Research

Permutation-validated principal components analysis of microarray data

Jobst Landgrebe*, Wolfgang Wurst*† and Gerhard Welzl‡

Addresses: *Max Planck Institute of Psychiatry, Molecular Neurogenetics, Kraepelinstrasse 2-10, 80804 Munich, Germany. †Institute of Mammalian Genetics and ‡Institute of Biomathematics and Biometry, GSF-National Research Center for Environment and Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany.

Correspondence: Gerhard Welzl. E-mail: welzl@gsf.de

Published: 22 March 2002

Genome ***Biology*** 2002, **3(4)**:research0019.1–0019.11

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/4/research/0019

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 12 December 2001
Revised: 31 January 2002
Accepted: 15 February 2002

Abstract

Background: In microarray data analysis, the comparison of gene-expression profiles with respect to different conditions and the selection of biologically interesting genes are crucial tasks. Multivariate statistical methods have been applied to analyze these large datasets. Less work has been published concerning the assessment of the reliability of gene-selection procedures. Here we describe a method to assess reliability in multivariate microarray data analysis using permutation-validated principal components analysis (PCA). The approach is designed for microarray data with a group structure.

Results: We used PCA to detect the major sources of variance underlying the hybridization conditions followed by gene selection based on PCA-derived and permutation-based test statistics. We validated our method by applying it to well characterized yeast cell-cycle data and to two datasets from our laboratory. We could describe the major sources of variance, select informative genes and visualize the relationship of genes and arrays. We observed differences in the level of the explained variance and the interpretability of the selected genes.

Conclusions: Combining data visualization and permutation-based gene selection, permutation-validated PCA enables one to illustrate gene-expression variance between several conditions and to select genes by taking into account the relationship of between-group to within-group variance of genes. The method can be used to extract the leading sources of variance from microarray data, to visualize relationships between genes and hybridizations and to select informative genes in a statistically reliable manner. This selection accounts for the level of reproducibility of replicates or group structure as well as gene-specific scatter. Visualization of the data can support a straightforward biological interpretation.

Background

Microarrays have become standard tools for gene expression analysis as the messenger RNA levels of thousands of genes can be measured in one assay. In a standard microarray experiment, total RNA or mRNA is extracted from cells or tissue, labeled by reverse transcription with radioactive or fluorescent-tag-labeled nucleotides and hybridized to the arrays. After hybridization and washing, the arrays are scanned and the hybridization intensities at each spot are determined by image-analysis software. Two-channel microarrays open up the possibility of carrying out many hybridizations in parallel using a common reference RNA. In

such experiments, different experimental conditions can be compared to each other. In many cases, different conditions are analyzed with some replications to allow variance analysis [1,2]. This procedure results in multivariate grouped data in which one group represents a condition with several replicates. Such data can be represented as a matrix with n rows (genes) and p columns (hybridizations) and a vector of length p containing the group labels. These data are characteristic of multi-condition microarray experiments.

To analyze such data, multivariate statistics are needed. Before carrying out the analysis, the data must be preprocessed by background subtraction, computation of ratios and array-wise normalization. After this step, the data can be analyzed using different multivariate approaches. These methods can be classified as supervised and unsupervised. A wide variety of supervised approaches have been described, for example, classification and regression trees [3] or support vector machines [4]. Among unsupervised methods, hierarchical clustering [5] and other clustering approaches [6,7], as well as projection methods such as multidimensional scaling [8], principal components analysis (PCA) [9-13] and correspondence analysis [14] have been described. Such projection techniques reduce the dimensionality of multivariate data to embed the variables and objects of the data in a visualizable (two- or three-dimensional) space. The projection aims to represent the objects and variables in the reduced space in such a way that they approximate their original distances in the high-dimensional space. This enables one to extract and visualize the dominant effects on variance from the data. With PCA, linear combinations (principal components) of the original variables can thus be functionally interpreted (for review see [15]). This enables a biological interpretation of the nature of coherent variation.

In microarray experiments, the identification of subsets of genes with large variation between groups is of primary interest. This process has to comprise a criterion that accounts for the variance within groups. Sometimes this selection is only the first step in the data analysis. Hastie *et al.* [16] carried out hierarchical clustering of gene-expression data and computed an average expression profile for each cluster, providing the input for a response model. A direct significance analysis to select genes from microarray data (SAM) was proposed by Tusher *et al.* [17]. This method is based on t-like (in the case of two conditions) or F-like statistics.

Several methods for gene selection involving PCA have been proposed. The 'gene shaving' approach of Hastie *et al.* [10] restricts PCA to the first principal component. Groups of genes are generated by iterative exclusion of fixed fractions of genes (typically 10%) with smallest absolute loadings according to the first principal component. After orthogonalization of the data with respect to an averaged expression profile of the first group, the procedure is repeated. Another PCA-based method of gene selection using PCA-derived gene coefficient vectors and F-statistics was described by Landgrebe *et al.* [18].

Although these methods allow the detection of patterns or 'characteristic nodes' by dimension reduction and the selection of gene subsets with large variation between condition groups, the reliability of the results has to be determined. Therefore, it is imperative to assess whether the results are statistically reliable relative to the level of noise in the data. Classical statistical parametric tests depend on the assumptions of normality and independence of variables (hybridizations). Yet, these assumptions do not hold for microarray data [1,19]. Consequently, computationally intensive methods such as permutation tests or bootstrap procedures as introduced by Efron [20] are preferable. Kerr *et al.* [1] show the application of bootstrap technique to clustering results. Ghosh [21] describes another approach based on mixture modeling to assess the reliability of clustering results. Other permutation-based approaches were published by Tusher *et al.* [17] and Dudoit *et al.* [3]. The method proposed by Hastie *et al.* [10] also contains bootstrap elements. An approach of Wall *et al.* [22] tries to combine PCA-based gene selection with a confidence measure using leave-one-out cross-validation.

Here we describe an approach combining PCA-directed gene selection with validation by permutation tests. We use a test statistic based on the genes' object scores to select genes with high variance with respect to the principal components. The method was developed for the analysis of microarray data having several conditions with a few replicates per condition or a group structure. We demonstrate this approach by applying it to the well-characterized data of Spellman *et al.* [23]. Although other methods are better adapted to the analysis of temporal effects (for example [24]), we chose the yeast data to allow comparison with other approaches applied to this dataset [14,23]. In addition, two datasets generated in our own laboratory were also analyzed. Our method was successfully applied to the different datasets. We revealed the main sources of variance in the data and described the genes related to this variance. This allowed the interpretation of variance and the permutation-validated selection of genes in a functional context.

Results

Permutation-test-validated PCA

We carried out permutation-test-validated PCA on grouped data with few replicates to study variation in gene expression across several conditions, to understand the structure of the data, to uncover patterns underlying the hybridization conditions and to identify subsets of genes with large variation across these patterns. PCA is primarily aimed at finding and interpreting complex relationships between variables in a dataset. Correlated variables are converted to factors that

are not correlated to each other. The central point of such analysis is to decompose the original n x p data matrix (n objects, p variables) in the following manner:

$$\mathbf{X} = \mathbf{A}\mathbf{F}^{\mathbf{T}},$$

where $\mathbf{X}$ is the n x p data matrix, $\mathbf{A}$ is the n x p matrix of factor scores and $\mathbf{F}$ is the p x p matrix of factor loadings. With $s = p$ factors the total variance of all variables is explained. The decomposition of $\mathbf{X}$ is done in such a way that the factors explain the total variance in a descending order. Therefore, it is possible to reduce the data to s dimensions with a minimum loss of information expressed by the matrix of residuals $\mathbf{E}$:

$$\mathbf{X} = \tilde{\mathbf{A}}\tilde{\mathbf{F}}^{\mathbf{T}} + \mathbf{E},$$

where $\tilde{\mathbf{A}}$ is the n x s matrix of factor scores, $\tilde{\mathbf{F}}$ the p x s matrix of factor loadings and $\mathbf{E}$ is the matrix of residuals as a result of dimension reduction. In this manner, PCA provides a projection of the objects from p-dimensional to s-dimensional space.

