Robust Local Normalization Of Gene Expression Microarray Data

Paul K. Wolber, Karen W. Shannon, Stephanie B. Fulmer-Smentek, Patrick J. Collins, Kapa Lenkov, Charles D. Troup, Scott D. Connell, Srinka Ghosh, Petula N. D'Andrade, Anne B. Lucas & Douglas A. Amorese
Agilent Technologies
M/S 25U-5, 3500 Deer Creek Road,
Palo Alto, CA 94304

Synopsis: A key step in the analysis of expression microarray data is normalization: the process of adjusting the signal from 2 different reporter channels on a single microarray or a single-reporter channel on multiple microarrays to a common scale. Current methods involve either normalization to some representative statistic of all of the data or to a statistic of some subset of the data, such as a set of "housekeeping" genes. The first method fails to correct any non-linearity between the data channels; the second method is sometimes undone by differential expression of genes that were thought to be unregulated. Agilent has invented a normalization method that uses robust statistical methods to establish the "central tendency" of a set of differential expression data. Normalization utilizes the data clustered near this central tendency; these points comprise an experimentally determined set of housekeeping genes. The resulting algorithm is rapid, robust and capable of correctly normalizing microarray data from different platforms, such as cDNA and in situ synthesized oligonucleotide microarrays. In addition, the method provides an easily interpreted measurement of the degree to which the normalization has altered the original data.

Agilent Technologies

ABSTRACT

A key step in the analysis of expression microarray data is normalization: the process of adjusting the signal from 2 different reporter channels on a sin channel on multiple arrays to a common scale. The process of normalization sets the reference point for subsequent determinations of differential exp in normalization will skew the reported expression ratios, and will invalidate the assumption of random errors which underlies statistical methods that observed ratios. Current normalization methods are usually global, and involve either normalization to some representative statistic of all of the data, s a statistic of some subset of the data, such as the mean of a set of "housekeeping" genes. The first method fails if the relationship between two data average degree of differential expression is not symmetric between the samples being compared; the second method is sometimes undone by differen were thought to be unregulated. We have constructed a normalization method that uses the "central tendency" of data of comparable intensity to est normalization. The resulting algorithm is rapid, robust and superior to existing procedures in several ways. First, the method completely eliminates sys model-independent fashion, which greatly improves the reliability of statistical tests for the significance of differential expression. Second, the method of normalization genes. Thus, it preserves the advantages of methods that utilize sets of "housekeeping" genes, while verifying for each experiment th normalization subset do not exhibit differential expression. Finally, the method provides several simple, quantitative measurements of the degree to wh been altered. Such measurements can be used to gauge the quality of a given microarray experiment. We have used the method to normalize data fro situ–synthesized oligonucleotide and cDNA-deposition array experiments. In all cases, the method eliminates systematic errors, correctly identifies th and assigns quality assessments that are in agreement with other methods of determining experiment quality.

Method

This description applies to a 2-color gene expression microarray experiment. First, the net signal for the ith color (i = 1, 2) of the jth of N total features, $S_{i,j}$, is globally normalized by dividing by the geometric mean net signal in that channel:

$$\Gamma_{i,j} = S_{i,j} \Bigg/ \left(\prod_{j=1}^{N} S_{i,j}\right)^{1/N}$$

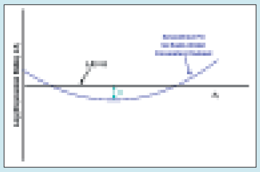
Next, the globally normalized signal $\Gamma_{i,j}$ is sorted by magnitude for each channel i, yielding a rank-order $R_{i,j}$ for each feature. For each feature j, a rank order distance δ_j is calculated and the features are filtered on the value of δ_j :

$$\delta_j = (|R_{1,j} - R_{2,j}|) \Big/ N \; ; \qquad j: \;\; \delta_j < \delta_{\max} .$$

The features that pass this filter are rank-order consistent: their relative signal rank does not change much between the two data channels. Together, these features define the central tendency of the data as a function of average signal intensity. For the purposes of a microarray experiment, they may be considered as an experimentally defined set of housekeeping genes.

