Deposited research article

A non-parametric approach for identifying differentially expressed genes in factorial microarray experiments

Qihua Tan*, Jesper Dahlgaard*, Werner Vach†, Basem M Abdallah‡, Moustapha Kassem‡ and Torben A Kruse*

Addresses: *Department of Clinical Biochemistry and Genetics, Odense University Hospital, Denmark. †Department of Statistics, University of Southern Denmark, Denmark. ‡Department of Endocrinology, Odense University Hospital, Denmark.

Correspondence: Qihua Tan. E-mail: qihua.tan@ouh.fyns-amt.dk

Posted: 10 March 2005

Genome ***Biology*** 2005, **6:**P5

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/4/P5

Received: 7 March 2005

This is the first version of this article to be made available publicly. This article was submitted to *Genome* ***Biology*** for peer review.

A non-parametric approach for identifying differentially expressed genes in factorial microarray experiments

Qihua Tan[*], Jesper Dahlgaard[*], Werner Vach[†], Basem M. Abdallah[‡], Moustapha Kassem[‡], Torben A. Kruse[*]

[*] Department of Clinical Biochemistry and Genetics, Odense University Hospital, Denmark

[†] Department of Statistics, University of Southern Denmark, Denmark

[‡] Department of Endocrinology, Odense University Hospital, Denmark

Address for correspondence:

Dr. Qihua Tan

Department of Clinical Biochemistry and Genetics (KKA)

Odense University Hospital

Sdr. Boulevard 29

DK-5000 Odense C

Denmark

Tel. +45 65412822 Fax: +45 65411911 e-mail: qihua.tan@ouh.fyns-amt.dk

Abstract

We introduce a non-parametric approach using bootstrap-assisted correspondence analysis to identify and validate genes that are differentially expressed in factorial microarray experiments. Model comparison showed that although both parametric and non-parametric methods capture the different profiles in the data, our method is less inclined to false positive results due to dimension reduction in data analysis.

Background

As a high-throughput technique, microarray capable of simultaneously measuring mRNA levels for thousands of genes is becoming an increasingly important tool for researchers in biomedical science. At the same time, interpreting the large amount of data produced in microarray experiments imposes a major challenge to bioinformaticians [1]. Among the major issues in data analysis is the clustering of genes that are co-regulated in a biological process (for example cell cycle, treatment response, disease development) in high dimensional microarray experiments. Many clustering algorithms have been proposed to cluster genes using unsupervised [2-4] and supervised or knowledge-based [5,6] approaches.

In unsupervised gene clustering, the classes are unknown *a priori* and need to be discovered from the observed data. This is especially true for microarray studies using complex experiment designs due to the intricate relationships both between and within the multiple genetic and experiment factors including interactions which can't be predefined. Factorial experiment design (FED), characterized by simultaneous measurement of the effects of multiple experiment factors (the main effects) and the effects of interactions between the factors, is an economic yet efficient complex design popular in use in biomedical studies [7]. The nice features of FED have also made it well accepted in microarray experiments [8-11]. At the same time, statistical methods that take into account the experiment complexity are demanding for dealing with data produced in factorial microarray experiments. Kerr *et al*. [12] and Pavlidis [13] applied the analysis of variance model (ANOVA) to factorial microarray data using the parametric linear regression approach by assuming (1) normality in the log intensity of gene expressions and (2) linear relationship between log intensity and the effects of main experiment factors and their interactions. In their approaches, multiple replicates are required to insure model identifiability and then statistical procedures applied to correct the significance for multiple testing.

The singular value decomposition (SVD) [14] and SVD-based multivariate statistical methods, for example, principal components analysis [4,15] and correspondence analysis (CA) [16,17] have been applied in analyzing multidimensional microarray time-series data. Although such exploratory methods can be used for dimension reduction and for pattern discovery through data visualization, validity of the clusters or the selected genes has rarely been examined [18]. By bootstrapping the gene contributions on the reduced dimensions, we combine the resampling method with CA to identify the various gene expression profiles and to validate the significance of the differentially expressed genes in replicated factorial microarray experiments. We show in this paper, together with comparison with ANOVA, how an application of our methods to a microarray study on stem cells has helped us to find genes that are differentially regulated by the experiment factors and by their interactions. Additional applications of the methods in biomedical studies are suggested at the end of the discussion.

Methods

Correspondence analysis

As a multivariate data analyzing method, CA has been widely applied to process high-dimensional data in, for example, sociology, environmental science, and marketing research. Recently, the method has been applied to analyze microarray time-series data in cell cycle [16] and in diabetes research [17] to look for genes displaying distinct time-course expression profiles. In microarray experiments using a factorial design, we are actually facing a more sophisticated situation where we are interested not only in the effects of the multiple factors but also in the effects of interactions between them. Because FED represents a different complexity in high-dimensional microarray experiments, we apply the SVD-based CA to identify genes that are differentially regulated due to the experiment factors or as a result of their interactions. The idea is that main effects of the

multiple factors together with their interactions which dominate the variance in the data can be captured by the reduced dimensions in the newly transformed data space.

Suppose in a factorial microarray experiment, there are two experiment factors *A* and *B* with *p* levels in *A* and *q* levels in *B*. Then there will be *p*x*q* hybridizations each representing an interactive variable [19] or combination of experiment factors in the design. If, after gene filtering, we have a total of *n* genes, the data can be summarized by a large *n*x(*p*x*q*) matrix with *n* stands for the number of rows (genes) and *p*x*q* for the number of columns (hybridizations or interactive variables). To carry out CA, we divide each entry in the matrix by the total of the matrix so that the sum of all the entries in the resulted matrix equals 1. We denote the new matrix by P and its elements by p_{ijk} (i stands for the genes from 1 to *n*, j for the levels of factor *A* from 1 to *p* and likewise, *k* for the levels of factor *B* from 1 to *q*). In matrix *P*, the sum of row *i*, $p_{i.} = \sum_j \sum_k p_{ijk}$, is the mass of row *i* and the sum of the column representing the interactive variable A_jB_k, $p_{.jk} = \sum_i n_{ijk}$, is the mass of that column. With the row and column masses, we derive a new matrix *C* with elements $c_{ijk} = (p_{ijk} - p'_{ijk}) / \sqrt{p'_{ijk}}$ where $p'_{ijk} = p_{i.} p_{.jk}$ is the expected value for each element in matrix *P*. By submitting matrix *C* to SVD, we get $C = U\Lambda V'$ where *U* is the eigenvectors of CC', *V* is the eigenvectors of $C'C$, Λ is a diagonal matrix containing the ranked eigenvalues of *C*, λ_l (*l*=1, 2, ….*p*x*q*). Since the total inertia $\sum_l \lambda_l^2$ equals the sum of c_{ijk}^2 in *C*, the major variance in the original data is captured by the dimensions corresponding to the top elements in Λ.

