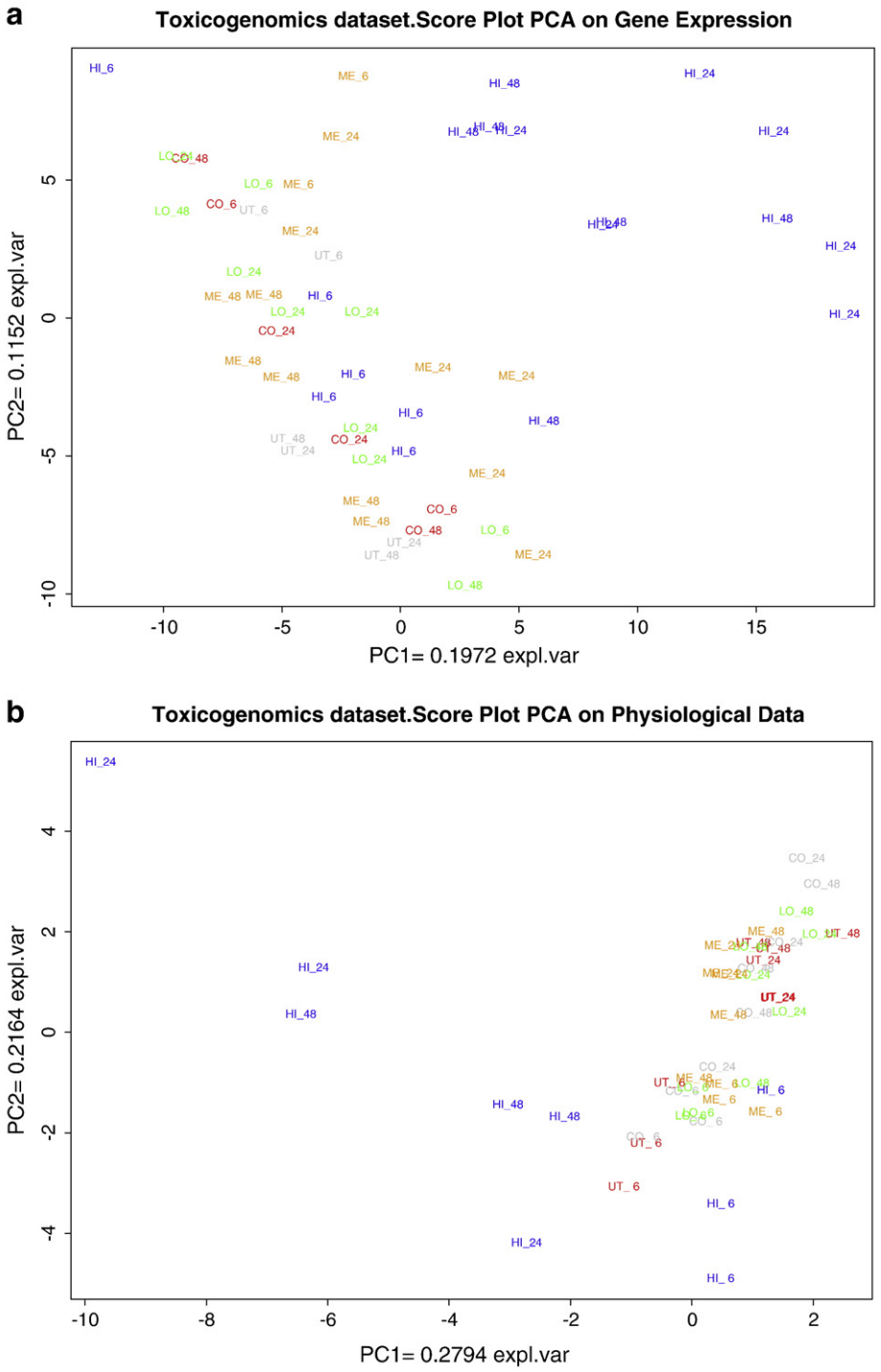

Fig. 1. PCA analysis of toxicogenomics data. Samples are labeled by the treatment group: HI, high bromobenzene dose; ME, medium bromobenzene dose; LO, low bromobenzene dose; CO, placebo; UT, untreated control. _6, _24 and _48 denote hours of administration. Closeness in the projected space indicates similarity between samples. (a) PCA score plot with gene expression data. (b) PCA score plot on physiological variables. (c) PCA score plot on functional variables.

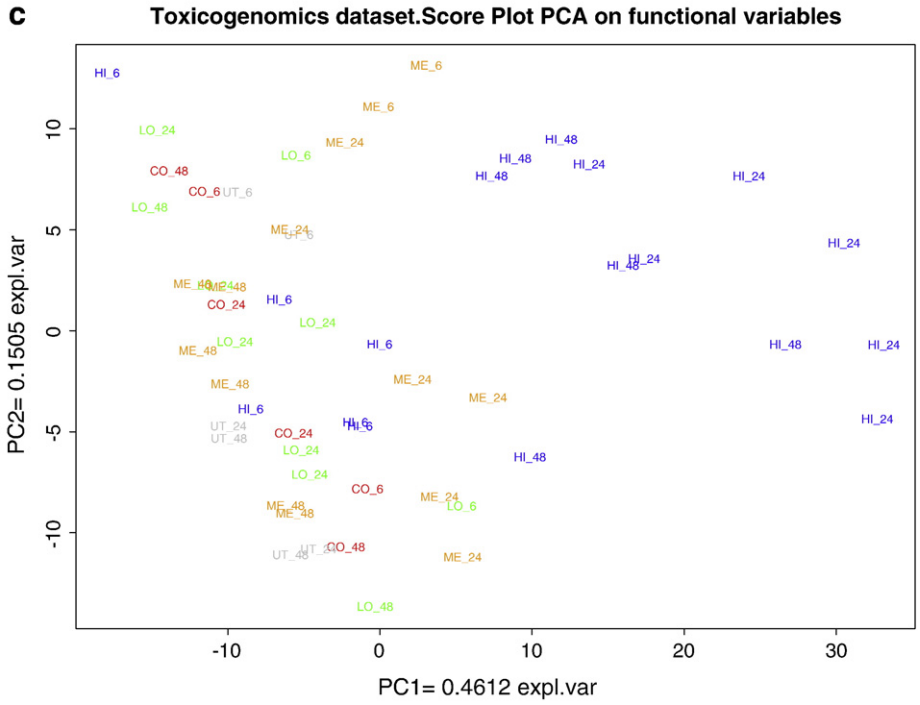

Fig. 1 (*continued*).

