VARAN: a web server for Variability Analysis of DNA microarray experiments

G. Golfier*, M. Tran Dang, L. Dauphinot, E. Graison, J. Rossier and M.-C. Potier

Neurobiologie et Diversité Cellulaire, CNRS UMR7637, Ecole Supérieure de Physique et de Chimie Industrielles, 10 rue Vauquelin, 75005 Paris, France

ABSTRACT

Summary: Here, we describe a tool for VARiability Analysis of DNA microarrays experiments (VARAN), a freely available Web server that performs a signal intensity based analysis of the log2 expression ratio variability deduced from DNA microarray data (one or two channels). Two modules are proposed: VARAN generator to compute a sliding windows analysis of the experimental variability (mean and SD) and VARAN analyzer to compare experimental data with an asymptotic variability model previously built with the generator module from control experiments. Both modules provide normalized intensity signals with five possible methods, log ratio values and a list of genes showing significant variations between conditions.

Availability: http://www.bionet.espci.fr/varan/

Contact: geoffroy.golfier@espci.fr

Supplementary information: http://www.bionet.espci.fr/varan/help.html

INTRODUCTION

The first step into DNA microarray analysis is the normalization of intensity signals. Various methods are then applied to select genes significantly differentially expressed. Since experimental variability is higher for a low signal intensity range than for a higher one, where fluorescence signal intensity measurements are more reliable, basic gene selection prior to data mining can be biased by keeping or eliminating genes that are, respectively, not significantly or significantly expressed regarding signal intensity or gene by gene expected variations. Furthermore, the intensity-related variability may introduce biases when clustering together genes associated to different magnitudes of signal intensity. This problem can be avoided by replicating experiments, eliminating low-intensity signal associated genes, applying variance stabilization methods (Huber et al., 2002; Rocke and Durbin, 2003) or performing sliding windows intensity analysis and building specific error models (Quackenbush, 2002; Hughes et al., 2000).

Here, we propose a Web tool allowing data normalization, generating and visualizing variability models predicting variability as a function of mean intensity that can be applied to datasets for rapidly selecting differentially expressed genes. This tool can also be used to transform log expression ratios into scaled log ratios (scaled fold option) in order to facilitate comparison (e.g. clustering and data mining) of genes associated with either low or high signal intensities.

SERVER OVERVIEW

Two types of analysis can be made with the VARAN server using either the generator or the analyzer module for, respectively, building variability models or comparing log ratio data with a variability model corresponding to the same conditions or protocols. Data before background correction (raw data) originating from any type of microarray (one or two channels) can be submitted in a simple specific text tabulated format containing six or seven columns, depending on the normalization method selected. Each row must report a gene ID, total and background intensity signals for the two conditions (e.g. Cy3 and Cy5 channels), a block flag for either print-tip normalization or files containing independent experiments or both and a non-mandatory flag restricting the normalization correction factor computation to a subset of reference genes. For Affymetrix®, the data background should be set to 0. The VARAN server uses Perl (http://www.perl.org) routines to create user adapted R (http://www.r-project.org) scripts from which results are accessible through an HTML interface.

RAW DATA PREPROCESSING

Data preprocessing is common to both generator and analyzer modules and includes for each gene the following: background correction, calculation of normalized specific signals, log2 ratios (M) and geometric means of Cy3 and Cy5 log2 specific signals (A). The default method for the Cy3 versus Cy5 normalization uses the R environment lowess function to correct

*To whom correspondence should be addressed.
Fig. 1. Example of graphical outputs from VARAN generator and VARAN analyzer. The public data set used comes from the comparison of gene expression in mouse embryo and placenta at embryonic day 12.5 (E12.5) (Carter et al., 2003). Experiments have been performed using the NIA 22K 60mer Oligo Microarray (Laboratory of Genetics—NIA). (A) Two plots from VARAN generator: log2 expression ratio versus geometric mean of log2 signal intensities (M versus A plot; window calculated quantiles, q_R and q_G, are drawn as triangles and used to perform two asymptotic fits generating the variability area delimited by both q_R and q_G asymptotes) and scaled fold plot (each log2 ratio value is divided by the absolute value of the q_G or the q_R asymptote at the corresponding A value for, respectively, a negative or a positive log2 ratio) at the top and bottom, respectively. The variability model was generated with three control experiments (placenta versus placenta) using the VARAN generator module. G stands for Green or Cy3 (condition 1) and R for Red or Cy5 (condition 2). In these control experiments R and G correspond to placenta RNA. Top figure (M versus A plot): the upper right box displays the parameters used to build the variability model, the bottom left box shows that normal distributions have been used to compute the quantiles. (B) Two plots from the VARAN analyzer module: M versus A plot (top) and scaled fold plot (bottom). The variability model from (A) was used to analyze differential expression between placenta and embryo at E12.5 using the VARAN analyzer. R and G correspond to embryo and placenta RNA, respectively. Genes significantly overexpressed in the embryo as compared with placenta are above the q_R asymptote on the M versus A plot and above 1 on the scaled fold plot (horizontal red line). Genes significantly underexpressed in the embryo are below the q_G asymptote on the M versus A plot and below −1 on the scaled fold plot (horizontal green line).