One big advantage of CA is that, with the SVD results, we can simultaneous project genes and interactive variables into a new space with the projection of gene *i* on axis *l* calculated as $g_{il} = \lambda_l u_{il} / \sqrt{p_{i.}}$ where u_{il} is the *i*-th row and the *l*-th column in *U*, and similarly the projection of

A_jB_k along axis l is $h_{jkl} = \lambda_l v_{jkl} / \sqrt{p_{.jk}}$ where v_{jkl} is the element in the l-th column in V that corresponds to A_jB_k. In practice, a biplot [20] is used to display the projections. The biplot is very useful for visualizing and inspecting the relationships between and within the genes and the interactive variables. In the biplot, genes projected to a cluster of interactive variables associated with one experiment factor are up-regulated due to that factor. Especially, genes projected to a single or standing-alone interactive variable A_jB_k are highly expressed as a result of interaction between the experiment factors A and B. As the inertia along the l-th axis can be decomposed into components for each gene, i.e. $\lambda_l^2 = \sum_i p_{i.} g_{il}^2$, we can calculate the proportion of the inertia of the l-th axis explained by the i-th gene as, $ac_{il} = p_{i.} g_{il}^2 / \lambda_l^2$ which is the absolute contribution of the i-th gene to the l-th axis. The sum of ac_{il} for a group of selected genes stands for the proportion of the total variance explained by these particular genes. If all the n genes are randomly distributed along the axis, the null contribution (random mean) by each gene would be expected as $1/n$. The random mean contribution will be used for calculating the bootstrap p-values in the next section.

Non-parametric bootstrapping

Since the top dimensions of CA can represent effects of both the experiment factors and their interactions, our aim is to identify the genes that make significant contributes to the dimensions. Although, for each dimension, the gene contribution can be ranked, directly picking up the top rank genes ignores variability in each of the estimated contributions and is thus unreliable. The bootstrap technique was applied by Kerr and Churchill [21] to assess pattern reliability based on the estimated error distribution in their ANOVA models applied in replicated microarray time-course experiments. Ghosh [18] introduced the resampling method to SVD analysis of time-course data to bootstrap the variability of the modes that characterize the time-course patterns in microarray data. Here we combine the non-parametric bootstrapping with the correspondence analysis of factorial

microarray data to evaluate the significance of genes in their contributions to the leading dimensions that feature the effects of main factors as well as the effects arising from their interactions. When there are *w* replicates available, we randomly pick up with replacement *w* arrays for each interactive variable to form a bootstrap sample of gene expression values which is of the same size as the real sample. The bootstrap distributions of the contributions on each dimension by each gene are obtained by repeating the bootstrapping for *B* times. Based on the distributions, we obtain the bootstrap p-value for comparing the estimated contributions with the random mean as $p \equiv \sum_{t=1}^{B} I(ac_t \le ac_o)/B$ where $I(\cdot)$ is the indicator function, ac_t is the absolute contribution estimated for each gene in bootstrap sample *t* and ac_o is the mean random contribution. Note that since we are restrictively resampling the replicate arrays for each interactive variable, the functional dependency among the genes are preserved in the bootstrap samples.

Clustering of significant genes

The selected significant genes can be clustered according to their observed expression profiles using gene clustering methods [22]. The different expression patterns in the clusters can be examined to look for genes that are differentially regulated in response to experiment factors (the main effects) or due to their interactions. Because some genes can significantly contribute to more than one top dimensions, the clustering is performed for all the genes that make significantly high contributions to at least one dimension in CA. The clustering of significant genes can help to establish biologically meaningful associations between the genes and the experiments.

Results

Application to a data in stem cell study

We use data from a microarray experiment (using Affymetrix HG-U133A 2.0 chips each containing 22,000 genes) on stem cells conducted in our lab as an example. In the experiment, two lines of

human mesenchymal stem cells (hMSC), telomerase-immortalized hMSC (hMSC-TERT) and hMSC-TERT stably transduced with the full length human delta-like 1 (Dlk1)/Pref-cDNA (hMSC-dlk1), were treated with vitamin D to examine the effects of Dlk1, vitamin D and their interaction on hMSC growth and differentiation and to look for genes that are differentially expressed in the process. The experiment was done using a 2x2 factorial design. Twelve hybridizations in total were conducted with each of the four interactive variables in triplicates: hMSC-TERT untreated by vitamin D or tert-control (designated as tC), hMSC-TERT treated with vitamin D (tD), hMSC-dlk1 untreated with vitamin D or dlk-control (dC), hMSC-dlk1 treated with vitamin D (dD). We first normalized our raw data (at probe level) using the quantile normalization method as described by Bolstad *et al*. [23]. Then we summarized the intensities for the probes in each probe-set using the robust multi-array average approach [24] to use as the expression value for each gene. Both data normalization and gene expression value calculation were done by the *affy* package in Bioconductor (http://www.bioconductor.org) for R (http://cran.r-project.org). Finally, genes are filtered by dropping those whose expressions failed to vary across the hybridizations or arrays (standard deviation/mean>0.03) which resulted in 2227 genes for subsequent analyses.