to be very sensitive as they do not require genes to be significantly declared, but still follow a prior univariate test to generate the gene rank which, on the other hand, can normally only be derived for two-class problems. Recent work has tried to combine multivariate statistics with functional assessment to overcome some of the limitations of the gene set methods. Kong et al. [18] proposed a methodology in which functional expression submatrices—i.e., the set of gene sharing the same function—were used to create projected functional subspaces which subsequently were tested by Hotelling T^2 statistics for their ability to separate between treatments. The same basic idea, but with a different statistical approach, is followed by Nettleton et al. [19]. These authors apply the multiresponse permutation procedure (MRPP) [20] to test against the null hypothesis of invariant distribution of gene expression among different treatment classes. Different in their purpose are the methods which combine multivariate expression with gene function to identify significant co-occurrences of functional classes [21,22] or predict gene function from transcriptomics data [23,24].

Still, all these methodologies deal with a single response variable—at two or more levels—if any, to evaluate significant functional associations. Very little has been done in the multivariate analysis of complex phenotypic determinations that follow transcriptomics profiling. Such data are of special relevance in biomedical research where composite clinical frames need to be understood in the light of gene activity [25,26].

In this work we present a novel approach for the analysis of transcriptomics data that integrates functional annotation and expression values in a multivariate fashion and directly assesses the relation of functional features to a multivariate space of response variables. Our method benefits from the correlation patterns between both gene and gene functions to identify a functional signature that best predicts the phenotypic outcome.

Results

Troxicogenomics dataset

PCA analysis of both gene expression data (Fig. 1a) and phenotypic variables (Fig. 1b) revealed a first component of variability that basically differentiates the high bromobenzene dose treatment at 24 and 48 h from the rest of the conditions. A pretty similar PCA score plot was found for gene expression and phenotypic variables, indicating that the major pattern of the variability in both datasets had similar structures and related to the intensive administration of the drug.

The initial GO term filter procedure generated a total of 1140 GO terms from the three main GO branches. After PCA-based transformations 823 functional components were created with an average explained variance of 40%. In most cases GO terms were represented by one functional variable and only in a few cases up to two variables were derived by a single functional class. Table 1 summarizes the results of the analysis procedure.

PCA analysis of the new matrix of functional variables showed a projected space similar to that obtained previously with gene expression and clinical data (Fig. 1c), but explained variance for the first and most discriminating component was clearly higher (46% with functional variables versus 20% with gene expression data), indicating a more compact signal in the transformed data. PLS (partial least squares) regression was then applied to relate the measured physiological parameters to the new space of functional variables. The number of components was selected by leave-one-out cross-validation, resulting in a 7-component model with maximal overall predicting value. Analysis of the R^2, Q^2, and VIP parameters for individual physiological variables permitted the identification of differences in relation to the functional data (Table 2). Variables such as ASAT, bilirubin, LDH, and ALAT showed important contributions (high VIP) and were well predicted by the model (high Q^2), which means that these parameters are highly related to the gene functional response triggered by the toxic compound. Other variables such as total protein or albumin were low contributors to the model and failed to be predicted, indicating their poor association to the gene expression pattern revealed in the analysis. These variables were removed from the final PLS model which obtained an average prediction error of 0.58 and a determination coefficient R^2 of 0.75, both highly significant (P value ~ 0). It is worth noting that R^2 values were in general high, also for poorly predicted variables, which

Table 1
Quantitive figures in the analysis procedure of the toxicogenomics and breast cancer datasets

	Origina data		PCA transformation to funcional data				PLS model			
	Probes	Annotated GO terms	GO term selection	Functional variables	Mean expl. var.	Mean GO level	No. comp.	Average R^2	Average Q^2	Significant funct. vars.
Toxicogenomics	2665	7411	1140	823	0.4	6.7	7	0.75	0.58	50
Breast cancer	22283	10940	3129	1901	0.44	6.2	4	0.45	0.3	65

illustrates the optimistic nature of this parameter to evaluate the validity of a regression model.

Graphical analysis of the PLS model discovers further aspects of the data. The score plots associated with the functional and physiological data matrices showed a stronger differentiation between the high bromobenzene doses at 24 and 48 h and the remaining samples (Figs. 2a and 2b), indicating that the physiological response to the gene expression effects of toxic compound is mainly concentrated under these conditions. Additionally, the **Y** biplot of the PLS model (Fig. 2c) revealed a positive correlation (same orientation in the projected space) of these high toxicity levels with the most responsive cellular compounds, while other parameters, such as glucose and kidney weight, showed a negative relationship. In fact, increased levels of ASAT, bilirubin, LDH, ALAT, and phospholipids have been associated with the response to xenobiotics and are considered as markers of toxicity [27], while plasma glucose concentrations tend to decrease for the imbalance in energy requirements [27].

Selection of significant functional variables in the PLS model was done by resampling methods. Fifty functional classes were selected at a P value < 0.05. Functional variables are represented in the **Y** biplot by open dots (Fig. 2c). Significant variables are depicted colored. Significant GO terms include *response to stimulus*, *heme binding*, *fatty acid metabolic process*, *oxidoreductase activity*, *glutathione transferase*, *apoptosis*, *ribosomal unit*, and *cytoskeleton* (see Supplemental Material T1 for a complete list). Fig. 3 shows the DAG of the significant functional terms corresponding to the Biological Process GO branch. Significant functional categories extensively explain the cellular adaptative response to drug administration which includes conjugation to glutathione by glutathione transferase, modification in oxidative, heme containing enzymes, activation of the ribosome machinery for protein synthesis, and cytoskeleton reorganization [27].