locally each log2 ratio value regarding A (Cleveland, 1979). The user can define the smoothness of the fit by specifying the proportion of values that will influence the regression at each point. Other normalization methods are available that perform Cy3 correction by either dividing the sum of Cy5 intensity signals by the sum of Cy3 intensity signals over the whole array or blocks or by applying a linear least square regression on the intensity signals for a subset of reference genes (specified in the seventh column of the input file). One can also choose to perform independent linear scaling of log2
intensity signals for the two conditions (Fielden et al., 2002), and for this purpose global and Z-score methods are available, the Z-score method reducing the variance of the log2 intensity signals to 1. More details are given in the help page of the VARAN server.

Finally the user can define the proportion of each Cy3 and Cy5 signal intensity range to be used for the correction factor calculation, thus excluding genes showing the lowest and the highest intensity signal values. This parameter does not affect both lowess fit and linear regression over reference gene normalization.

DATA SUBMISSION

There are basically two ways of submitting data to the VARAN server. If the experimenter has performed several control experiments (same RNA labeled with Cy3 and Cy5 or replicates of Affymetrix® chips), these sets of data can be loaded on the VARAN generator. The variability model generated can be saved and further used in the VARAN analyzer module for analyzing differential experiments. One model will be specific for one type of DNA microarray (group of genes), one sample (tissue, cell type) and one microarray platform and should be generated from several control experiments performed on different arrays from various batches. If the user changes one of these parameters (type of array, sample, platform), a new model needs to be generated from a new set of control experiments.

If the experimenter has not performed control experiments, the VARAN generator can thus be used to generate the asymptotic model as well as to run analyses of datasets from differential experiments. In this particular case, datasets should be obtained from a relatively dense microarray where a low proportion of differentially expressed genes is expected.

VARAN GENERATOR

The VARAN generator module computes the asymptotic variability model based on a sliding windows analysis of the M versus A distribution. The confidence interval (CI) that will define the variability area is set by the user. Each window refers to the log2 ratio values included in a short range of A. The way the windows will cover the whole M versus A distribution is defined by setting the number of windows or the number of log2 ratios per window. Inside each window, the distribution of log2 ratios is used to compute two quantiles (q_G and q_R) corresponding to the probabilities of 1−(C.I./2) (q_G) and 1−[1−(C.I./2)] (q_R) regarding a normal distribution or a t-distribution defined by the local mean and SD of the log2 ratios. A least mean square regression is then applied over both sets of q_G and q_R values in order to compute two asymptotes defining the variability model. For each gene, a scaled fold value is then calculated by dividing the log2 ratio by the absolute value of the q_G or the q_R asymptote at the corresponding A value for, respectively, a negative or a positive log2 ratio.

Available results of this explorative analysis of the M versus A distribution include the following: a M versus A plot, (Fig. 1), normalized data (intensity signals, M and A values and scaled fold transformed log2 ratios), a file defining the parameters of the asymptotic models evaluated and, if specified, a list of genes outside the variability area. However, data obtained from relatively dense microarray and where a low proportion of differentially expressed genes is expected are also suitable for generating a variability model.

VARAN ANALYZER

VARAN analyzer compares the M versus A distribution deduced from a DNA microarray data file with a variability model obtained with the generator module in order to select genes for which log2 ratios are not included in the variability area and to apply the scaled fold transformation. The file containing the variability model is also used to define all the normalization parameters to be used. The results obtained include the M versus A plot, normalized data and the list of genes significantly outside the variability area.

ACKNOWLEDGEMENTS

The authors wish to thank GeneScore company for providing DNA microarrays. This project was supported by EEC grant BMH4-CT96-0554 and the Fondation Jérôme Lejeune. G.G. has a fellowship from the Fondation Jérôme Lejeune.

REFERENCES