The biplots from the correspondence analysis of our stem cell data is shown in Figure 1 where projections of both the genes and the four combinatory variables (between cell lines and vitamin D treatments, the suffix number indicates replicate) along the first dimension or axis are plotted against that along the second (Figure 1a) and the third (Figure 1b) axes. In Figure 1 the first axis, which accounts for 64.7% of the total variance, separates the two cell lines in the data. It is interesting to see that both tC and tD are projected to the left panel and closely coordinated on the first axis while both dC and dD are projected to the right although with some distance between them. It is easy to find that the second axis (accounting for 21% of the total variance) mainly represents the effect of vitamin D treatment in the hMSC-dlk1 cell line (Figure 1a). Unlike Figure

1a, inspection on Figure 1b does not reveal any biological significance. This is understandable because the third axis explains only 4.8% of the total variance. Since the variance in the data is overwhelmingly dominated by the first and the second axes, Figure 1 reveals that significance in the experiment is represented firstly by genes differentially expressed in the two cell lines, and secondly by genes regulated in response to vitamin D treatment in the hMSC-dlk1 cell line. In addition, note that our gene filtering procedure has left a hole in the cloud of genes in the center of Figure 1a.

We use the described bootstrap procedure to obtain the empirical distributions of gene contribution on the different axes and calculate their bootstrap p-values for significance inferences. By resampling for 1000 times, we find highly significant genes ($p<0.001$) that contribute to the first (274 genes) and the second (203 genes) axes. These genes explain 50.5% and 41.7% of the total variance along each of the two axes. For a significance level of $p<0.01$, we have 294 genes contributing to the first axis and 260 genes to the second axis which account for about half (51.9% and 47%) of the total variance carried by the first two axes. The procedure detected only 4 genes contributing to the third axis with $p<0.05$ but no gene with $p<0.01$. Figure 2 is the boxplot showing the bootstrap distribution of gene contribution on the first (Figure 2a) and the second (Figure 2b) axes by the selected highly significant genes ($p<0.001$). The distributions of the bootstrap contribution are all well above the random contribution ($1/2227=0.00045$) indicated by the dashed horizontal line. Because the genes are ranked according to their observed contributions in CA, Figure 2 also shows that it is important to take into account the variations in gene contribution in evaluating their significances because high rank genes tend to exhibit big variations. Figure 3a displays expression profiles for genes highly significantly ($p<0.001$, 439 genes) contribute to the first two axes. It is easy to see that genes in blocks 1 and 2 are mainly up or down-regulated in the hMSC-TERT cell line which represents a cell line effect. Genes in blocks 3-5 are genes showing

interaction effects between the cell lines and vitamin D treatment with genes highly expressed in the hMSC-dlk1 cell line but without vitamin D treatment (block 3), and down expressed when administrated with vitamin D (block 4). Contrary to block 4, block 5 represents another interaction pattern for genes highly expressed in the hMSC-dlk1 cell line with vitamin D treatment. Finally, at the bottom of Figure 3a (block 6), we have a small cluster of genes exhibiting the main effects of vitamin D which are up-regulated in both cell lines. It is necessary to point out that, although genes in the upper part of block 1 are up-regulated in the hMSC-TERT cell line, their activities are suppressed in the hMSC-dlk1 cell line conditionally on the vitamin D treatment effect which may as well be seen as interactions. Such a situation tells us that, in practice, there may not always be a black and white distinction between the main and the interaction effects as predefined by the linear parametric model in ANOVA.

Comparison with ANOVA

We also analyzed the same data set using the existing parametric approach, i.e. ANOVA model, with aim at comparing the performances of the two methods. In the analysis, we fit the expression level of a gene (E) as a linear function of the cell line effect (C), the treatment effect (D) and their interaction (C·D) (Pavlidis, 2003), i.e. we fit

$$E = \mu + C + D + C \cdot D + \varepsilon$$

where μ is the mean expression level of the gene, ε is the random error. Because for each of the 2227 genes, the model independently tests the main effects and their interactions, we introduce the false discovery rate (FDR) [25] to establish the p value threshold and to help to correct for multiple testing. Our analysis detected highly significant genes ($p<0.001$) that are differentially expressed between the two cell lines (601 genes), between the vitamin D treated and untreated groups (56 genes) and as a result of interaction effects (220 genes). The expression profiles of these genes are shown in Figure 3b for cell line effect, Figure 3c for interaction effect, and Figure 3d for the vitamin

D treatment effect. Although for the same significance level, we obtain much higher number of genes in the ANOVA model, the main patterns revealed by the parametric model are captured by our non-parametric approach with Figure 3b corresponds to blocks 1 and 2 in Figure 3a, Figure 3c to blocks 3-5 and top of block 1 in Figure 3a, Figure 3d to blocks 6 and 5 in Figure 3a. The correspondence in the results produced by both methods indicate that our non-parametric approach can be used as an alternative to the parametric ANOVA model to identify differentially expressed genes in factorial microarray experiments.

To further compare with our non-parametric approach, we calculated the total contributions of the highly significant genes in ANOVA on the top two axes in CA. The 601 genes in Figure 3b explain 36.39% of the total variance in the first axis and 16.05% of that in the second axis. The 220 genes in Figure 3c contribute to 17.02% of the variance in the first axis, 25.12% of that in the second axis and the 56 genes in Figure 3d account for only 1.91% of the total variance in the first axis and 3.58% of that in the second axis. These results reflect that, the interaction effect in ANOVA is represented by both the first and mainly the second axes but the cell line effect by the first axis which is in consistency with our non-parametric approach. Note that although both methods detected a relatively small number of genes showing a vitamin D treatment effect independent of the cell lines, such a main effect was not revealed by the biplots in Figure 1 where both genes and the samples are projected onto the most important dimensions. This is sensible given their very small contributions to the major axes. To further link the ANOVA results with that from our non-parametric approach, we examine the variations in the contribution of the highly significant genes in ANOVA on the different dimensions of CA. In Figure 4, we show the boxplots of bootstrap contributions (ranked according to their observed contributions in CA) on the first two axes by the highly significant genes in the ANOVA model that show cell line effect (Figure 4a and b, $p<0.000001$, 60 genes), interaction effect (Figure 4c and d, $p<0.0001$, 62 genes), and effect of

vitamin D treatment (Figure 4e and e, $p<0.001$, 56 genes). Figure 4 reconfirms that very highly significant genes displaying cell line effect in ANOVA mainly significantly contribute to the first axis in CA. Meanwhile, genes estimated as showing highly significant interaction effect in ANOVA can significantly contribute to both the first and the second axes. Moreover, genes as detected to display the effect of vitamin D treatment mainly contribute to the second axis in CA.