Table 2
PLS model parameters for the **Y** data structure (physiological variables) of the toxicogenomics dataset

Physiological variable	R^2	Q^2	VIP
ASAT	0.94	0.81	5.54
Bilirubin.tot	0.89	0.67	5.02
LDH	0.92	0.67	5.66
ALAT	0.90	0.66	5.27
Phospholipids	0.81	0.61	4.16
Liver.BW	0.79	0.51	3.23
Liver	0.72	0.48	2.38
Body.Weight	0.59	0.43	2.14
Glucose	0.58	0.43	2.69
Creatin	0.72	0.43	4.41
Kidney.BW	0.61	0.39	2.76
GSH.corr	0.57	0.37	3.03
Triglycerides	0.63	0.36	1.77
Cholesterol	0.67	0.35	3.64
Urea	0.54	0.27	1.97
ALP	0.72	0.27	3.05
AG.ratio	0.59	0.15	3.05
Tot.Protein	0.53	−0.03	2.85
Kidneys.weight	0.20	−0.09	0.78
Albumin	0.38	−0.12	1.98

Breast cancer dataset

The breast cancer dataset contained nearly 10 times the number of probes of the toxicogenomics dataset. Still data transformation by PCA on functional classes rendered a not very different compression result: 1901 functional variables with an averaged explained variance of 44% (Table 1). PCA and PLS score plots of both gene expression and functional data showed a different distribution of p53+ and p53- samples along the first component, which was more pronounced when samples were labeled by their ER status (Figs. 4a and 4b). This is in agreement with observations in the original work on the incompleteness of the p53 sequence determinations to establish the p53 deficiency status in breast tumors [28]. The PLS analysis with functional variables resulted in a 3-component model with significant Q^2 and R^2 parameters. Again, only a subset of clinical variables was well predicted by the model, namely the p53seq, ER status, histological grade, and PgR status for which the mean R^2 and Q^2 were 0.45 and 0.30, respectively (Supplemental Table T2). Furthermore, the **Y** loading plot of the PLS model showed a negative correlation between the p53 genotype and the ER status and histological grade (Fig. 5), which has been described in previous reports [29]. Sixty-five significant functional variables were detected by resampling. The corresponding Gene Ontology terms pointed to functions related to the immune response, cell division and proliferation, cytoskeleton organization, estrogen receptor signaling—already highlighted by Miller and co-workers [28]—and also to novel functional activities such as *activation of JNK activity*, *fiber development*, and *chemokine activity*. The complete list of significant functional terms in the breast cancer study is provided as supplemental material T1.

Comparison with other functional assessment methods

We compared the functional class results in the toxicogenomics and breast cancer examples, respectively, to two traditional univariate pathway analysis methods, namely the enrichment analysis by the Fisher exact test [30] and the Gene Set Enrichment Analysis provided by the FatiScan [17]. Additionally, we compared our results in both data examples to those obtained with the multivariate approach proposed by Kong et al. [18] where the Hotelling T^2 statistics is used to find treatment-associated significant differences between functional class-defined gene expression submatrices. GO term comparisons were done using the Blast2GO software [31]. In contrast to our strategy, all comparing methodologies required the selection of two contrasting conditions—HI bromobenze treatment vs control in the toxicogenomics example, and p53seq label for the breast cancer study—to define the analysis. In both study cases, traditional univariate approaches provided a reduced and semantically less rich, i.e., consisting of more general terms, set of significant functional classes (see Supplemental Table T1). On the contrary, the Hotelling T^2 method by Kong et al. consistently generated a far too large selection of GO classes (256 and 1520 GO terms for toxicogenomics and breast cancer datasets, respectively) which included most of the functions detected as significant by our method and many others suspiciously false positives, such as neural activity-associated processes in the case of the toxicogenomics liver samples and eye and bone specific functions in the case of the breast cancer data. Detailed information in functional results is provided in Supplemental Table T1.

Discussion

The proposed method integrates in one analysis three basic elements of transcriptomics studies: gene expression data, functional annotation, and phenotype characteristics, providing a direct relationships between gene function and response variables. The integrative analysis of transcriptomics data has been the subject of recent statistical developments [18,32–34]. Our method differs from other approaches in that it translates gene expression to a distinct expression signature of the functional class. By applying PCA on gene sets that share a function, the major expression patterns associated to the functional class can be identified and used as novel variables to study the phenotype. With this approach, two potentially critical problems can be overcome based on the assumption that important genes are correlated to similar genes. First of all, the unimportant genes are dramatically reduced in numbers which can be decisive to be able to detect significant variations. Secondly, the important (as well as unimportant) variation is expressed in a reduced form by scores from principal component analysis. Hence, ideally, each phenomenon appears only once and therefore has a better chance of influencing the further analysis. A key element to achieve this is the criterion for selecting Gene Ontology terms and functional components. We applied a simple filtering procedure on the set of initial GO terms to avoid annotation redundancies that arise from the hierarchical structure of the Gene Ontology. In this way candidate GO terms are guaranteed to collect at least partially different annotation sets. More important even is the criterion for selecting functional components. Component selection in dimension reduction approaches are habitually based on cross-validation or scree-plot analysis [35]. These procedures consist of building and evaluating different models by leaving out one or more observations that are then predicted by the model built, or using as many components as needed to reach a given amount of explained variance. In our case, the purpose of component selection is to identify a relevant expression features of the functional class, rather than to test prediction ability or sufficiently explain the functional submatrix. Therefore we choose a criterion that would select functional variables when they collect an amount of variance above what could be considered random noise. The effect is an important reduction in functional classes from the original GO set and a selection of terms of medium hierarchy depth level (mean value around 6.5) with a sufficient explanatory capacity (~40% on average).

Compared with common univariate statistical approaches for the assessment of gene functional enrichments [5,16,17], the method proposed in this work differentiates for its consideration of the coordinative behavior between functional classes—not only within—and therefore potentially capturing the cooperative activity of functional processes. This implies that covariance between genes is particularly stressed in our approach, since a functional class of differentially expressed genes but not correlated gene members might not be detected by our method but could be identified by a univariate strategy. Compared to another published multivariate method, our approach seems to achieve a good trade-off between sensitivity and selectivity in the selection of significant functional classes. We postulate that the two-step strategy of our method—creation of functional variables followed by PLS inference—and significance criterion based on the distribution of VIP values of the randomized PLS models are key for obtaining a sensible selection of functional variables. The simple randomization of expression values in functional submatrices would tend to create in too compacted Hotelling T^2 null distributions that would declare as significant an excessive number of variables in the Kong et al. method.