In grouped data with replicates per group (condition), there is additional information about the columns of the data matrix: $y' = (y_1, y_2, \ldots, y_p)$ is a set of class labels identifying the replicates for each condition. Although PCA is generally not considered to be appropriate for grouped data, the method has been adapted for this data type (rank-ordered PCA [25]).

The consecutive steps of the permutation validated PCA procedure are shown in Figure 1. In step 1, we perform rank-ordered PCA based on the polished gene expression matrix $\mathbf{X}$ (see Materials and methods) by computing separate one-way ANOVAs on the principal components loadings for each of the components. If the between-group variance dominates the total variability in the data, PCA discriminates between groups. In this situation, the first components of the PCA and components with high F-values are identical. Thus, following the order of explained variance, we select the components with significant F-statistics ($p \leq 0.01$). At the occurrence of a component with nonsignificant F-statistics, we terminate the selection. This process results in k components (step 2). Data approximated in the space based on these components reflect the between-group variance. Thus, in step 3 of the procedure, we compute components from the group-averaged data and derive the exact between-group variance for each gene, which can be estimated by:

$$s_g^2 = \frac{1}{p-1}\sum_{i=1}^{k} a_{gi}^2,$$

where a_{gi} is the factor score for gene g and component i. As a test value, we use $t_g = (p-1)s_g^2$ (step 3). Genes with a high value are candidates for selection.

1. Rank-ordered PCA for grouped data based on polished expression matrix $\mathbf{X}$ with rows centered to have zero mean and standardized columns.
2. Selection of components with significant F-statistics (ANOVA, $p \leq 0.01$) following the order of explained variance. Termination of the selection process at the occurrence of a component with non significant F-statistics resulting in k components.
3. PCA based on group-averaged data, computation of the test value
$$t_g = \sum_{i=1}^{k} a_{gi}^2$$
4. Random permutation of the columns of $\mathbf{X}$, PCA based on group-averaged resampled data, computation of
$$T_g = \sum_{i=1}^{k} a_{gi}^2$$
for each gene g.
5. Repeat step 4 to get permutation distributions of T_g.
6. Select genes with t_g greater than the 95% quantile of the permutation distributions T_g.
7. Visualize arrays and genes in the reduced space of k components.

Figure 1
The permutation-validated PCA procedure for grouped data.

To assess the reliability of the results we perform a permutation analysis (steps 4-6). Under the hypothesis of no effect of different conditions on gene expression profiles, the class labels given by y' are randomly sampled to determine the permutation distribution of the required test statistics. Computing PCAs from randomized group-averaged data yields the distribution of the test statistic T_g for each gene g (step 4):

$$T_g = (p-1)s_g^2 = \sum_{i=1}^{k} a_{gi}^2.$$

We compute 1,000 permutation distributions for each gene (step 5). In step 6 we select the genes for which t_g is greater than the 95% quantile of the permutation distribution of T_g. The last step is the visualization of the arrays and selected genes in the reduced k-components space. If $k = 2$, a twofold visualization is suggested. The biplot with marked selected genes can be used to relate genes and conditions. Genes lying near an axis of a condition are upregulated in this hybridization and genes lying in the opposite direction are repressed. With several conditions, this relation is generally not unique. Therefore, the visualization may be supported by color-coded expression-profile tables. Here, the data matrix is rearranged according to the angular distance from the x-axis for each gene (rearranging n rows). The same is done for hybridizations (rearranging p columns). If $k > 2$ several biplots and color-coded tables must be constructed.

Application to yeast cell-cycle data

To demonstrate our approach, we applied it to the yeast cell-cycle data published by Spellman *et al.* [23]. These authors synchronized the yeast cell cycle using independent methods of cell-cycle arrest and measured the expression of all yeast open reading frames (ORFs) at different time points after the cell-cycle synchronization. They identified genes related to the cell cycle using Fourier transformation and correlation measures. We analyzed the cell-cycle-related genes selected by Spellman *et al.* [23] to demonstrate the relationship between cell-cycle phases and gene-expression patterns and to select a subset of genes that show the highest and most reproducible variance across the cell-cycle phases. We analyzed the expression patterns of 773 selected genes over all 73 hybridizations.

The cell cycle is a temporal continuum that is generally grouped into cell-cycle phases. This classification was also carried out by Spellman *et al.* [23]. The classification of genes in cell-cycle phase groups enables one to analyze the variance of gene expression across the cell-cycle phases and to select genes with different and robust regulation. We analyzed the data using permutation-validated PCA. A first PCA was based on the polished logarithmic ratios including all arrays. An analysis of variance (ANOVA) using the variable loadings as dependent variables and the classification-derived cell-cycle phase groups as factors was carried out. The first two components were highly significant whereas the others were not. Figure 2 shows a plot of the first two component loadings against each other. Of the data's variance, 37.2% was explained by the first two components. The plot shows that the resulting ordination of the hybridizations corresponds to the assignment of cell-cycle phases by Spellman *et al.* [23]. The angular position of the hybridizations in the plot reflects their correct temporal situation in the cycle. The first seven arrays of the CDC series seem to be misclassified. They are shown with colored labels in Figure 2 and show a counterclockwise shift in expression in the cell cycle. Thus,