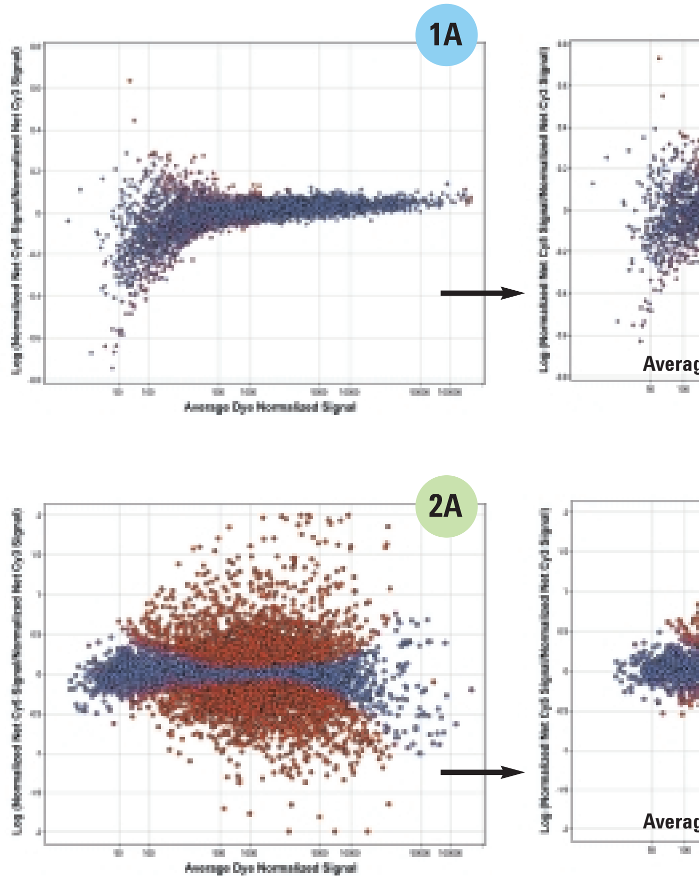
If global normalization was the correct normalization model for a particular data set, then a smooth curve fit to a graph of log expression ratios $LR_j = \log(\Gamma_{2,j} / \Gamma_{1,j})$ for points in the rank-order consistent subset (plotted as a function of average globally normalized signal $\Lambda_j = (\Gamma_{2,j}\,\Gamma_{1,j})^{1/2}$ would be a horizontal line at LR = 0. The extent to which the observed graph differs from LR = 0 defines the degree to which global normalization has systematically failed, and suggests a simple method for correcting this failure. Suppose that the smooth curve fit to the values of LR versus Λ has value $LR_j = C$ for average signal Λ_j. Then the simple transformation

$$\log(\Omega_{2,j}) \equiv \log(\Gamma_{2,j}) - C/2 \; ; \;\; \log(\Omega_{1,j}) \equiv \log(\Gamma_{1,j}) + C/2$$



will yield a new set of normalized signals $\{\Omega_{i,j}\}$ with the property that a smooth curve fit to the rank-order consistent subset of the features will now be a horizontal line at LR = 0. Note that this normalization is local: it normalizes relative to a subset of points in some neighborhood of the particular average intensity being considered. The details of the makeup of the neighborhood depend upon the method used to obtain the smooth curve fit.

By performing this procedure, one runs the danger of "correcting" a real aspect of the data. For example, if the genes probed by features on a given array are systematically up-regulated at low expression levels and down-regulated at high expression levels relative to some control sample, then this behavior will be corrected out of the data by the rank-order consistent normalization procedure. However, the root-mean-square average of the log ratio displacement C of the rank-order invariant subset (or any similar quantity, referred to as a **normalization change metric**) can be used to measure the degree to which the data has been altered, with global normalization as the reference point. Thus, this method naturally yields an experimentally-determined indication of the degree to which the data meets the assumptions underlying the method via the rms degree to which the data has been changed, versus global normalization.