Furthermore an additional aspect in our approach is that the analysis is not restricted to pairwise comparisons between conditions, but it can evaluate the composite phenotype dynamically and for the relationships within outcome parameters. This last consideration of multiple phenotypic characterizations in microarray datasets was likewise addressed by Fellenberg et al. [36]. In this work, Correspondence Analysis was used to study relationships between transcriptomics data and extensive sample annotations. The authors developed an interesting method to extract relevant phenotypic characteristics and map them to gene expression features by multivariate projection methods. However, this work does not incorporate the gene functional information which can provide a more interpretable result, in terms of biological processes, to the relationship phenotype-transcriptome, and also does not exploit an inferential relationship, such as PLS does, to achieve an optimized projection of the gene expression and the phenotypic spaces.

All together, our results indicated that the proposed method is effective in extracting informative functional signatures that differentially correlate with diverse aspects of the phenotype. We believe that this approach will be of great help in the study of the molecular mechanisms behind the observed characteristics of organisms and to unravel genotype-phenotype functional relationships.

Material and methods

The proposed method

Schematically, our proposal uses multivariate projection methods to obtain new correlated variables for gene sets which share a given function. These new "functional variables" are then used to perform a multivariate regression on a set of response variables. The analysis of the principal directions of the multivariate regression allows for the identification of gene function features correlated with the phenotype. The proposed method consists of the following steps:

1. Find the functional annotation of the genes in the transcriptomics dataset. For each functional term, create a "subexpression matrix" with all associated genes.
2. Perform principal component analysis in each of the new expression matrices and select a number of components that collect nonrandom variation.
3. Collect the PCA scores of the selected components into a new matrix of "functional variables." These functional variables represent coordinative expression patterns of genes associated by a functional label.
4. Use this new matrix to perform partial least square (PLS) regression on the response variables.
5. Select significant functional variables in the PLS by bootstrap.

In principle, any functional vocabulary can be used to elaborate functional variables. In this work we have taken the Gene Ontology scheme (http://www.geneontology.org) as it is the most extensive vocabulary for the description of gene function. We considered all terms present in the Direct Acyclic Graph (DAG) encompassed by the gene collection of the transcriptomics datasets but removing all annotation redundant terms. A term is considered annotation redundant within a given gene collection if it has a child term with identical gene annotation set. For example, if GO:0006915 (*apoptosis*) has 15 annotated genes and parent term GO:0012501 (*programmed cell death*) back-inherits these and only these 15 genes, then *programmed cell death* is considered annotation redundant and removed from the initial set of functional classes.

Principal component analysis projects a data matrix into a space of lower dimension while keeping most of the variability in the original data [9].

The PCA model for each functional class can be expressed in matrix notation as

$$\mathbf{X} = \mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{R},$$

where $\mathbf{A}$ ($\mathbf{I}\times\mathbf{F}$) is the matrix collecting the F functional variables, $\mathbf{B}$ ($\mathbf{J}\times\mathbf{F}$) is the loading matrix that indicates the importance of each gene on each functional variable, and $\mathbf{R}$ is the residual matrix. Dimension reduction is possible when there exists a correlation structure in the dataset, i.e., where there is a sufficient number of genes with

correlated expressions. In this sense, PCA can be considered as a summarizing method and in our approach the profile given by the observations scores of each principal component reflects a coordinated behavior of a group of genes within the functional class and defines the so-called functional variables. Selection of functional variables is done based on the amount of variance explained by the corresponding component, normalized by the number of genes in the functional class. Components—i.e., functional variables—are selected in this case if their normalized variance is greater than the average gene variance of the complete dataset.

The relationship between functional variables and phenotypic variables is analyzed by partial least squares [10]. PLS is a dimension reduction regression approach which finds a projected space that maximizes the correlation between independent and dependent data