It is necessary to point out that even though some of the selected genes in ANOVA make significant contributions to the top dimensions in CA, there are also others that show only random contributions. One obvious example is the genes detected to show significant vitamin D treatment effect in Figure 4e and f. We think that the situation reflects the problem of false positive results in ANOVA even after adjusting for multiple testing.

Discussion

We have presented a non-parametric approach for analyzing high-dimensional microarray data produced in replicated factorial experiments. Application of the method to our stem cell data has helped us to find genes that display contrasting expression profiles in the two cell lines. Our method also detected genes turned on/off due to vitamin D treatment in the hMSC-dlk1 cell line. The results are important in deepening our understanding in the genetic control of stem cell differentiations. As a widely used exploratory method for visualizing multi-dimensional data, CA displays the associations of gene expression with the effects of experiment factors as well as with the interaction effects between the factors. In the linear regression based ANOVA model, unsupervised analysis of FED data requires that parameters be assigned to each of the factors as well as to each of the interaction terms which can easily run into model identifiability problem and false positive results due to increased multiple testing. By data visualization using the biplot, CA reveals the main effects and interactions that dominate the major variations in the data and thus results in increased

efficiency in data analysis through dimension reduction. Although our example data is in a 2x2 factorial design, generalization of our method to more factors is just straightforward.

In the ANOVA model, parameters are assigned to stand for either the main effects or the interactions. However, such a black-and-white assertion may not always hold in biological reality. In Figure 3a, although genes in the upper part of bock 1 display a cell line effect (high expression in the hMSC-TERT cell line), they are also up or down-regulated in the hMSC-dlk1 cell line but conditional on vitamin D treatment, a situation which may reflect an interaction effect. On the contrary, the interaction effects (up and down regulation) between hMSC-dlk1 cell line and vitamin D treatment are clearly represented by genes in blocks 3 and 4 in Figure 3a. The example illustrates the necessity of model-free approach in modeling biological data.

Kerr and Churchill [21] emphasized the importance of replicates in microarray experiments. In their linear regression based ANOVA model [12], sufficient replications are needed to ensure model identifiability and accuracy of the parameter estimates as well as to examine their model assumptions (normality, linearity, etc). In our non-parametric approach, replicates are solely used for assessing the distribution of gene contributions on the major dimensions that dominate the variance in the observed data. This operating characteristic should naturally enable our method to deal with data in high order FEDs in an efficient manner. Most importantly, in our bootstrap resampling procedure, the inherent functional dependency among the genes is kept intact. This is different from the ANOVA model which ignores the correlation in gene activities by testing the genes independently.

Another nice feature in our non-parametric approach is that CA can also help to standardize the variance in the data. Because the individual elements in matrix C which is submitted to SVD can be viewed as the standardized residuals, the algorithm helps to compensate for the larger variance in genes with stronger signals and the smaller variance in genes with weaker signals. This feature thus

serves as an additional way to alleviate the intensity-dependent variance problem in microarray data [26].

Although in this paper we focus on applying our method to analyze microarray data from complex factorial experiments, we are planning to introduce the same approach to other types of clinical investigations, for example, tumor classifications. In that case, the bootstrap-assisted CA could help us to cluster the genes while associating them with the clustered tumor subclasses and moreover to validate the differences between the tumor classes. Such practice is important because the global gene expression profiles characterized by the significant marker genes can provide useful information for tumor diagnosis, treatment strategies and outcome predictions.

Conclusion

Factorial experiments have the advantage of giving greater precision for estimating overall factor effects, of enabling interactions between different factors to be explored [27]. These nice features promote the use of FED in miroarray studies [11]. We have shown how our non-parametric procedures can be applied to identify the clusters of genes that exhibit differential expression profiles induced by the main factors or by interactions between the factors, and meanwhile to validate their significances. We hope our model-free procedures introduced in this paper can serve as an alternative to the existing ANOVA model in analyzing microarray gene expression data in factorial design.

Acknowledgements

This work was financed by the Danish Biotechnology Instrument Center (DABIC) under the biotechnological research program of the Danish Research Agency, and supported by the Clinical Institute at OUH and by the Danish Center for Stem Cell Research.

References

1. Lander ES: **Array of hope**. *Nat Genet*. 1999, **21:** S3-S4.

2. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome wide expression patterns.** *Proc Natl Acad Sci U S A*., 1998, **95:** 14863-14868.

3. Toronen P, Kolehmainen M, Wong G, Castren E: **Analysis of gene expression data using self-organizing maps**. *FEBS Lett*., 1999, **451:** 142-146.

4. Raychaudhuri S, Stuart JM, Altman RB: **Principal components analysis to summarize microarray experiments: application to sporulation time series**. *Pac Symp Biocomput*., 2000, 455-466.

5. Byvatov E, Schneider G: **Support vector machine applications in bioinformatics**. *Appl Bioinformatics*, 2003, **2:** 67-77.

6. Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data**. *Journal of the American Statistical Association*, 2002*,* **97:** 77-87.

7. Shaw R, Festing MF, Peers I, Furlong L: **Use of factorial designs to optimize animal experiments and reduce animal use**. *ILAR J*., 2002, **43:** 223-232.

8. Wildsmith SE, Archer GE, Winkley AJ, Lane PW, Bugelski PJ: **Maximization of signal derived from cDNA microarrays**. *Biotechniques*, 2001, **30:** 202-208.

9. Yang YH, Speed T: **Design issues for cDNA microarray experiments**. *Nat Rev Genet*., 2002, **3:** 579-588.

10. Churchill GA: **Fundamentals of experimental design for cDNA microarrays**. *Nat Genet*., 2002, **32:** S490-S495.

11. Glonek GF, Solomon PJ: **Factorial and time course designs for cDNA microarray experiments**. *Biostatistics*, 2004, **5:** 89-111.

12. Kerr MK, Martin M, Churchill GA: **Analysis of variance for gene expression microarray data**. *J Comput Biol*., 2000, **7:** 819-837.

13. Pavlidis P: **Using ANOVA for gene selection from microarray studies of the nervous system**. *Methods.*, 2003, **31:** 282-289.

14. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci U S A*., 2000, **97:** 10101-10106.

15. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV: **Fundamental patterns underlying gene expression profiles: simplicity from complexity**. *Proc Natl Acad Sci U S A*., 2000, **97:** 8409-8414.

16. Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M: **Correspondence analysis applied to microarray data**. *Proc Natl Acad Sci U S A*., 2001, **98:** 10781-10786.

17. Tan Q, Brusgaard K, Kruse TA, Oakeley E, Hemmings B, Beck-Nielsen H, Hansen L, Gaster M: **Correspondence analysis of microarray time-course data in case–control design**, *Journal of Biomedical Informatics,* 2004, **37:** 358-365.

18. Ghosh D: **Resampling methods for variance estimation of singular value decomposition analyses from microarray experiments**. *Funct Integr Genomics*., 2002, **2:** 92-97.

19. Clausen SE: **Applied correspondence analysis: An introduction**. Sage publications. 1988.

20. Gabriel KR, Odoroff CL: **Biplots in biomedical research**. *Stat Med*., 1990, **9:** 469-485.

21. Kerr MK, Churchill GA: **Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments**. *Proc Natl Acad Sci U S A*., 2001, **98:** 8961-8965.

22. Shannon W, Culverhouse R, Duncan J: **Analyzing microarray data using cluster analysis**. *Pharmacogenomics*, 2003, **4:** 41-51.

23. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics*, 2003, **19:** 185-193.

24. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics*, 2003, **4:** 249-264.

25. Reiner A, Yekutieli D, Benjamini Y: **Identifying differentially expressed genes using false discovery rate controlling procedures**. *Bioinformatics*, 2003, **19:** 368-375.

26. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics*, 2002, **18:** S96-S104.

27. Cox DR: **Planning experiments**. New York: John Wiley and Sons. 1958.

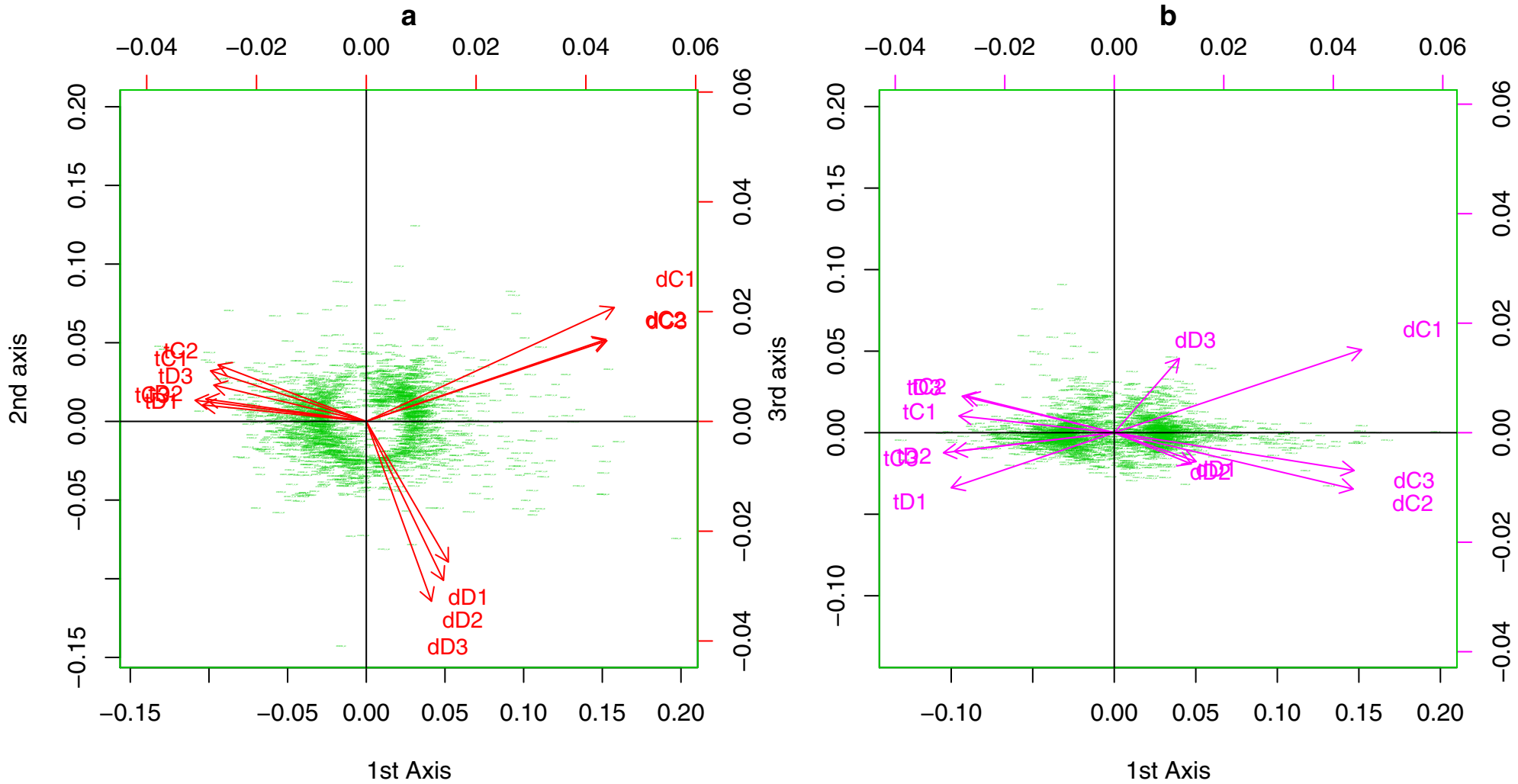
Figure captions:

