Bioinformatics Advance Access published February 25, 2005

© The Author (2005). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

Bioinformatics: Application note.

A web-based tool for principal component and significance analysis of microarray data

Alexei A. Sharov, Dawood B. Dudekula, and Minoru S.H. Ko*

Developmental Genomics and Aging Section, Laboratory of Genetics, National Institute on Aging, National Institutes of Health, Baltimore, MD 21224, USA.

*To whom correspondence should be addressed.

Running title: Microarray analysis tool

Keywords: PCA, SVD, clustering, FDR, microarrays, gene expression

ABSTRACT

Summary: We have developed a program for microarray data analysis, which features the False Discovery Rate (FDR) for testing statistical significance and the Principal Component Analysis (PCA) using the Singular Value Decomposition (SVD) method for detecting the global trends of gene expression patterns. Additional features include analysis of variance (ANOVA) with multiple methods for error variance adjustment, correction of cross-channel correlation for two-color microarrays, identification of genes specific to each cluster of tissue samples, biplot of tissues and corresponding tissue-specific genes, clustering of genes that are correlated with each principal component (PC), 3-dimensional graphics based on virtual reality modeling language (VRML), and sharing of PC between different experiments. The software also supports parameter adjustment, gene search, and graphical output of results. The software is implemented as a web tool, and thus the speed of analysis does not depend on the power of a client computer.

Availability: The tool can be used on-line or downloaded at http://lgsun.grc.nia.nih.gov/ANOVA/.

Contact: kom@mail.nih.gov

Global gene expression analysis with microarrays becomes a routine procedure in biomedical research. Although many programs have been developed to support the statistical analysis of microarray results (e.g., Kim*, et al* 2001, Theilhaber*, et al* 2004, TIGR 2004, Tusher*, et al* 2001), they do not necessarily contain all the advanced analysis methods. To facilitate the use of these relatively new methods, we developed NIA Array Analysis software. Full description of the software as well as the glossary of technical and statistical terms can be found at http://lgsun.grc.nia.nih.gov/ANOVA/. Here we describe the main features of this software.

The NIA Array Analysis software can be used for both single-color and two-color microarrays with or without a dye-swap. It uses a tab-delimited text file as an input and generates outputs in both graphics and

text formats. An additional tool (Arrayjoin) assembles multiple input files from different experiments into one input file. The software can also take an annotation file, which hyperlinks each microarray probe to various web resources, including Unigene, TIGR, MGI, and NIA Mouse Gene Index. These gene links allow users to incorporate microarray data into other programs, e.g., the GenMAPP for Gene Ontology analysis. All results can be saved as a stand-alone web page for sharing or releasing the data.

The software offers an optional adjustment of signal intensities, when two-color hybridizations are used. This is based on our observation that signal intensities in one channel (e.g., red) often increase with the increasing signal intensities in the other channel (e.g., green), even when the same reference RNA is always used for the red channel. If readings from these two channels are independent, signal intensities in red channel should not vary among experiments and should be corrected if it changes.

We have implemented the single-factor analysis of variance (ANOVA) for testing statistical significance. Testing multiple hypotheses with the ANOVA requires some modifications such as error variance averaging and False Discovery Rate (FDR). The average error variance for genes with similar signal intensities is estimated using the sliding window of adjustable size applied to genes sorted by their average signal intensities. Because some genes (outliers) may have unusually high error variance, genes with the highest variance values (a top 1% by default) are not used for the error variance averaging. To obtain an estimate for the true error variance, the software provides the following 5 different error models as options: (i) actual error variance (this option processes each gene independently); (ii) intensity-specific average error variance; (iii) Bayesian error model (Baldi and Long 2001); (iv) maximum between intensity-specific average error variance and actual error variance; and (v) maximum between intensity-specific average error variance and Bayesian error variances. The option (iv), the most conservative model, is used as default. However, if error variance is too high, then none of these models is reliable. Thus, we tag and visually examine genes with high error variance (5 times greater than the average). Users can also select more stringent criteria for removing outliers (i.e., a lower *z*-threshold level). The

default threshold ($z = 8$) removes only the most deviating outliers. Estimation of the z-value is based on the ANOVA results; thus, ANOVA is applied iteratively with outlier removal in each cycle until no new outliers are detected.