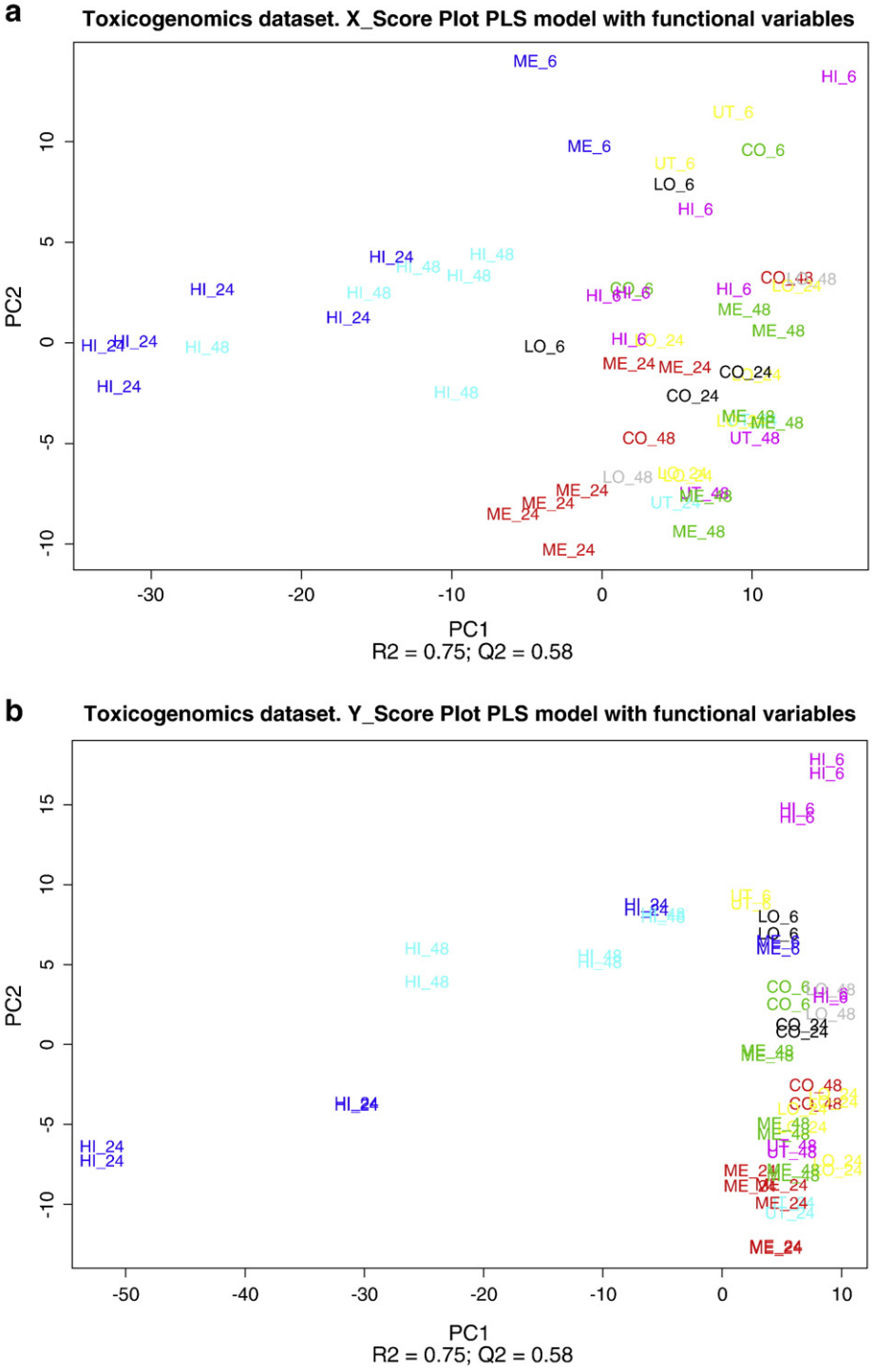

Fig. 2. PLS analysis of toxicogenomics data. Samples are labeled by the treatment group: HI, high bromobenzene dose; ME, medium bromobenzene dose; LO, low bromobenzene dose; CO, placebo; UT, untreated control. _6, _24 and _48 denote hours of administration. (a) **X**_score plot showing the relationships between treatments according to the dimension reduction of the functional data. (b) **Y**_score plot showing the relationships between treatments according to the dimension reduction of the physiological variables. (c) **Y**_ biplot shows the projection of both functional and physiological variables. Variables poorly explained by the model are given in gray. Functional variables are represented by dots and colored when significant.

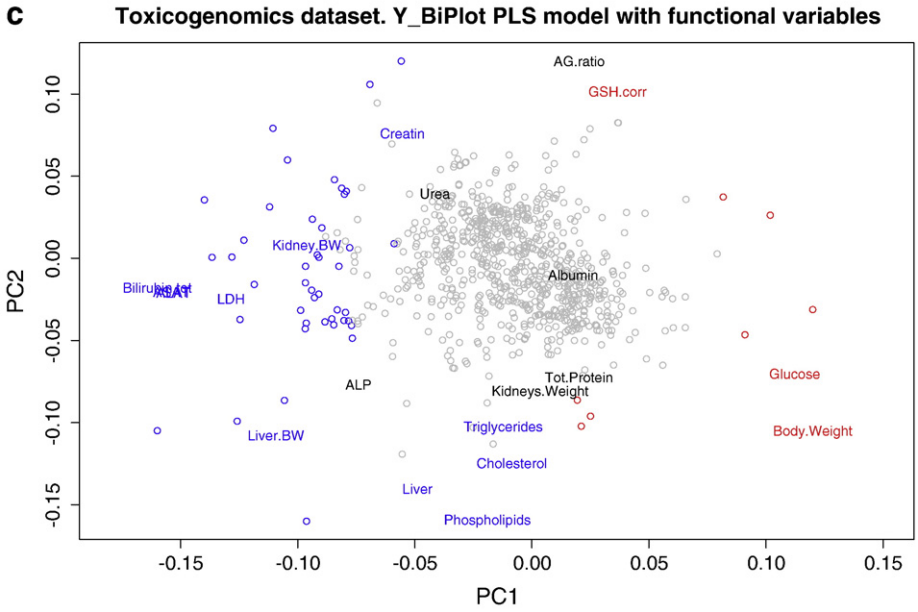

Fig. 2 (*continued*).

structures, as well as the explained variability within both data matrices.

The PLS models the data through the use of the following expressions,

$$\mathbf{T} = \mathbf{X}\mathbf{W}^{*} = \mathbf{X}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}$$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{C}^{\mathrm{T}} + \mathbf{F}$$

where **X** and **Y** are the matrices of functional and physiological variables, respectively. **T** is the matrix that maximizes the covariance between **X** and **Y**, **P** the loading matrix for **X**, **C** the loading matrix for **Y**, **W** and **W*** are weighting matrices that indicate the importance of each functional variable in the new projected space, and **E** and **F** the residual matrices for **X** and **Y**, respectively.

Each component of the PLS model represents a pattern of variation that relates independent and dependent variables. Therefore, by analyzing the weights of functional and response variables in the PLS model we can identify gene function features that are associated with the observed phenotype. The significance of the PLS model is habitually given by the R^2 and Q^2 statistics, which indicate respectively the explanatory and predictive power of the model.

The R^2 is defined as the fraction of the total sum of squares which is captured by the model. For a model with F components,

$$\mathrm{R}^2 = \frac{\mathrm{SSM_F}}{\mathrm{SST}},$$

where SSM is the sum of squares of the model with F components and SST is the total sum of squares.

Q^2 parameter is given by

$$Q^2_{cum}(F) = 1 - \frac{\mathrm{PRESS}(F)}{\mathrm{SST}}$$

$$\mathrm{PRESS}_F = \sum_{t=1}^{I} r_i^2,$$

which indicates the sum of squares of the prediction errors "r" for the observations not included in the model during the cross-validation procedure.

Furthermore, the importance of each functional variable in the model can be computed by the VIP parameter, which is the sum of the contributions of the variable to the model components moderated by the weight of the component. This VIP parameter computes the influence on **Y** of every term x_k in the model, according to

$$\mathrm{VIP}_{FK} = \sqrt{\sum_{f=1}^{F} \left(w_{fk}^2 * (\mathrm{SSY}_{f-1} - \mathrm{SSY}_f) \right) * \frac{K}{\mathrm{SSY}_0 - \mathrm{SSY}_F}}.$$

Finally, we include a permutation test to determine the probability of the computed model parameters to occur by chance and to select significant functional terms. This permutation is performed on the original data matrix and therefore affects the steps of generation (PCA) and selection (PLS) of functional variables.

Datasets

We have applied the proposed method to two different datasets

The first dataset corresponds to a toxicogenomics study in which the effect of bromobenzene in liver toxicity in rats is analyzed [27]. In this experiment, rats are administrated the drug bromobenzene at three different doses (high, medium, and low) and liver/blood/urine samples are taken after 6, 24, and 48 h of treatment. There are control (no administration) and placebo (only drug vehicle administration) rat groups. For each experimental condition one to three rats were taken for gene expression profiling and microarray experiments were done with a dye-swap design on a custom cDNA microarray. Gene expression information is available for 2665 genes. Additionally, physiological and morphological determinations were conducted on the same rats, including body weight (g), kidneys weight (g), kidney/BW (g/kg), liver (g), liver/BW, bilirubin tot, ASAT, ALAT, LDH, albumin g/L, ALP (U/L), creatin umol/L, cholesterol (mmol/L), glucose (mmol/L), phospholipids (mmol/L), triglycerides (mmol/L), tot.protein (g/L), urea (mmol/L, A/G ratio, GSH corr. (M) [27].