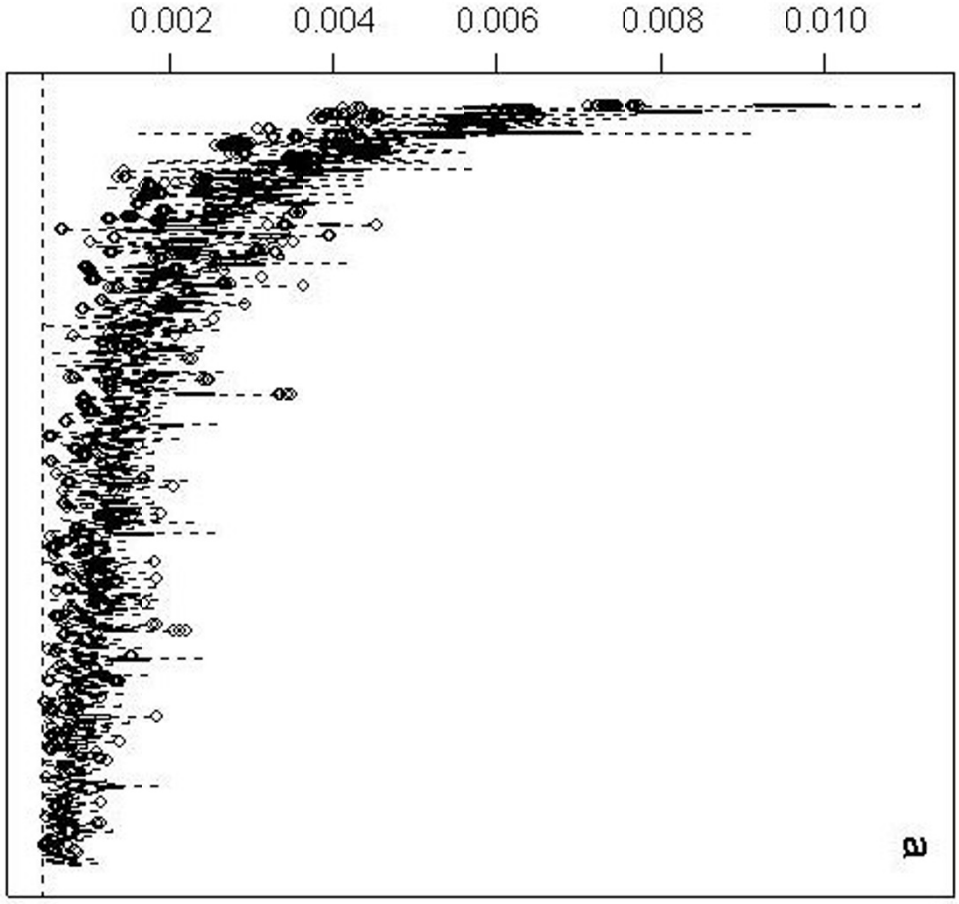
Figure 1. Biplots showing the projections by both genes and the interactive variables (samples) on the first axis against that on the second (1a) and the third (1b) axes. The first axis is mainly dominated by the variance in gene expression in the two cell lines while the second axis by the interaction effect between vitamin D treatment and the hMSC-dlk1 cell line. However, the pattern in the third axis is not meaningful.

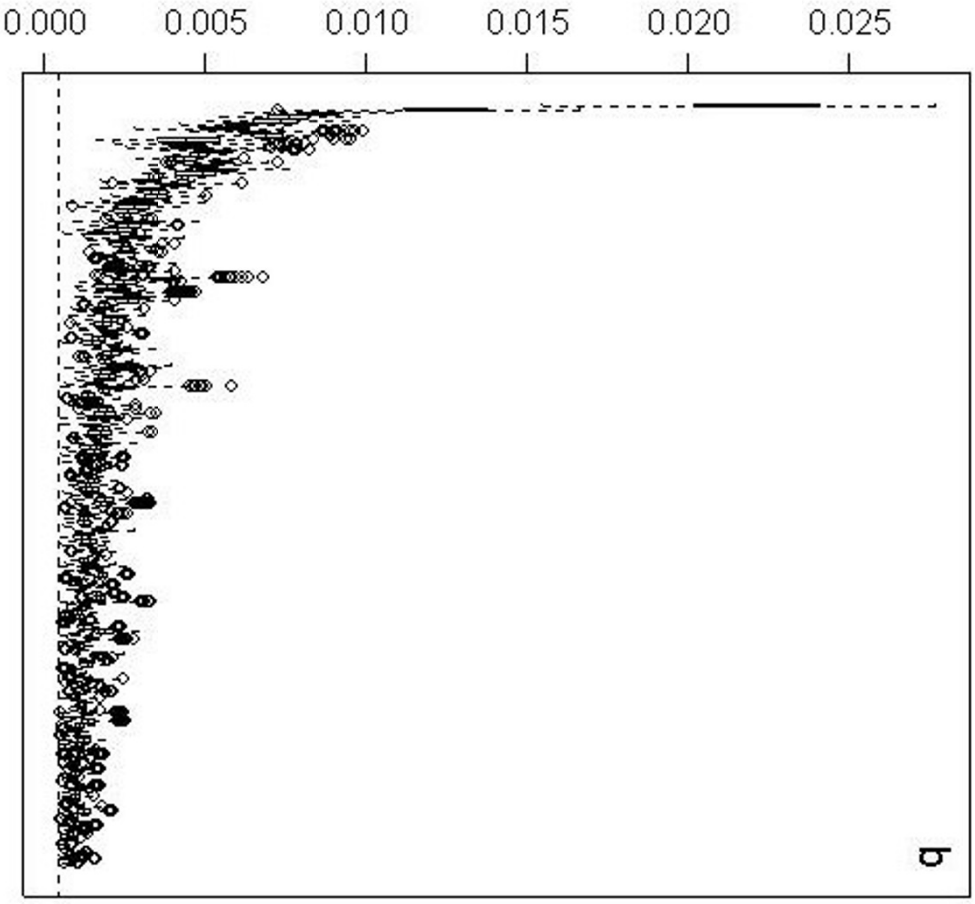
Figure 2. Boxplots showing the bootstrap distributions of gene contribution on the first (2a) and the second (2b) axes for genes whose bootstrap p-value<0.001.