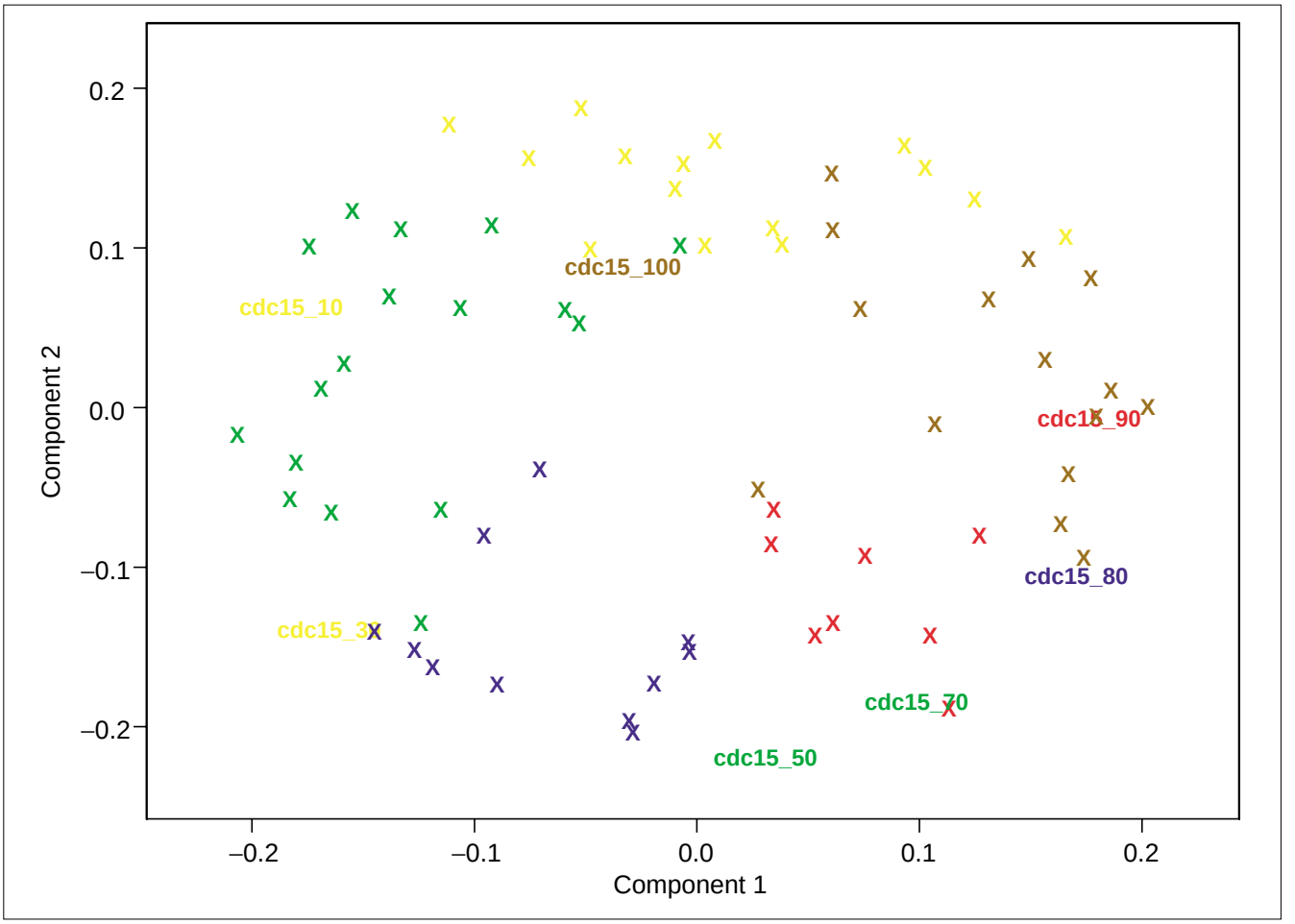

Figure 2
PCA of the cell-cycle data. Plot of the first two components' loadings against each other. Crosses or gene names represent the variable (hybridization) loadings. Colors indicate the cycle phase to which the hybridizations were classified by Spellman *et al.* [23]. M/G1, yellow; G1, green; S, blue; G2, red; M, brown. The labeled arrays are misclassified CDC hybridizations assigned to new cell-cycle phases by PCA.

the first two components approximate the cell-cycle aspect of the data quite well and could be interpreted as cell-cycle components. They reflect the main temporal aspect of the data.

A second PCA was carried out on the group-averaged data using the original cell-cycle classification with two components (94.4% explained variance). Figure 3 shows a biplot of the 773 gene scores and the five cell-cycle phase group loadings. For each gene the distance to the origin indicates the variance in the reduced two-dimensional space. The hole in the middle of the plot reflects the fact that only genes related to the cell cycle were chosen by Spellman *et al.* [23]. Genes without variance with respect to the cell cycle (equally transcribed in most cell-cycle phases) would lie in the middle of the biplot. In Figure 3, 60 genes are labeled with gene symbols. These genes had a test value above the 95% percentile of the permutation distribution of the related test statistic. Figure 3 allows the assignment of the genes to the cell-cycle phases in which they are regulated. As illustrated by Figures 3 and 4, in the cell-cycle phase M/G1, *CDC46* (encoding part of the replication complex) was selected as an upregulated gene, whereas the histone genes *HTB2*, *HTA2* and *HHO1* (also marked by Spellman *et al.* [23] and Fellenberg *et al.* [14]) were selected as downregulated genes. In phases G1 and S, *POL30* (replication complex) and *RAD51* (cell-cycle-related protein kinase) were selected. The histone genes repressed in M/G1 were upregulated in S. In G2 and M, *CLB1* (G2/M-specific cyclin involved in mitotic induction), *CDC5* (mitotic DNA replication) and *CLB2* (G2/M-specific cyclin involved in mitotic induction) were selected as upregulated, in phase M *CDC20* (cyclin degradation, part of the anaphase-promoting complex). Thus, among the known genes selected by our algorithm, many play a crucial role in the cell cycle. As described by Spellman *et al.* [23], the microarray expression data confirm the results of other gene expression studies.