The second dataset is a breast cancer study by Miller and co-workers [28]. This work explores the relationship between the p53 (TP53) pathway and breast tumor severity. The Affymetrix U133 A and B human GeneChips (~25,000 probes) were used to assess the genome-wide transcriptome profile of 251 primary invasive breast tumors for which detailed information on p53 status (p53+, mutant;

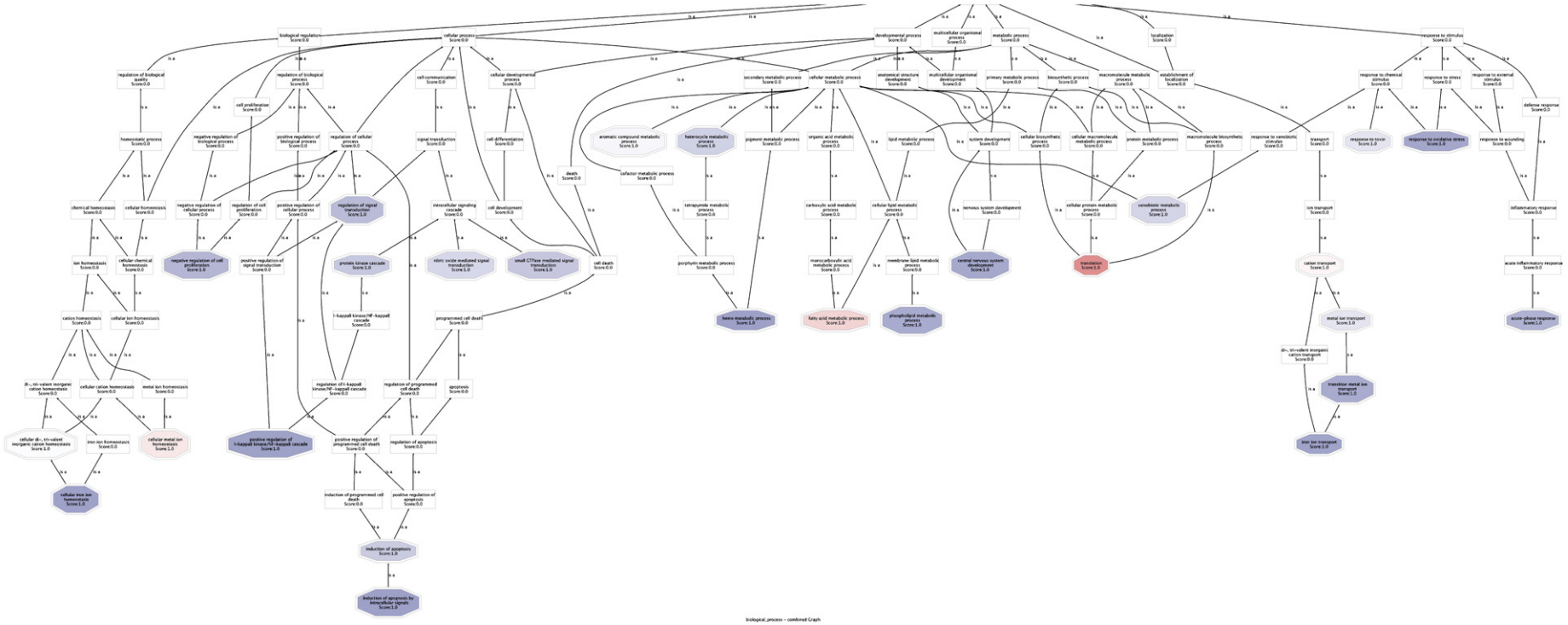

Fig. 3. Gene Ontology Direct Acyclic graph of the pool of Biological Process significant terms detected by the PLS model on functional variables. Term color intensity is proportional to the importance of the functional class in the PLS model. Hexagonal nodes are the actual selected GO terms. For a fuller view of this figure, please see Appendix A.

p53-, wt) was available. Additionally tumors were characterized by their estrogen-receptor (ER) status, Elston histological grade, PgR status, age at diagnosis, tumor size (mm), lymph node status, DSS TIME (disease-specific survival time in years), and DSS EVENT (disease-specific survival event; 1=death from breast cancer, 0=alive or censored).

These two datasets represent two analysis scenarios. The toxicogenomics dataset contains a multifactorial experimental design with strong gene expression signals associated to the treatments. A relative low number of genes and a wide array of response variables are present. The breast cancer dataset illustrates a typical cancer study with a large number of cases and a genome-wide transcriptomics profiling. A few clinical parameters were evaluated for each patient and gene expression signals are expected to be more diluted.

Data preprocessing and analysis

The toxicogenomics dataset was obtained directly from the authors, normalized by lowess, and centered genewise for each dye-swap pair as in [37]. Breast Cancer Affymetrix data were downloaded from the GEO database as global mean normalized data. Physiological/clinical variables were scaled in all cases and missing values were imputed by the *k*th nearest neighbors algorithm [38]. Gene Ontology functional annotations were obtained from public repositories. Annotated Gene Ontology DAG structures were generated with the Blast2GO software [31]. Noninformative reference distributions for the toxicogenomics and breast cancer dataset were generated by bootstrap. One thousand bootstrap runs were executed, in each case resampling both column- (samples) and row-wise (genes). Resampling by columns