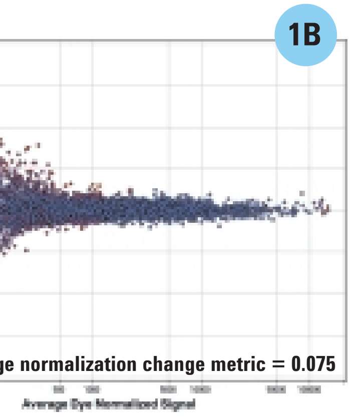


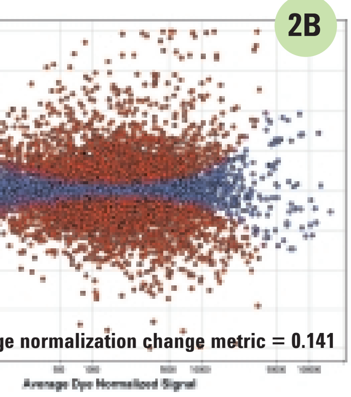
Graphical Illustration of Rank-Order Consistency Normal

Figures 1A and 1B: The figures show plots of log(expression ratio) versus for the average of 4 self-comparison gene expression microarray experim were prepared from brewer's yeast grown in synthetic complete medium; mer oligonucleotide probe to every annotated ORF in the yeast genome (s & -5022EN). Data are shown for global normalization (Figure 1A) and ran (Figure 1B) of the same arrays. In both figures, points are colored accord tained in the rank-order consistent normalization set: blue points were al points were never used and purple points were sometimes used. Note th expression) is known for this experiment, and that rank-order consistent systematic error in the global normalization.

Figures 2A and 2B: The figures show plots of log(expression ratio) versus for the average of 4 fluor-reversed pairs of gene expression microarray ex plete medium versus yeast in sporulation medium). All other details are a the rank-order consistent subset (blue) clearly delineates the central tend order consistency normalization corrects subtle systematic displacements average log(expression ratio) of zero.

ngle array or a single-reporter
pression ratios. Systematic errors
test the significance of the
such as the mean or median, or to
sets is non-linear or if the
tial expression of genes that
ablish the reference point for
stematic normalization errors in a
l experimentally defines a subset
hat the members of the
hich the underlying data has
om several types of in
e local central subset of the data

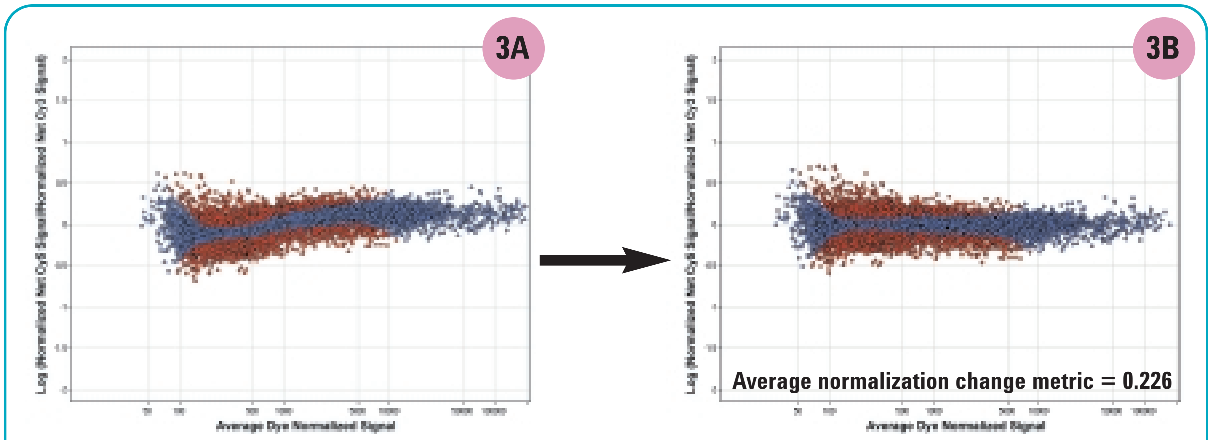




ization (Oligo Arrays)

s average net normalized intensity
ents. The hybridization samples
the array contained at least one 60-
see Agilent App. Notes 5988-5063EN
k-order consistent normalization
ing to how often they were con-
ways used for normalization, red
hat the true answer (no differential
normalization corrects an obvious

s average net normalized intensity
periments (yeast in synthetic com-
as in Figures 1A and 1B. Note that
dency of the data, and that rank-
s of this data subset away from an