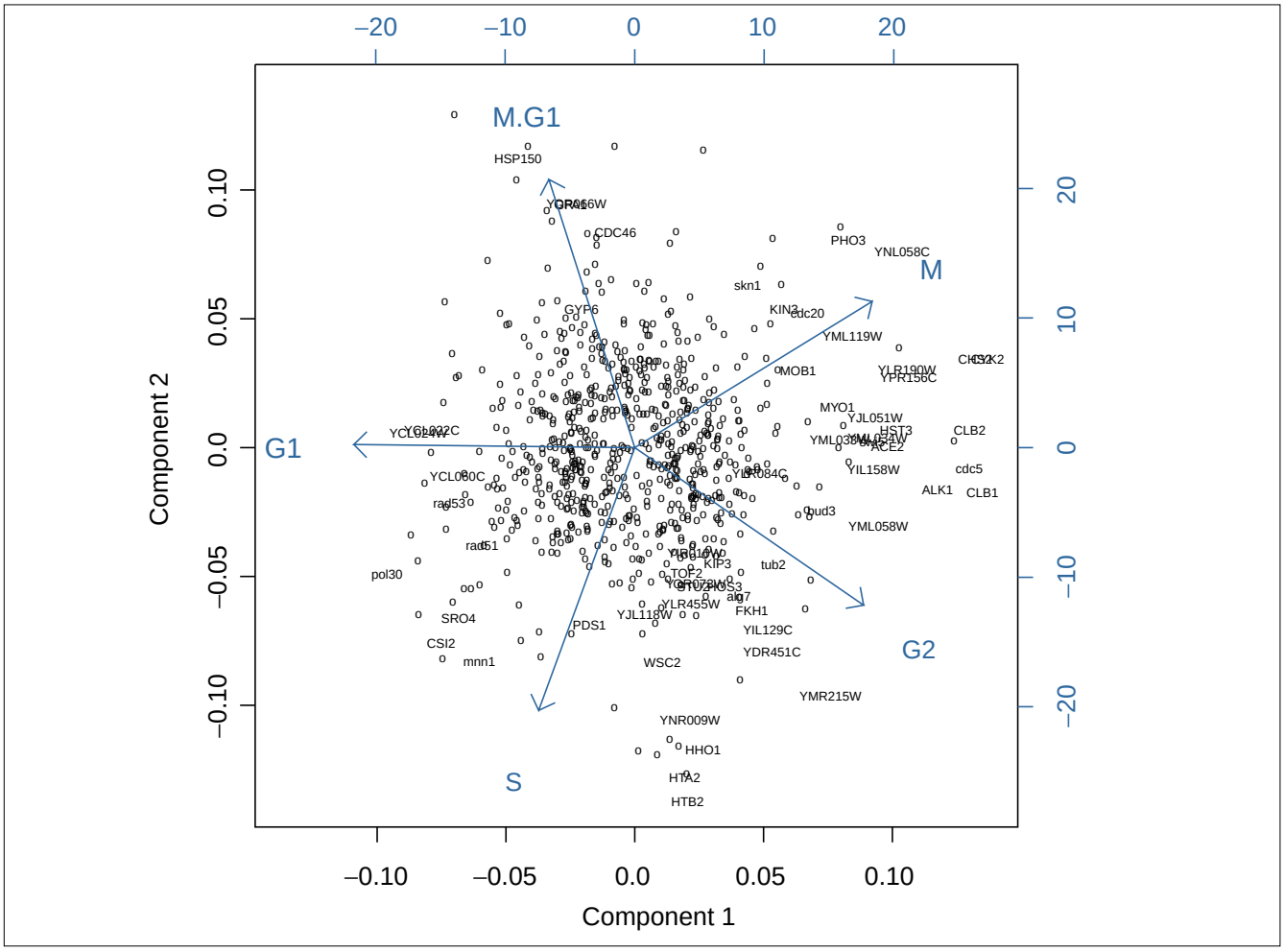

Figure 3
PCA of the grouped cell-cycle data. Biplot of gene scores and cell-cycle phase group loadings according to the first two components of the PCA. The open circles or gene names represent the gene scores. The vectors represent the cell-cycle phase group (variable) loadings. The biplot enables the association of genes with the cell-cycle phase groups. Labeled genes were selected by permutation test.

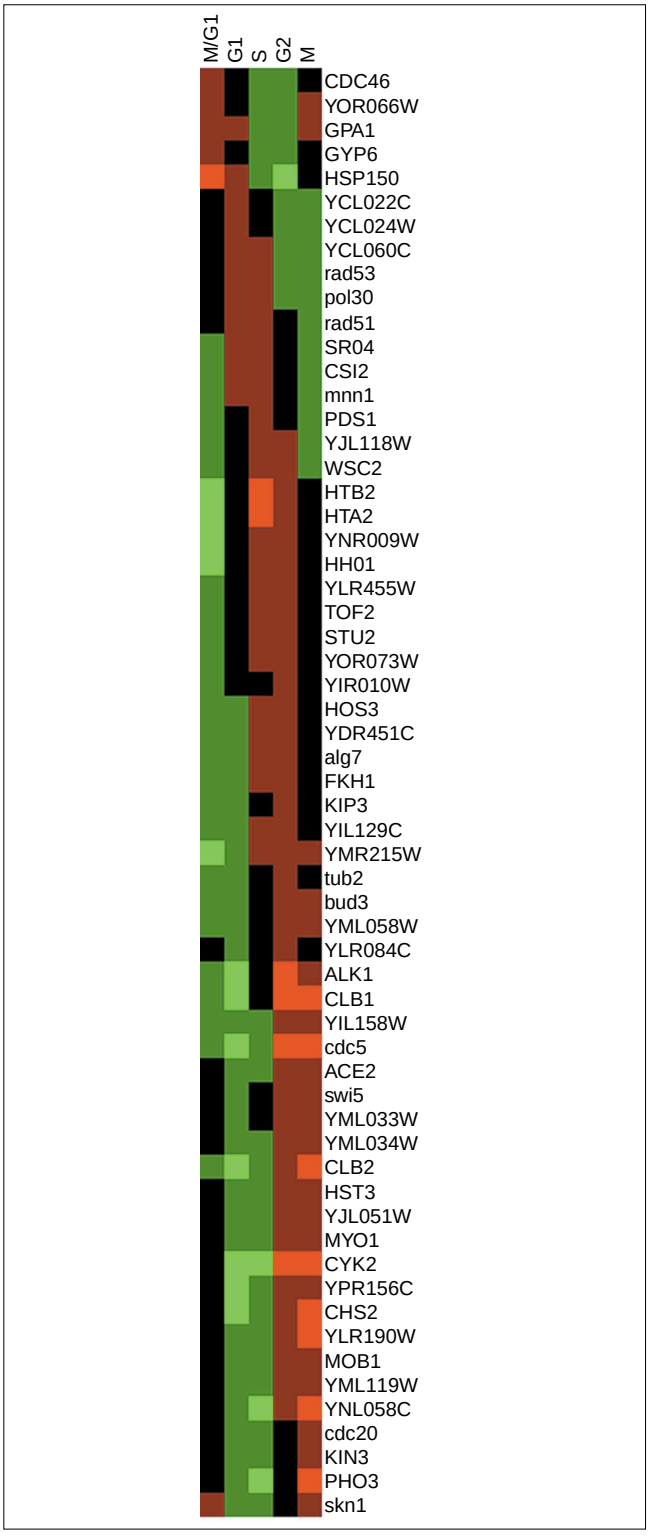

Figure 4
Color-coded expression-profile table of the genes selected from the cell-cycle data by permutation test. Abscissa, cell-cycle phase groups; ordinate, genes. The cycle phase groups and the genes are arranged according to the angular position in the PCA biplot (see Figure 3). Upregulation, red; downregulation, green. The figure was generated using Michael Eisen's software TreeView [4].

Application to abolition of CRH-R1 function data

In this experiment, mRNAs from whole brains of mice of different genetic backgrounds which had been treated with an antagonist directed against the ligand-binding domain of a seven-transmembrane neuropeptide receptor [26] (the corticotropin-releasing hormone receptor 1 (CRH-R1) [27]) were compared to mRNAs from brains of mice lacking a functional CRH-R1 (CRH-R1 knockout mice [28]) using cDNA-microarrays (Table 1). The data consist of $\log_2$ ratios. The matrix had 1,810 complete observations (genes) and 21 hybridizations. We computed a PCA based on the polished matrix of single hybridizations to show that the treatment group members clustered together (data not shown). We performed an ANOVA using the variable loadings as dependent variables and the treatment groups as factors. The first two components were highly significant, whereas the third component was not. The first two components explained 37.5% of the data's variance. We carried out PCA for group-averaged data with two components (54.8% explained variance). Figure 5 shows a biplot of these components. The two components describe a gradient effect of the abolition of CRH-R1 function in different genetic backgrounds. Component 1 (abscissa) distinguishes the CRH-R1 abolition (null mutant) from relatively mild CRH-R1 function impairment (h1, w1: 1 day of treatment with antagonist). With the increasing effect on the animals of gene function impairment, the animals' loadings on the first component become more similar to the genetic CRH-R1 inactivation. Component 2 (ordinate) distinguishes between impairment of heterozygotes treated for 1 day and wild-type animals treated for 7 days (both of 129Ola/CD1 background).

Table 1

Hybridizations performed in the CRH-R1 abolition experiment

Genotype (symbol)	Treatment (days)	Group ID	N
129Ola/CD1 knockout (k)	0	k0	4
129Ola/CD1 heterozygous (h)	0	h0	4
129Ola/CD1 heterozygous (h)	1	h1	5
129Ola/CD1 wild type (w)	0	w0	4
129Ola/CD1 wild type (w)	1	w1	4
129Ola/CD1 wild type (w)	7	w7	3
129SvJ wild type (s)	0	s0	4
129SvJ wild type (s)	1	s1	3
129SvJ wild type (s)	7	s7	2
Total			33

Thirty-three 12-week-old male mice with the genotypes shown in column 1 (k were CRH-R1 deficient animals, h and w were their littermates) were treated orally with an antagonist directed against the CRH-R1 (Janssen compound R121919) for 0, 1 or 7 days at a dose of 40 mg per kg body weight. Untreated groups (h0, w0, s0) were used to normalize the data and do not appear in Figures 5 and 6.