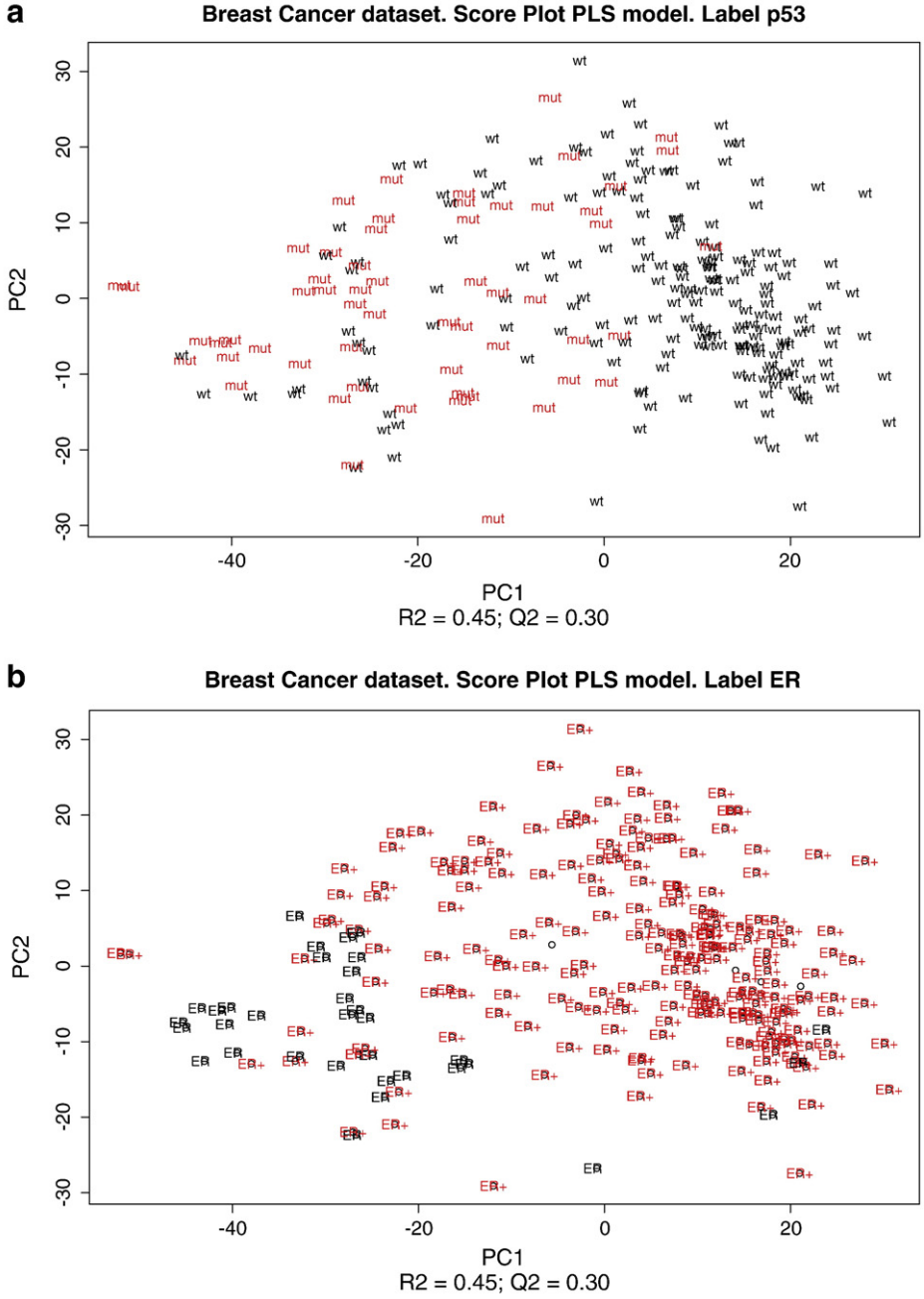

Fig. 4. X_score plot PLS model for breast cancer data. PLS model computed with functional variables. Tumor samples are labeled either for their p53 genotype (a) or ER status (b).

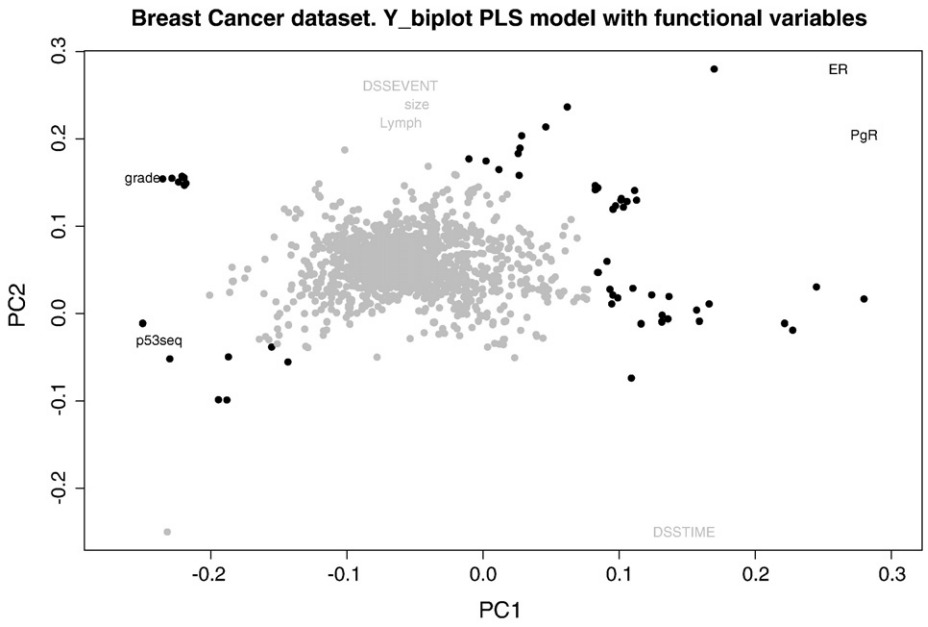

Fig. 5. Breast Cancer PLS **Y**_biplot. Projection of functional and clinical parameters into the first two components of the PLS model. Variables poorly explained by the model given in gray. Functional variables are represented by dots and colored dark when significant.

eliminates the relationship between the gene expression and the phenotype, while rearrangements by rows will destroy the coordinative structures within each functional class. The *P* value corresponding to the PLS model parameters (R^2 and Q^2) and the importance of functional variables (VIP) were computed as the frequency of occurrence of true data values in the respective reference null distributions. Significance threshold was set to 0.05.

All computations were performed in *R*, using limma [3], pls [39] and EMV packages. Scripts are available on request to the authors.

Acknowledgments

This work was funded by the Spanish Ramon y Cajal Program, grants from the Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER) ISCIII, and grant BIO2005-01078 from the Spanish Ministry of Education and Science. Publication costs were granted by the National Institute of Bioinformatics (www.inab.org) a platform of Genoma España.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.05.015.