The FDR identifies the proportion of false-positives among significant genes (Benjamini and Hochberg 1995, Reiner*, et al* 2003). Traditional *p*-values, which are designed for testing a single hypothesis, are not suited to the comparison of several thousand genes. The Bonferroni correction is not relevant either, because it is too stringent and allows no false positives among significant genes. We have implemented the original method (Benjamini and Hochberg 1995):

$$FDR_r = \min_{i \geq r} \left[\frac{p_i N}{i} \right], \quad (1)$$

where r is the rank of a gene ordered by increasing p-values, p_i is the p-value for a gene with rank i, and N is the total number of genes tested. It indicates the proportion of false positives among all genes with p-values lower or equal to the p-value of the gene of interest.

The software offers two methods of clustering tissue samples and subsequent identification of correlated genes. First, hierarchical clustering of samples (e.g., tissues and cells) is done by using the average distance method. A set of genes that are unique to each cluster is identified in the following manner. For each gene, g, we first identify a sample $T_1(g)$ with the lowest average expression, $E(T_1(g))$, within the cluster, and a sample $T_2(g)$ with the highest average expression, $E(T_2(g))$, outside the cluster. If K genes satisfy $E(T_1(g)) > E(T_2(g))$, these genes always have higher expressions in samples within the cluster than in all samples outside the cluster. To determine if the difference $E(T_1(g)) - E(T_2(g))$ is statistically significant, we calculate z-values based on the error model, and p-values based on single-tail normal distribution. Finally we estimate FDR values using equation (1), in which N is the minimum between $2K$ and the total number of genes. The set of K genes represents only a half (the positive part) of the normal distribution, and thus K is doubled for estimating the FDR.

Second, we have implemented the Principal Component Analysis (PCA). One advantage of PCA is that the principal components are always orthogonal (uncorrelated), whereas other methods (e.g., *K*-means clustering) often produce redundant correlated clusters. We have also implemented the Singular Value Decomposition (SVD) method, which reduces the dimension in both columns and rows of the data matrix. The method combines samples and genes in a single graph (called "biplot") so that their association can be analyzed visually (Chapman*, et al* 2002, Gabriel 1971). The NIA Array Analysis tool generates interactive 2-dimensional and 3-dimensional biplots (Figure 1). Each gene in a biplot is hyperlinked to its annotation and histogram showing the expression levels in each sample. We identify two sets of genes that are positively and negatively correlated with each principal component (PC). If the degree of a gene expression change associated with a specific PC exceeds a user-defined threshold, then the gene is considered correlated with the PC.

The NIA Array Analysis tool has been successfully used for last two years (e.g., Hamatani*, et al* 2004, Sharov*, et al* 2003). This open-source non-restricted software will be a valuable resource for the research community.

REFERENCES

Baldi, P. and Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics,* **17,** 509-519.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society, B,* **57,** 289-300.

Chapman, S., *et al.* (2002) Using biplots to interpret gene expression patterns in plants. *Bioinformatics,* **18,** 202-204.

Gabriel, R. (1971) The biplot graphical display of matrices with application to principal component analysis. *Biometrika,* **58,** 453-467.

Hamatani, T., *et al.* (2004) Dynamics of global gene expression changes during mouse preimplantation development. *Dev Cell,* **6,** 117-131.

Kim, S.K., *et al.*. (2001) A gene expression map for Caenorhabditis elegans. *Science,* **293,** 2087-2092.

Reiner, A., *et al.* (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics,* **19,** 368-375.

Sharov, A.A., *et al.* (2003) Transcriptome analysis of mouse stem cells and early embryos. *PLoS Biol,* **1,** E74.