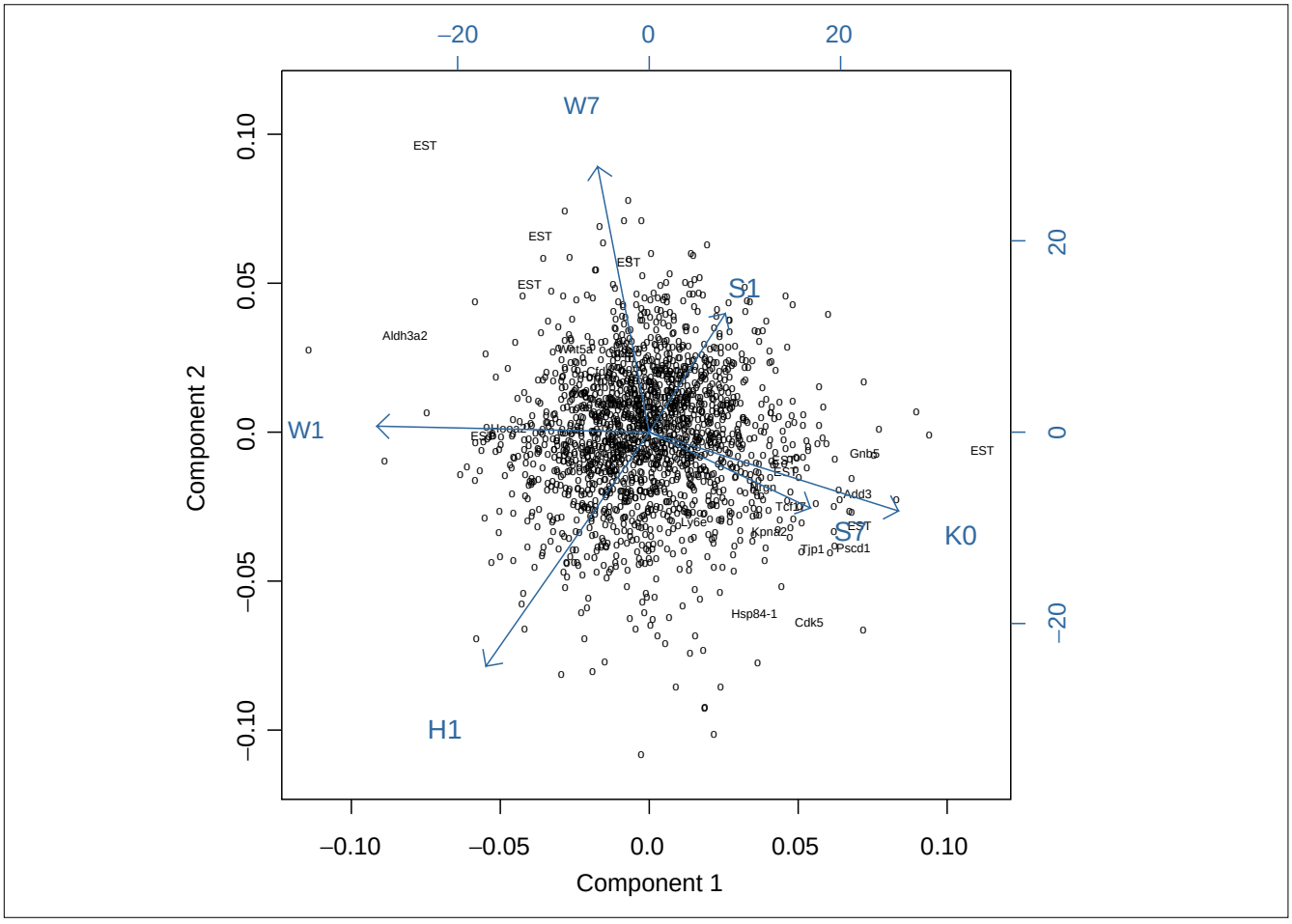

Figure 5
PCA of the grouped CRH-R1 function abolition data. Biplot of gene scores and treatment group loadings according to the first two components of the PCA. The open circles or gene names represent the gene scores. The vectors represent the treatment group (variable) loadings. The biplot enables the association of genes with the treatment groups. Labeled genes were selected by permutation test.

Because long-term-treated wild-type animals of 129SvJ background (s7) are similar to the knockout animals, treatment seems to have a strong effect in these animals. Animals with a 129Ola/CD1 background (group w7) show a weaker response to treatment with the antagonist. Both components describe abolition-of-function effects in a background-dependent manner. Thus, given a particular genetic background, treatment with an antagonist against CRH-R1 can mimic the genetic abolition of gene function. A comparable phenomenon was shown in yeast by Hughes *et al*. [29].

Only 25 genes were selected by permutation tests and are labeled with gene symbols in Figure 5. These genes show high variance across the treatment groups and are highly reproducible. Only genes that contrast the groups k0 and s7 on one side and in the groups w1 and w7 on the other side are selected. The profiles of these genes are illustrated in Figure 6 and support the interpretation of the biplot.

Application to antidepressant data

In this experiment, 29 12-week-old male mice of 129SvJ background were treated with mirtazapine, paroxetine or vehicle for 1, 7 or 28 days (Table 2). cDNA microarrays were used to measure the mRNA expression in total brain homogenates of these animals. The data consist of $\log_2$ ratios. The matrix had 2,190 complete observations (genes) and 24 hybridizations. We computed a PCA based on the polished matrix of single hybridizations to show that the treatment groups were ordinated together (data not shown). We performed an ANOVA using the variable loadings as dependent variables and the treatment groups as factors. The first two components were highly significant. They explained 36.3% of the data's variance and the object (gene) scores with respect to these two components were used to compute the variance of genes according to the group differences.

We carried out a PCA with group-averaged data and two components (72.1% explained variance). Figure 7 shows a