References

[1] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. USA 98 (2001) 5116–5121.
[2] M.K. Kerr, M. Martin, G.A. Churchill, Analysis of variance for gene expression microarray data, J. Comput. Biol. 7 (2000) 819–837.
[3] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (2004) 3.
[4] T. Speed, Statistical Analysis of Gene Expression Microarray Data, Chapman and Hall/CRC, London, 2003.
[5] F. Al-Shahrour, R. Diaz-Uriarte, J. Dopazo, Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information, Bioinformatics 21 (2005) 2988–2993.
[6] J. Dopazo, Functional interpretation of microarray experiments, Omics 10 (2006) 398–410.
[7] I. Rivals, L. Personnaz, L. Taing, M.C. Potier, Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23 (2007) 401–407.
[8] S.B. Pounds, Estimation and control of multiple testing error rates for microarray studies, Brief Bioinform. 7 (2006) 25–36.
[9] J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.
[10] A. Smilde, R. Bro, P. Geladi, Multiway Analysis: Applications to the Chemical Sciences, Wiley, Chichester, England, 2004.
[11] R. Kustra, R. Shioda, M. Zhu, A factor analysis model for functional genomics, BMC Bioinform. 7 (2006) 216.
[12] P. Carmona-Saez, M. Chagoyen, F. Tirado F, J.M. Carazo, A. Pascual-Montano, GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists, Genome Biol. 8 (2007) R3.
[13] Y. Lu, P.Y. Liu, P. Xiao, H.W. Deng, Hotelling's T2 multivariate profiling for detecting differential expression in microarrays, Bioinformatics 21 (2005) 3105–3113.
[14] J. Landgrebe, W. Wolfgang, G. Welzl, Permutation-validated principal components analysis of microarray data, Genome Biol. 3 (2002) 19.1-19.1.
[15] M.J. Nueda, A. Conesa, J.A. Westerhuis, H.C. Hoefsloot, A.K. Smilde, M. Taln, A. Ferrer, Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA, Bioinformatics 23 (2007) 1792–1800.
[16] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. USA 102 (2005) 15545–15550.
[17] F. Al-Shahrour, L. Arbiza, H. Dopazo, J. Huerta-Cepas, P. Minguez, D. Montaner, J. Dopazo, From genes to functional classes in the study of biological systems, BMC Bioinform. 8 (2007) 114.
[18] S.W. Kong, W.T. Pu, P.J. Park, A multivariate approach for integrating genome-wide expression data and biological knowledge, Bioinformatics 22 (2006) 2373–2380.
[19] D. Nettleton, J. Recknor, J.M. Reecy. Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis, Bioinformatics 24 (2008) 192–201.
[20] P. Mielke Jr, K. Berry, Permutation methods: a distance function approach, Springer-Verlag, New York, 2001.
[21] H.K. Lee, A.K. Hsu, J. Sajdak, J. Qin, P. Pavlidis, Coexpression analysis of human genes across many microarray data sets, BMC Bioinformatics 25 (5) (2004) 18.
[22] D.J. Allocco, I.S. Kohane, A.J. Butte, Quantifying the relationship between co-expression, co-regulation and gene function, Genome Res. 14 (2004) 1085–1094.
[23] F. Luo, Y. Yang, J. Zhong, H. Gao, L. Khan, D.K. Thompson, J. Zhou, Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory, BMC Bioinform (8) (2007) 299.
[24] T.M. Murali, C.J. Wu, S. Kasif. The art of gene function prediction. Nat. Biotechnol. 24 (2006) 1474–1475.
[25] T.O. Khor, Toxicogenomics in drug discovery and drug development: potential applications and future challenges, Pharm. Res. 23 (2006) 1659–1664.
[26] R. Clarke, H.W. Ressom, A. Wang, J. Xuan, M.C. Liu, E.A. Gehan, Y. Wang, The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, Nat. Rev. Cancer 8 (2008) 37–49.
[27] W.H. Heijne, R.H. Stierum, M. Slijper, P.J. van Bladeren, B. van Ommen, Toxicogenomics of bromobenzene hepatotoxicity: a combined transcriptomics and proteomics approach, Biochem. Pharmacol. 65 (2003) 857–875.
[28] L.D. Miller, J. Smeds, J. George, V.B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E.T. Liu, J. Bergh, An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival, Proc. Natl. Acad. Sci. USA 102 (2005) 13550–13555.

[29] G. Cattoretti, F. Rilke, S. Andreola, L. D'Amato, D. Delia, P53 expression in breast cancer, Int. J. Cancer 41 (1988) 178–183.
[30] N. Blthgen, K. Brand, B. Cajavec, M. Swat, H. Herzel, D. Beule, Biological profiling of gene groups utilizing Gene Ontology, Genome Inform. 16 (2005) 106–115.
[31] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing Value estimation methods for DNA microarrays, Bioinformatics 17 (2001) 520–525.
[32] A. Fagan, A.C. Culhane, D.G. Higgins, A multivariate analysis approach to the integration of proteomic and gene expression data, Proteomics 7 (2007) 2162–2171.
[33] P. Wei, W. Pan, Incorporating gene networks into statistical tests for genomics data via spatially correlated mixture model, Bioinformatics 24 (2008) 404–411.
[34] S. Matsui, M. Ito, H. Nishiyama, H. Uno, H. Kotani, J. Watanabe, P. Guilford, A. Reeve, M. Fukushima, O. Ogawa, Genomic characterization of multiple clinical phenotypes of cancer using multivariate linear regression models, Bioinformatics 15 (2007) 732–738.
[35] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Multi- and Megavariate Data Analysis, UMETRICS AB, Umea, 2001.
[36] K. Fellenberg, C.H. Busold, O. Witt, A. Bauer, B. Beckmann, N.C. Hauser, M. Frohme, S. Winter, J. Dippon, J.D. Hoheisel. Systematic interpretation of microarray data using experiment annotations, BMC Genomics 7 (2006) 319.
[37] A. Conesa, M.J. Nueda, A. Ferrer, M. Talon, maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments, Bioinformatics 22 (2006) 1096–1102.
[38] A. Conesa, S. Gotz, J.M. Garcia-Gomez, J. Terol, M. Talon, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (2005) 3674–3676.
[39] B.H. Mevik, R. Wehrens, The pls Package: Principal Component and Partial Least Squares Regression in R, J. Stat. Software 18 (2007) 1–24.