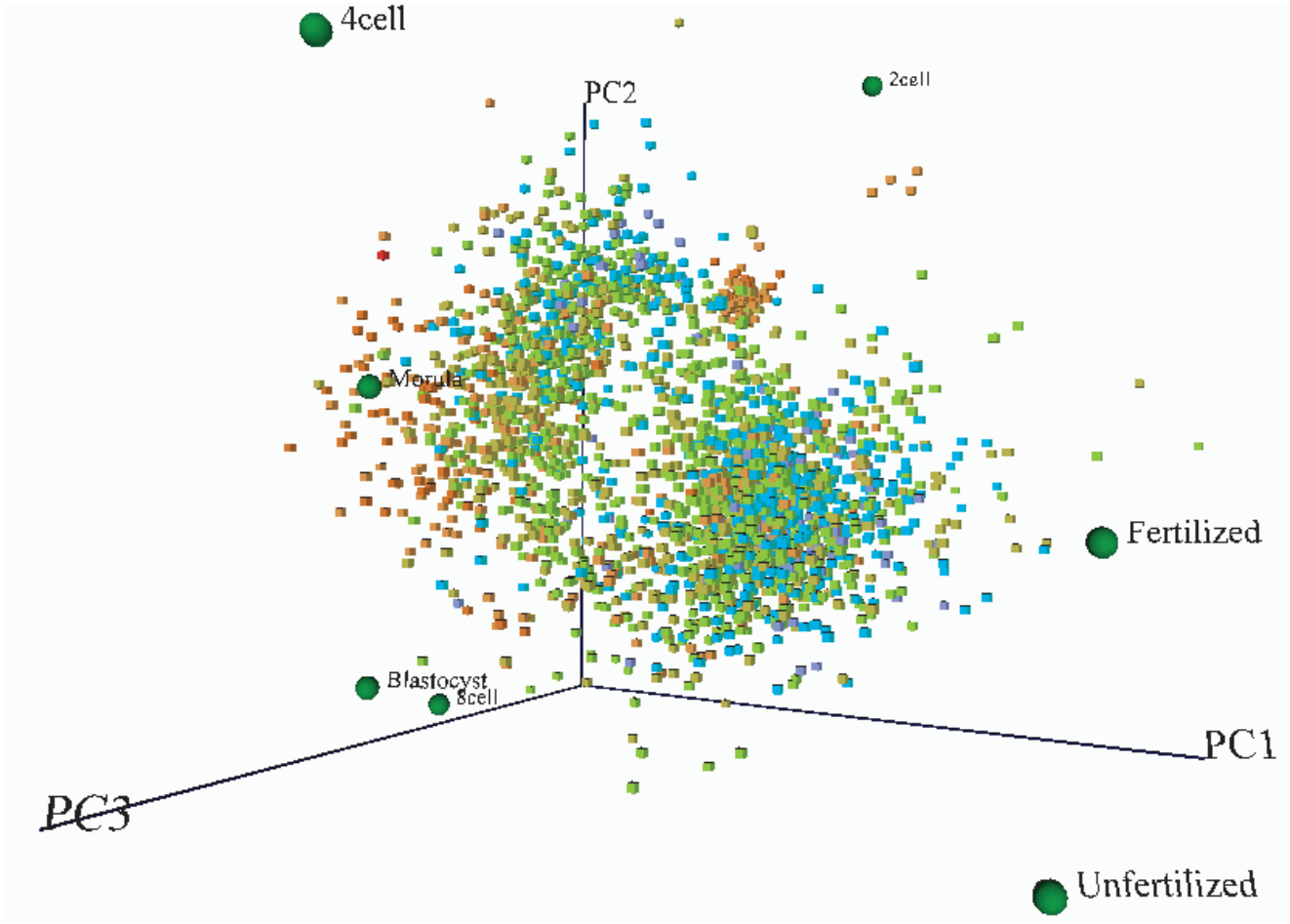
Theilhaber, J., *et al.* (2004) GECKO: a complete large-scale gene expression analysis platform. *BMC Bioinformatics,* **5,** 195.

TIGR (2004) TM4 microarray software suite. *http://www.tigr.org/software/tm4/*.

Tusher, V.G., *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A,* **98,** 5116-5121.

Figure Legends

Fig. 1. 3D biplot of gene expression during preimplantation mouse development (data from Hamatani et al. 2004).


4cell
PC2
2cell
Morula
Fertilized
Blastocyst
8cell
PC1
PC3
Unfertilized