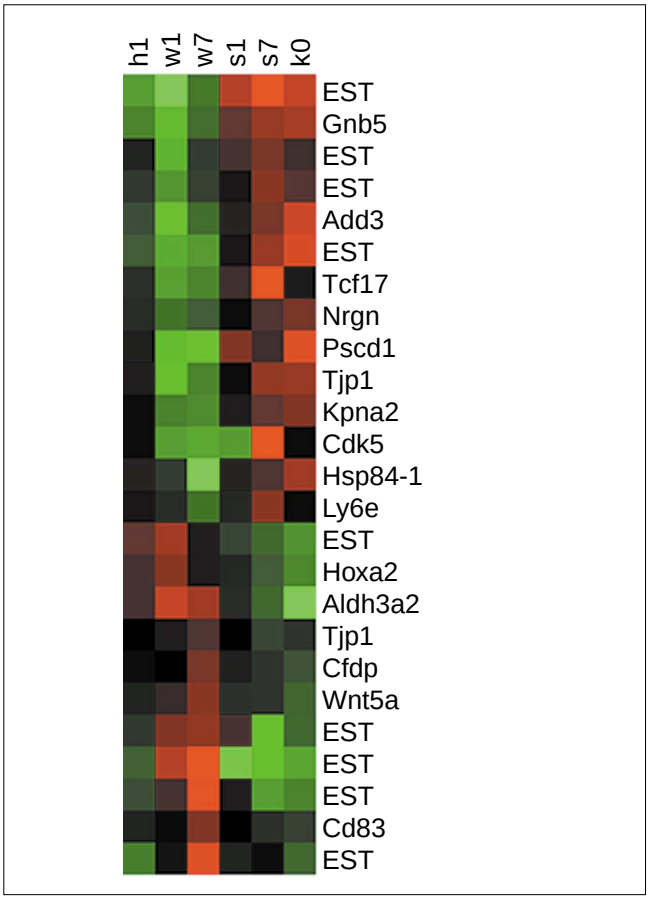

Figure 6
Color-coded expression-profile table of the genes selected from the CRH-R1 function abolition data by permutation test. Abscissa, treatment groups; ordinate, genes. The treatment groups and the genes are arranged according to the angular position in the PCA biplot (see Figure 5). Upregulation, red; downregulation, green.

biplot of these components with 127 selected genes at the 95% percentile (labeled with numbers) and 30 genes at the 99% percentile (labeled with gene symbols). The components describe the effects of antidepressant treatment on the mouse brain. The first component (abscissa) discerns treatment with mirtazapine (a) from treatment with paroxetine (p). This component can be interpreted as the drug-type effect. The second component discerns short (1 day) from longer (7 or 28 days) treatment and can be interpreted as the treatment duration effect. Figure 8 shows the 30 genes selected at the 99% percentile in a color-coded expression table. These genes strongly reflect the treatment type and duration effect.

Discussion

Here we propose a method for analyzing microarray data with group structure imposed by different conditions. We combine the visualization focused on the variance of genes between groups and gene selection, taking into account the within-group variance. Based on PCA, this method is able to visualize relationships between hybridizations by dimension reduction. Yet, data visualization via a biplot allows more than biological interpretation of the components. After appropriate data preprocessing, searching for genes with changes in expression patterns across the groups can be based on the genes' (objects') distance from the centroid of the biplot. This distance is proportional to the variance of genes in the dimension-reduced space. A correspondence analysis would give a similar result [14]. But a selection of genes must be accompanied by an assessment of whether the results are statistically reliable relative to the level of noise in the data. Whereas classic statistical tests (like *t*- and F-statistics) are based on assumptions concerning distribution and variable independence that do not hold for microarray data [1,19] the permutation-validation procedure presented here makes no assumption about the dependence of gene-expression measurement within the expression matrix **X**. Therefore, gene-specific scatter is taken into consideration by calculating the test-value permutation distributions for each gene under the null hypothesis of no group-structure effect in the expression profiles. Another method for validating PCA results using a leave-one-out approach (Wall *et al.* [22]) is very global, and can only be applied when the conditions correspond to a continuous parameter, such as time or dose.

The last step of the permutation-validated PCA procedure concerns the visualization and the interpretation of the selected genes according to their importance in a biological context. In the case of two dimensions ($k = 2$), a color-coded expression profile can be generated by rearranging the selected genes and the arrays with respect to angular distances in the biplot. When looking at a biplot showing several variable loadings, a given object (gene) has to be projected on all

Table 2

Design of the antidepressant experiment

Drug	Treatment (days)	Group ID	N
Mirtazapine	1	a1	3
Mirtazapine	7	a7	3
Mirtazapine	28	a28	5
Paroxetine	1	p1	5
Paroxetine	7	p7	5
Paroxetine	28	p28	3
Vehicle	28	c28	5
Total			29

Twenty-nine 12-week-old male mice of genotype 129SvJ were treated with the drug mirtazapine (a), paroxetine (p) or vehicle (c) by mouth for 1, 7 or 28 days at a dose of 40 mg per kg body weight. The untreated control group (c28) was used to normalize the data and does not appear in Figures 7 and 8.

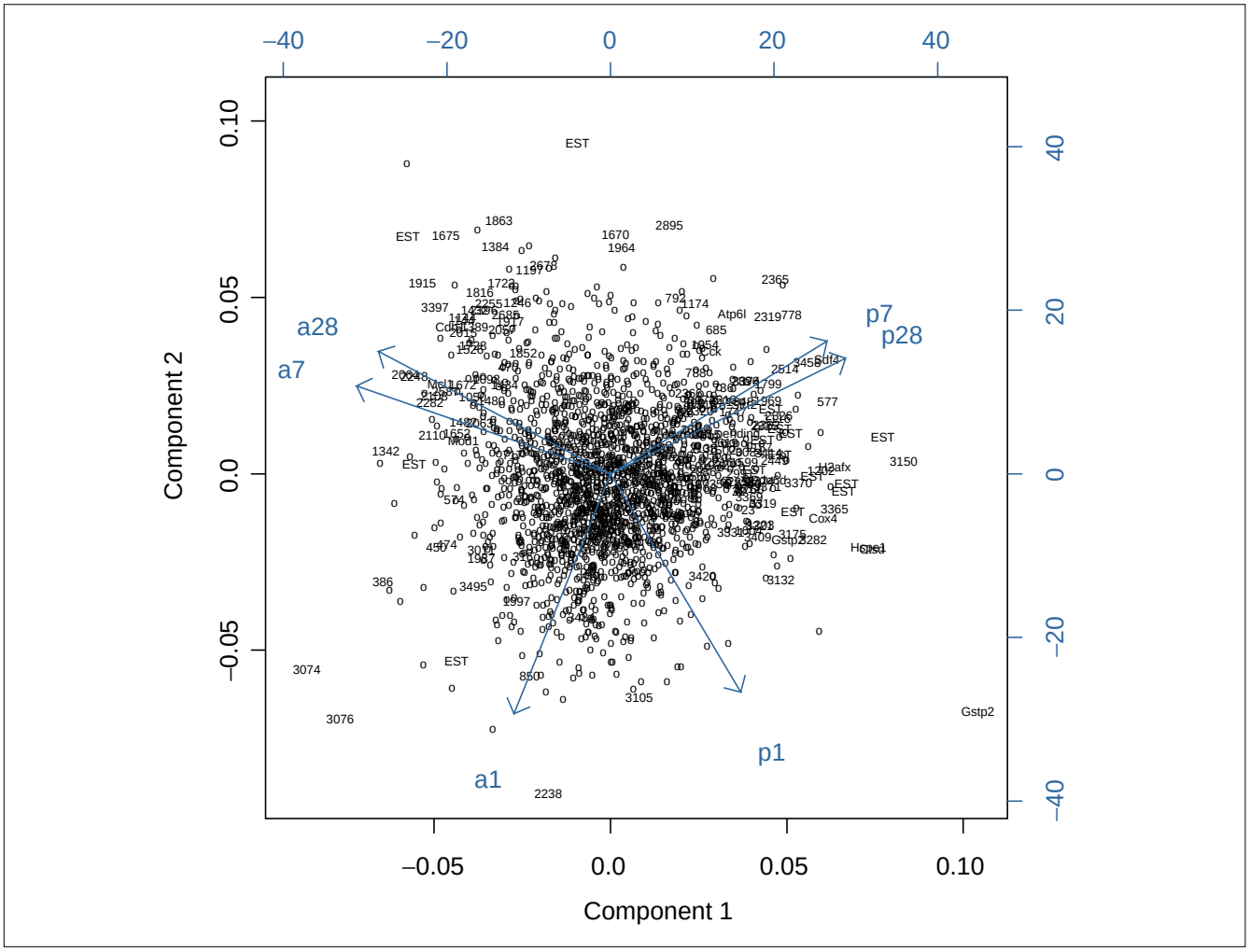

Figure 7
PCA of the grouped antidepressant data. Biplot of gene scores and treatment groups loadings according to the first two components of the PCA. The open circles represent the gene scores. The vectors represent the treatment groups (variable) loadings. The biplot enables the association of genes with the treatment groups. Genes labeled with numbers have test values above the 95% percentile of the permutation distribution; genes labeled with gene symbols have test values above the 99% percentile.

