SNOMAD (Standardization and NOrmalization of MicroArray Data): web-accessible gene expression data analysis

Carlo Colantuoni[1, 2, 3], George Henry[1,†], Scott Zeger[3] and Jonathan Pevsner[1, 2,*]

[1]Department of Neurology, Kennedy Krieger Institute, 707 North Broadway, Baltimore, MD 21205, USA, [2]Department of Neuroscience, Johns Hopkins University School of Medicine, 725 N. Wolfe Street, Baltimore, MD 21205, USA and [3]Department of Biostatistics, Johns Hopkins School of Public Health, Baltimore, MD 21205, USA

Received on September 11, 2001; revised on February 18, 2002; accepted on April 25, 2002

ABSTRACT

Summary: SNOMAD is a collection of algorithms for the normalization and standardization of gene expression datasets derived from diverse biological and technological sources. In addition to conventional transformations and visualization tools, SNOMAD includes two non-linear transformations which correct for bias and variance which are non-uniformly distributed across the range of microarray element signal intensities: (1) Local mean normalization; and (2) Local variance correction (Z-score generation using a locally calculated standard deviation).

Availability: The SNOMAD tools were developed in the R statistical language (http://www.r-project.org/). SNOMAD is an interactive, user-friendly web-application which can be accessed freely via the internet with any standard HTML browser: http://pevsnerlab.kennedykrieger.org/snomad.htm.

Contact: ccolantu@jhmi.edu or pevsner@jhmi.edu

The target of gene expression analysis, i.e. variation in actual gene expression levels, can be obscured by the many sources of variation inherent in microarray technologies. These sources can produce artifactual noise and/or bias in gene expression data. This complicates gene expression analysis within individual datasets as well as between datasets derived from diverse technologies and biological systems. Several groups have addressed basic normalization processes (see web site for references). Terrence Speed and collaborators have developed methods for the normalization of data both within and between hybridization experiments (http://www.stat.Berkeley.EDU/users/terry/zarray/Html/papersindex.html). Rocke and Durbin (2001) have addressed the differential levels of variance in differential gene expression observations at different absolute gene expression levels.

A number of plots are invaluable for the transformation, quality control, and analysis of gene expression datasets (Figure 1a). The plot of gene expression intensity versus gene expression ratio (Figure 1a, Right) is a plot which effectively shows differences in gene expression (Y axis) across the range of gene expression levels (X axis). The normalization processes detailed in this report are carried out across the axes of this plot, and are therefore designed for application to two-channel data or paired one-channel data (the Y-axis involves ratios, hence requiring two sets of element intensities).

Most investigators apply a global mean normalization to raw intensities, ensuring that the mean ratio for all array elements will be equivalent across multiple microarray experiments. However, much artifactual variation present in gene expression data is not constant across the range of element signal intensities, and hence cannot be addressed by global normalization. We present two local normalizations which address bias and variance which are non-uniformly distributed across absolute signal intensity.

The first is a local mean normalization which corrects for systematic bias in gene expression ratios between two samples. In gene expression datasets, the mean expression ratio often deviates significantly from a value of one (Figure 1a, last panel). We use a robust local regression ('loess' in the R statistical language) to calculate the local mean gene expression ratio as it varies across the range of gene expression intensity (Figure 1b, gray line Top). The distance of each gene expression ratio from this local mean (Figure 1b, Top 'residuals') is then used as the corrected ratio. This process ensures a mean intensity ratio of one

[*]To whom correspondence should be addressed.
[†] Present address: Wadham College, Oxford University, UK.