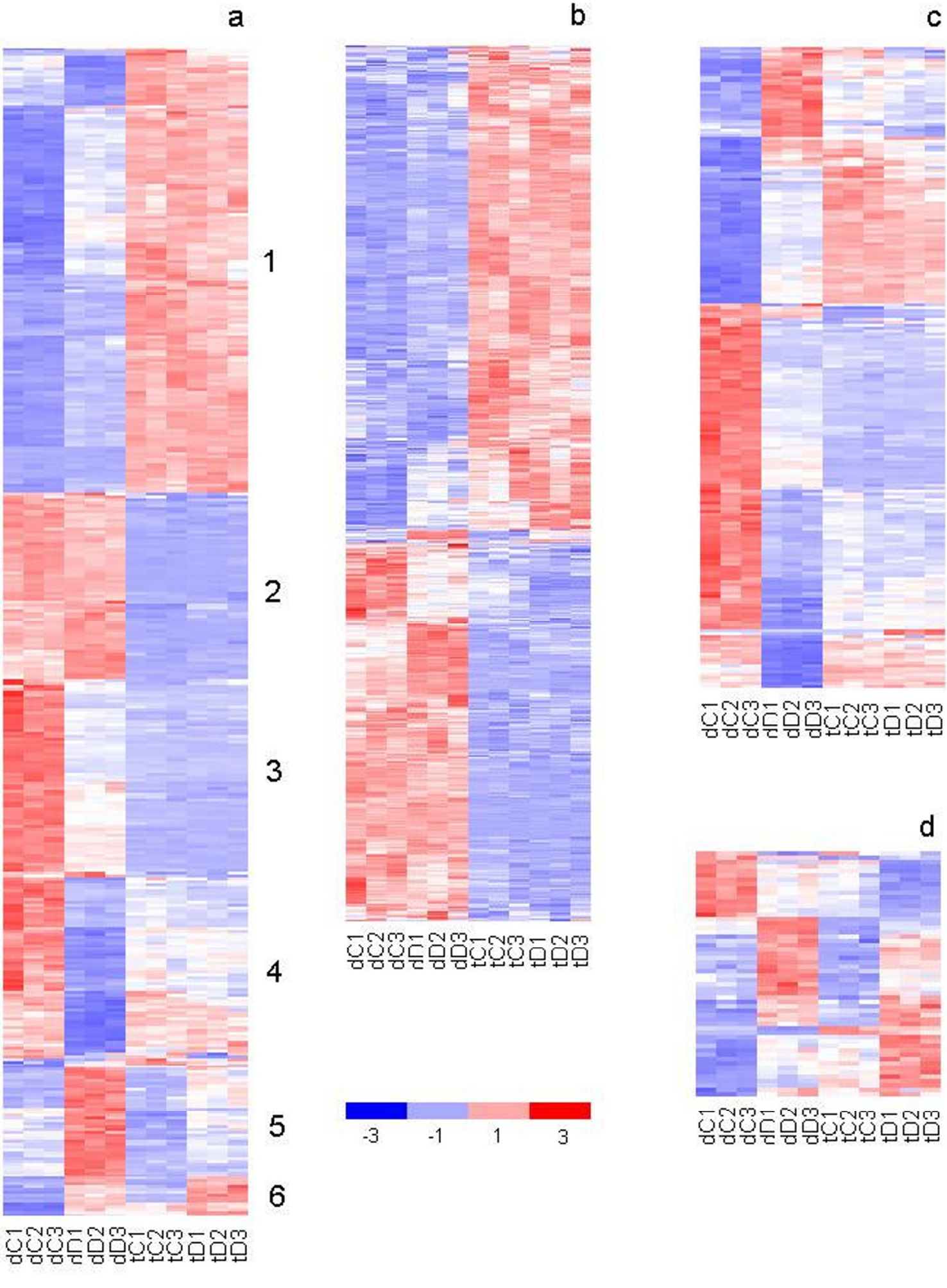
Figure 3. Expression profiles for genes that significantly contribute to the first tow axes (3a) ($p<0.001$) in CA and for significant genes ($p<0.001$) detected as displaying the cell line effect (3b), the interaction effect (3c), and effect of vitamin D treatment (3d) in the ANOVA model.

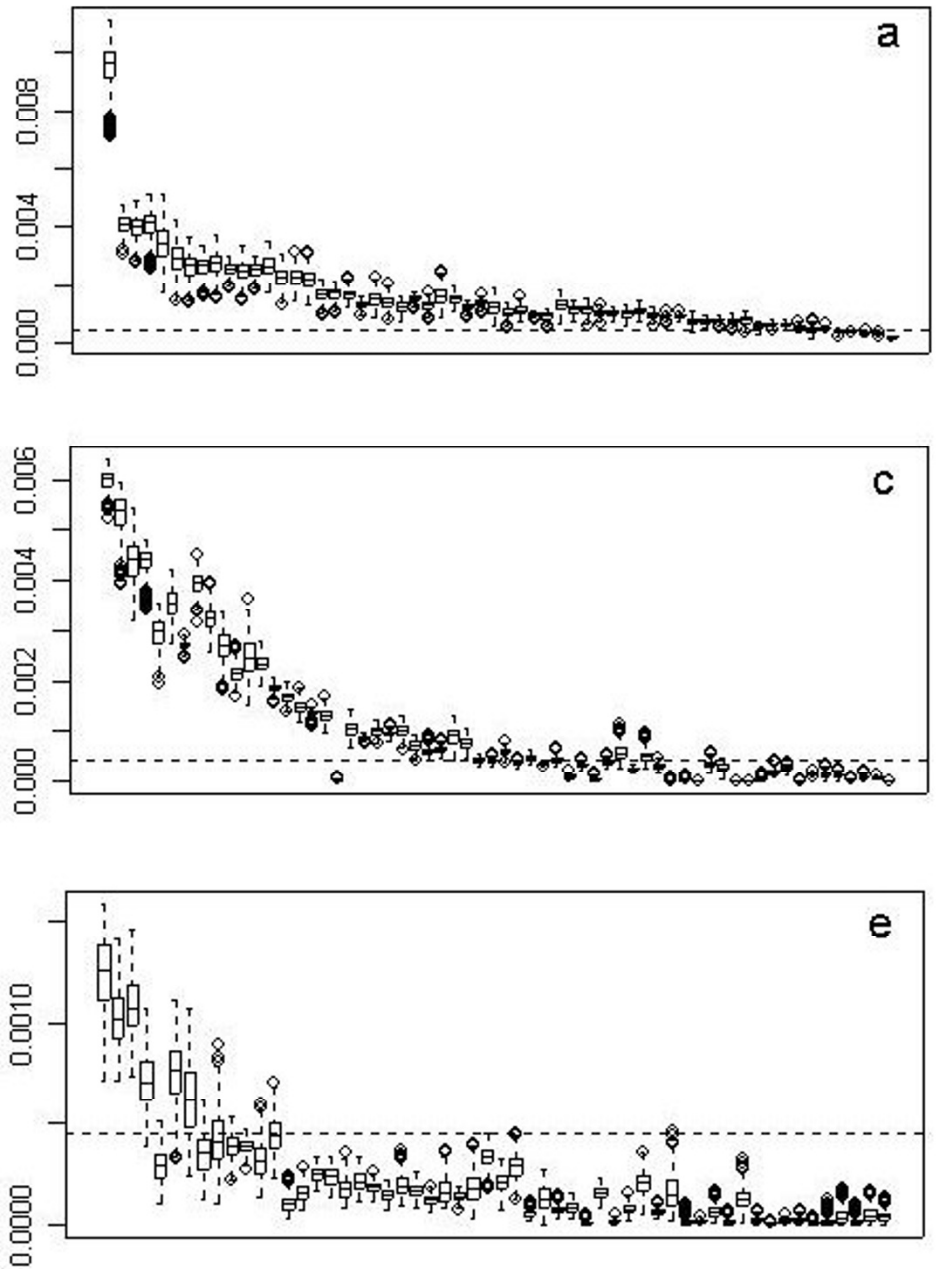
Figure 4. Boxplots showing the bootstrap distributions of gene contribution on the first two axes for significant genes that display cell line effect (4a and b, $p<0.000001$), interaction effect (4c and d, $p<0.0001$), and vitamin D treatment effect (4e,f, $p<0.001$) in the ANOVA model.