different variables (conditions) to understand its pattern with regard to all of them. A color-coded expression-profile table may support this visual interpretation. As a further development of the method described here, we envisage cluster analysis of the selected genes for higher dimensions ($k > 2$).

The application of permutation-validated PCA to microarray data shows that the basic sources of variance could be extracted from all datasets: The components computed from the Spellman *et al.* [23] yeast data described the cell cycle and allowed ordinations of the hybridizations according to their temporal situation in the cell cycle. Arrays misclassified by the Fourier transformation [23] were assigned to shifted positions in the cell cycle (this was also achieved by correspondence analysis [14]). The components computed from the abolition of CRH-R1 function experiment described a gradient of increasing functional impairment depending on the genetic background of the animals. The analysis of the antidepressant data also shows how principal components led to an understanding of the fundamental biological phenomena captured by the data: here, they discern the types of treatment and the treatment duration.

But there were important differences in the results: whereas the grouped PCA of the cell-cycle data explained 94.4% of the data's variance, the corresponding rates were 72.1% explained variance for the antidepressant data and 54.8% for the CRH-R1 abolition data. In a situation with homogeneous array groups and preselected genes such as the cell-cycle data, the level of explained variance was very high as the components explained the kind of variance the genes were preselected for. For the antidepressant data, no *a priori* information about

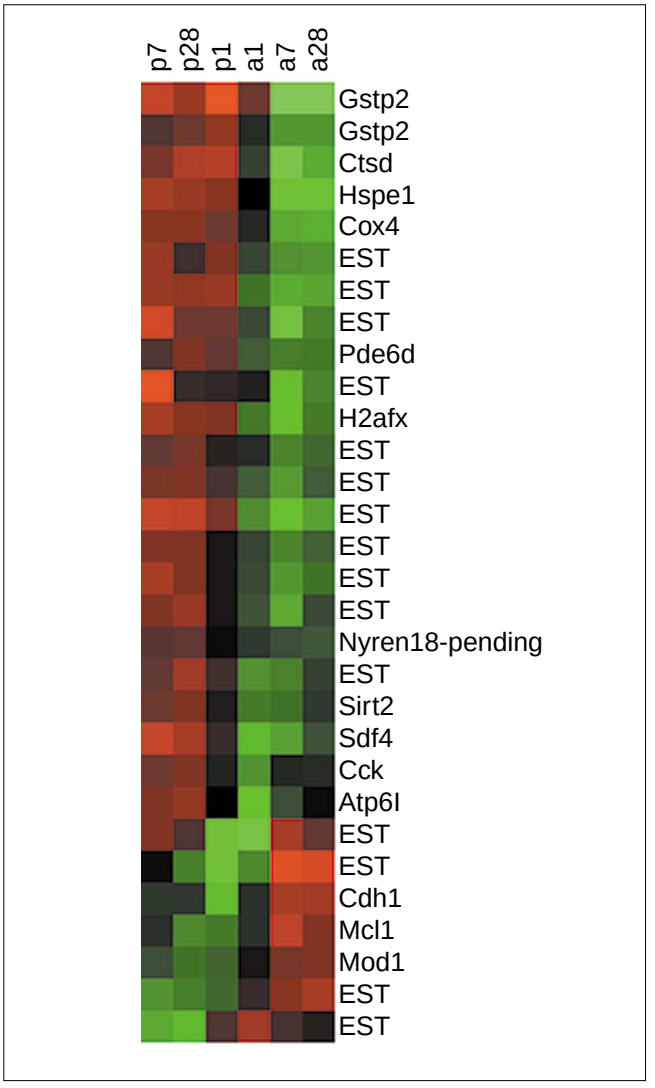

Figure 8
Color-coded expression-profile table of the genes selected from the antidepressant data by permutation test above the 99% percentile. Abscissa, treatment groups; ordinate, genes. The treatment groups and the genes are arranged according to the angular position in the PCA biplot (see Figure 7). Upregulation, red; downregulation, green.

the relation of genes to treatment type and duration was available. Thus the level of explained variance was lower (72.1%), although the two components used to build the test statistics still captured a big part of the variance present in the data. Although the material used for the antidepressant treatment data (RNA from total mouse brain homogenates containing a variety of cell types) was more heterogeneous than the clonal yeast cell lines, antidepressant effects on the brain's mRNA transcription were so important that clear variance structures emerged in the data. In contrast, the level of explained variance for the CRH-R1 function abolition data was 54.8%. In this experiment, different methods of impairing or abolishing the function of CRH-R1 were used on different mouse genetic backgrounds. The variance in the experimental design was thus quite high. Heterogeneity within the groups was also higher than in the antidepressant experiment, probably because a selective pharmacologic antagonization of the neuromodulating peptide CRH [27] had a less pronounced and stereotypic effect on transcription in the brain than the antidepressant drugs acting on a wide spectrum of neurotransmitter receptors, transporters and related enzymes [30]. In this situation of high variance in the experimental design, and a relatively high rate of heterogeneity in the treatment groups, permutation-validated PCA only selected genes reflecting the contrast between the groups w1, w7 on one side and s7 and ko on the other side because this contrast was captured by the first two components. Other aspects of the data were not captured by PCA. Thus, a multivariate approach trying to compare very different gene expression patterns at the same time might lead to loss of information. In such a case, the selection of genes should be treated with caution and cross-validation by independent methods should be applied if hypotheses are to be derived from the selected genes. Pairwise comparisons of groups might be more appropriate in such a situation.

In conclusion, permutation-validated PCA can be used to extract the leading source of variance from microarray data, to visualize relationships between genes and hybridizations and to select informative genes in a statistically reliable manner. This selection accounts for the level of reproducibility of replicates or group structure as well as gene-specific scatter.

Materials and methods

Sample processing and hybridization

A subset of the data from Spellman *et al.* [23] was used. To acquire our own data, microarrays were manufactured, mice treated and total brain RNA extracted, labeled and hybridized as described in [18]. Briefly, mice were killed after the end of treatment, RNA was extracted by RNeasy and TRIZOL procedure. Total RNA (100 μg) was fluorescence-labeled by oligo-dT-primed reverse transcription to cDNA in the presence of Cy3-dUTP as described by Eisen and Brown [31]. After reverse transcription, total brain Cy3-labeled cDNA from each animal was hybridized to a microarray. Fluorescence intensity was detected using the Genetic Microsystems GMS 418 Array Scanner. Raw data were assessed with the Spectrum vs.3.2 image-analysis software developed by Chen *et al.* [32].