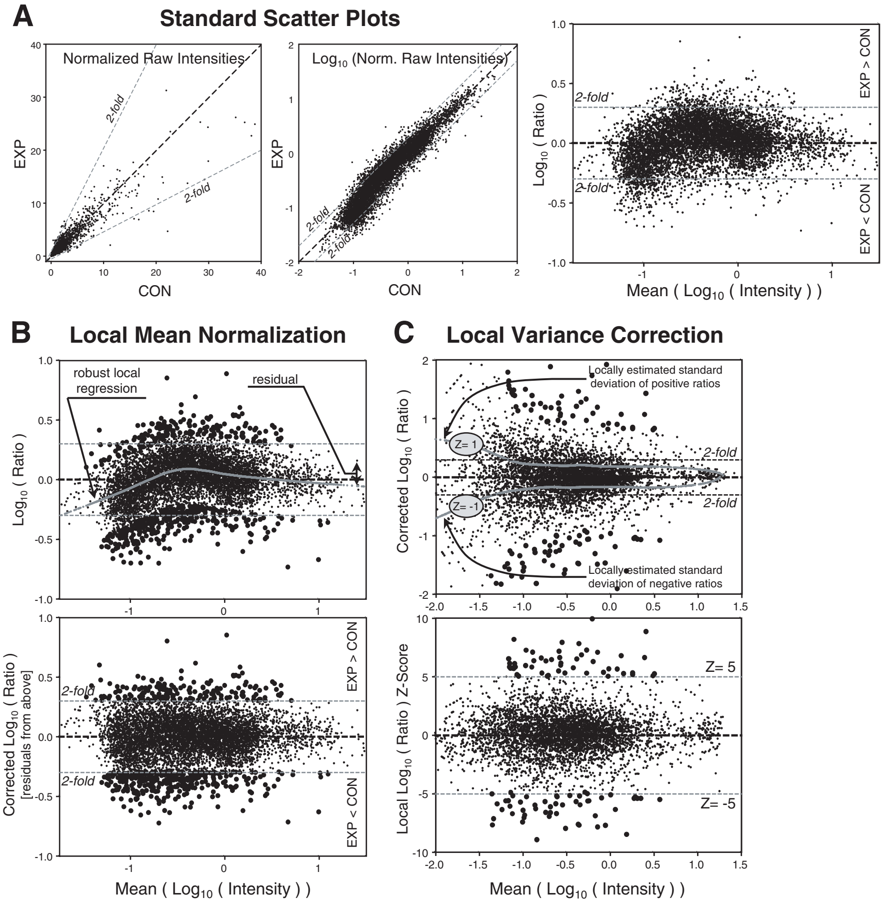

Fig. 1. (a) Standard scatter plots in gene expression analysis. (b) Local mean normalization across microarray element signal intensity (same data as depicted in Figure 1a). Top: Mean Log_{10} Intensity versus Log_{10} Ratio with overlaid local mean (gray line). Bottom: Mean Log_{10} Intensity versus the Corrected Log_{10} Ratios (residuals from Figure 1b Top). (c) Local variance correction across microarray element signal intensity. Top: Mean Log_{10} Intensity versus the Corrected Log_{10} Ratios (same plot as that is depicted in Figure 1b, but for a new dataset). A local standard deviation is calculated independently for the positive and negative ratios (gray lines) and is used to generate local Z-scores: Corrected Log_{10} Ratio/Local STDEV = Local Z-score. Bottom: Mean Log_{10} Intensity versus the Local Z-scores.

at all points across the range of element signal intensities (Figure 1b, Bottom).

A second independent parameter which varies non-uniformly across the range of absolute signal intensities is the variance in the observed gene expression ratios (Figure 1c). When variance in the observed gene expression ratios is non-uniform, it is inappropriate to apply uniform cut-off values for the selection of interesting expression changes (e.g. 2-fold cut-offs). Here, we propose the standardization of expression ratios to a standard deviation calculated locally (again using the ‘loess’ function) across the range of element signal intensities. The resulting differential gene expression metric is expressed in local standard deviation units—a local Z-score. Z-scores reflect the position of a differential gene expression value relative to the distribution of all values obtained in a particular comparison. Local Z-scores extend this by representing the position of the differential expression value relative to other values in its local neighborhood (proximal in level of gene expression).

The SNOMAD gene expression analysis tools are available via the internet: http://pevsnerlab.kennedykrieger.org/snomad.htm. No programming expertise or software download/installation is required. Users can upload their gene expression data and specify the transformations they wish to have applied to their data. Results include both a text file containing numeric values and image files depicting graphs of the data at all stages of transformation.

ACKNOWLEDGEMENTS

Funding for this work was provided by MRRC grant HD24061-12 (to J.P.) and NIH RO1 MH56639 (to S.L.Z.).

REFERENCES

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.