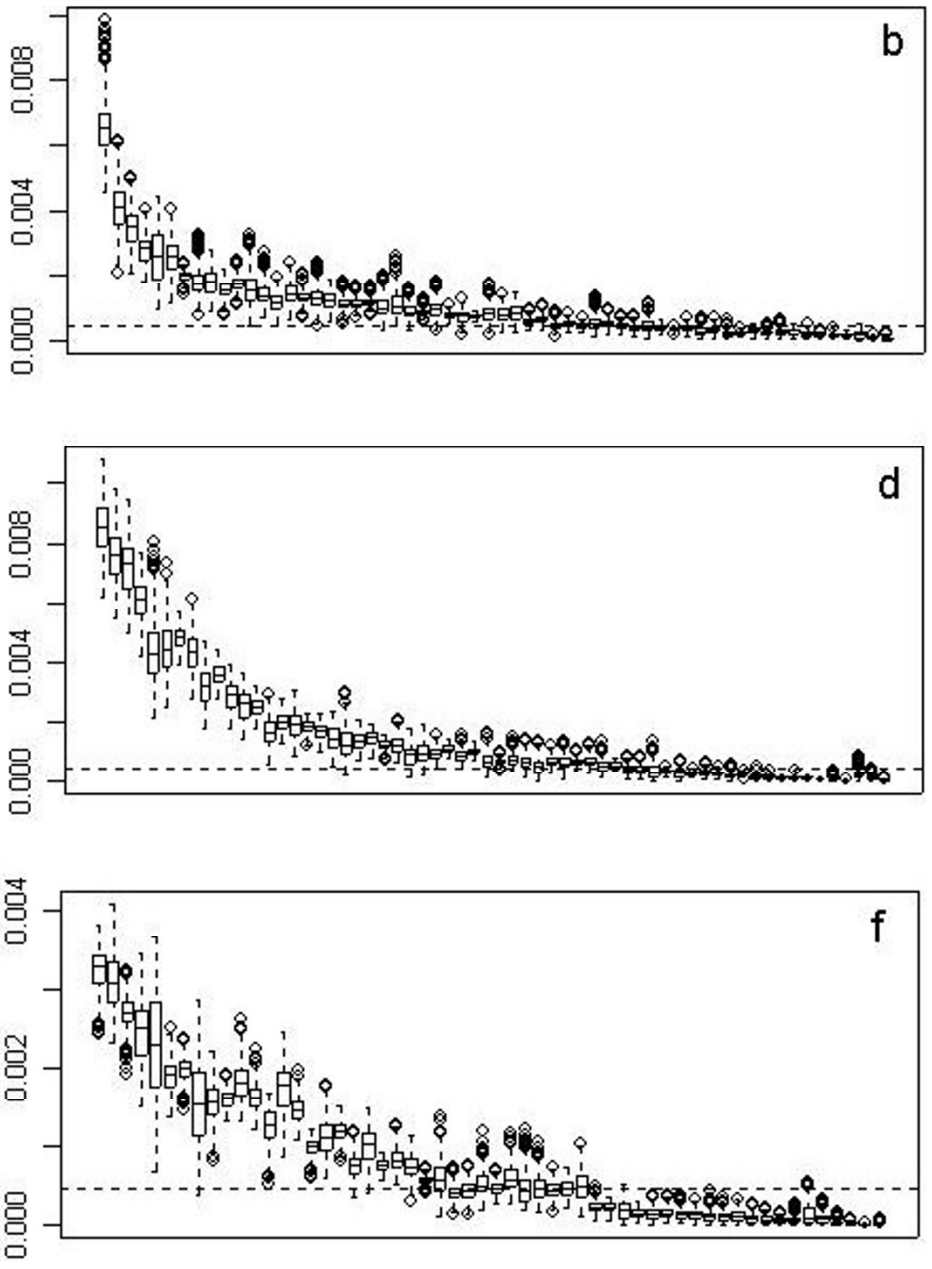

a
b
2nd axis
3rd axis
1st Axis
1st Axis
dC1
dC2
dC3
dD1
dD2
dD3
tC1
tC2
tD1
tD2
tD3


0.002
0.004
0.006
0.008
0.010
a


0.000
0.005
0.010
0.015
0.020
0.025
b


a
b
c
d
1
2
3
4
5
6
dC1
dC2
dC3
dN1
dD2
dD3
tC1
tC2
tC3
tD1
tD2
tD3
-3
-1
1
3


a
0.000
0.004
0.008


b
0.000
0.004
0.008

c
0.000
0.002
0.004
0.006

d
0.000
0.004
0.008

e
0.0000
0.0010

f
0.000
0.002
0.004