Data preprocessing

Data from Spellman *et al.* [23] were also used by Fellenberg *et al.* [14]; we did not modify the described preprocessing. The two datasets from our lab were preprocessed in the following manner. Matrix rows (genes) with missing observations were excluded from the datasets, resulting in data without missing values. To normalize and compare the different hybridizations to each other, the intensity measured at each spot of the arrays was divided by the centered median of the intensities measured at the corresponding spot in the reference groups. Thus,

every single hybridization was normalized against the reference groups by computing the $\log_2$ of the ratios (the mean of groups s0, h0 and w0 for the CRH-R1 data and group c28 for the antidepressant data). Therefore, these groups do not appear in Figures 5 to 8. Given an n x p data matrix, the following model [1,2] can be stated:

$$X_{gj} = \mu + \alpha_g + \beta_j + \delta_{gj} + \varepsilon_{gj}$$

In this model, X_{gi} is the log-ratio of gene g under experimental condition j, α_g is the normalizing effect for gene g (row), β_j is the experimental variance effect for j (column), δ_{gj} is the differential gene expression for gene g under experimental condition j and ε_{gj} is the random error.

To estimate the interaction term δ_{gj}, several other effects must be controlled: as α_g reflects the relation of experiment RNA to normalizing RNA and is of no biological interest, it can be controlled by mean centering rows. β_j reflects the global variance in RNA preparation, labeling efficacy and hybridization quality as well as other sources of experimental variance between the arrays and can be controlled by standardizing the matrix columns. Doing replicates enables control of ε_{gj} . The term δ_{gj} can thus be obtained by data polishing [9], that is, the matrix is iteratively subjected to column standardization and row mean centering until convergence is reached. This polished matrix was used as the basis for multivariate analysis.

Acknowledgements

We thank Claudia Kühne for technical assistance. We thank the GSF-Research Center, the Max-Planck-Gesellschaft and the Volkswagenstiftung for funding.

References

1. Kerr M, Martin M, Churchill G: **Analysis of variance in microarray data.** *J Comp Biol* 2000, **7:**819-837.
2. Ting Lee M, Kuo C, Whitnore G, Sklar F: **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridization.** *Proc Natl Acad Sci USA* 2000, **97:**9834-9839.
3. Dudoit S, Fridlyand J, Speed T: **Comparison of discrimination methods for the classification of tumours by using gene expression data expression data processing and modeling.** *Technical Report 576*, Berkeley, CA: University of California at Berkeley, 2000. Available at [http://www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html]
4. Brown M, Grundy W, Lin D, Cristianini N, Sugnet C, Furey T, Ares MJ, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci USA* 2000, **97:**262-267.
5. Eisen M, Spellman P, Brown P, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
6. Tavazoie S, Hughes J, Campbell M, Cho R, Church G: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285.
7. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E, Golub T: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96:**2907-2912.
8. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, *et al.*: **Molecular classification of cutaneous malignant melanoma by gene expression profiling.** *Nature* 2000, **406:**536-540.
9. Holter N, Mitra M, Maritan A, Cieplak M, Banavar J, Fedoroff N: **Fundamental patterns underlying gene expression profiles: simplicity from complexity.** *Proc Natl Acad Sci USA* 2000, **97:**8409-8414.
10. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan W, Botstein D, Brown P: **Gene shaving as a method for identifying distinct sets of genes with similar expression patterns.** *Genome Biol* 2000, **1:**research0003.1-0003.21.
11. Alter O, Brown P, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling**. *Proc Natl Acad Sci USA* 2000, **97:**10101-10106.
12. Hilsenbeck S, Friedrichs W, Schiff R, O'Connell P, Hansen R, Osborne C, Fuqua S: **Statistical analysis of array expression data as applied to the problem of tamoxifen resistance.** *J Natl Cancer Inst* 1999, **91:**453-459.
13. Raychaudhuri S, Stuart J, Altman R: **Principal components analysis to summarize micorarray experiments: application to sporulation time series.** *Pac Symp Biocomput* 2000:455-466.
14. Fellenberg K, Hauser N, Brors B, Neutzner A, Hoheisel J, Vingron M: **Correspondence analysis applied to microarray data.** *Proc Natl Acad Sci USA* 2001, **98:**10781-10786.
15. Krzanowski W: *Principles of Multivariate Analysis*. Oxford: Oxford University Press, 2000.
16. Hastie T, Tibshirani R, Botstein D, Brown P: **Supervised harvesting of expression trees.** *Genome Biology* 2001, **2:**research0003.1-0003.12.
17. Tusher V, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98:**5116-5121.
18. Landgrebe J, Welzl G, Metz T, van Gaalen M, Ropers H, Holsboer F, Wurst W: **Molecular characterization of antidepressant effects in the mouse brain using gene expression profiling.** *J Psychiat Res* 2002, **36:**119-129.
19. Dudoit S, Yang Y, Callow MJ, Speed T: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** *Technical Report 578*. Berkeley, CA: University of California at Berkeley, 2000. [http://www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html]
20. Efron B: **The bootstrap and modern statistics.** *J Amer Stat Assoc* 2000, **95:**1293-1296.
21. Ghosh D, Chinnaiyan AM: **Mixture modelling of gene expression data from microarray experiments.** *Bioinformatics* 2002, **18:**275. [http://www.sph.umich.edu/~ghoshd/COMPBIO/mixture1/]
22. Wall M, Dyck P, Brettin T: **SVDMAN — singular value decomposition analysis of microarray data.** *Bioinformatics* 2001, **17:**566-568.
23. Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.
24. Zhao L, Prentice R, Breeden L: **Statistical modeling of large microarray data sets to identify stimulus-response profiles.** *Proc Natl Acad Sci USA* 2001, **98:**5631-5636.
25. Krzanowski W: **Ranking principal components to reflect group structure.** *J Chemometrics* 1992, **9:**509-520.
26. Keck M, Welt T, Wigger A, Renner U, Engelmann M, Holsboer F, Landgraf R: **The anxiolytic effect of the CRH1 receptor antagonist R121919 depends on innate emotionality in rats.** *J Neurosci* 2001, **13:**373-380.
27. De Souza E: **Corticotropin-releasing factor receptors: physiology, pharmacology, biochemistry and role in central nervous system and immune disorders.** *Psychoneuroendocrinology* 1995, **20:**789-819.
28. Timpl P, Spanagel R, Sillaber I, Kresse A, Reul J, Stalla G, Blanquet V, Steckler T, Holsboer F, Wurst W: **Impaired stress response and reduced anxiety in mice lacking a functional crh-r1.** *Nat Genet* 1998, **19:**162-166.
29. Hughes T, Marton M, Jones A, Roberts C, Stoughton R, Armour C, Bennett H, Coffey E, Dai H, He Y, *et al.*: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**102-126.
30. Duman R, Heninger G, Nestler E: **A molecular and cellular theory of depression.** *Arch Gen Psychiatry* 1997, **54:**597-606.
31. Eisen M, Brown P: **DNA arrays for analysis of gene expression.** *Methods Enzymol* 1999, **303:**179-205.
32. Chen Y, Dougherty E, Bittner M: **Ratio-based decisions and the quantitative analysis of cDNA micro-array images.** *J Biomed Optics* 1997, **2:**364-374.