Graphical Illustration of Rank-Order Consistency Normalization (cDNA Arrays)

Figures 3A and 3B: The figures show plots of log(expression ratio) versus average net normalized intensity for a single self-comparison gene expression microarray experiment. The hybridization samples were prepared from human HeLa cell RNA; the array was Agilent P/N G4100A (Human 1 cDNA array). Global normalization was used in Figure 3A; rank-order consistent normalization of the same array was used in Figure 3B. In both figures, points are colored according to how often they were contained in the rank-order consistent normalization set: blue points were used for normalization, while red points were not. Note that the true answer (no differential expression) is known for this experiment, and that rank-order consistent normalization corrects an obvious systematic error in the global normalization.

Which Genes Exhibit Rank-Order Invariance?

The rank-order consistent subset of features in a gene expression microarray experiment act like housekeeping genes: their expression levels are relatively insensitive to the physiological state of the cell. A potential use of rank-order consistent normalization is the experimental determination of the subset of genes that appear to be housekeeping genes.

In order to test this idea, we examined the subset of genes that were rank-order consistent in each of the 8 arrays in the yeast comparison experiment described in Figure 2. This experiment measured differential expression between yeast grown in sporulation medium (nitrogen starvation) versus vegetative growth in a defined complete medium. The genes in the rank-order consistent subset were grouped according to either their Gene Ontology (GO) molecular function or their GO biological process. The results of this analysis are shown in the 2 tables to the right.

The resulting lists are clearly enriched in basic cellular functions. Thus, rank-order consistency shows great promise as a method for experimentally defining sets of housekeeping genes.

Top 10 Normalization Groups (GO Molecular Function)

Gene Ontology Molecular Function	Number of Genes	% in Normalization Set
DNA-directed RNA polymerase II	7	57.14
proteasome endopeptidase	24	45.83
peptidylprolyl cis-trans isomerase	13	38.46
RNA polymerase III transcription factor	9	33.33
cochaperone	6	33.33
signal transducer	19	31.58
v-SNARE	16	31.25
protein serine/threonine kinase	18	27.78
glucose transporter	15	26.67
structural constituent of cytoskeleton	44	25.00

Top 10 Normalization Groups (GO Biological Process)

Gene Ontology Biological Process	Number of Genes	% in Normalization Set
ubiquitin-dependent protein degradation	59	33.90
chromatin silencing at HML and HMR	6	33.33
fatty acid biosynthesis	6	33.33
RNA processing	6	33.33
chromatin modeling	18	33.33
proteolysis and peptidolysis	9	33.33
transcription initiation, from Pol III promoter	9	33.33
chromatin assembly/disassembly	10	30.00
vacuolar acidification	17	29.41
establishment of cell polarity	55	29.09

Discussion and Conclusions

Rank-order consistency normalization (available as a component of Agilent's Microarray Feature Extraction software, 5/2002 release) offers the following advantages as a normalization method:

- The method is robust and model-independent
  - based on rank-order statistics
  - capable of correcting arbitrary non-linear systematic distortions
  - capable of recognizing and correcting distortions in datasets exhibiting high degrees of differential expression (see Figs. 2A & 2B).
- The method naturally provides an experimentally-based measurement of the degree to which the data has been altered
- The method is capable of experimentally defining sets of housekeeping genes

The authors wish to acknowledge the assistance of the array fabrication, QC and scanner groups at Agilent Technologies, as well as many helpful discussions with Dr. Nicholas Sampas (Agilent).

Agilent's Printed Microarray Solutions

Design it!	**Microarray Design Services**
G2560A	Microarray Design and Basic QC
G2561A	Probe Selection
G2562A	Probe Curation
G2563A	Professional Consulting Service
Print it!	**cDNA and Custom Microarrays**
G2506A	25-mer Custom *in situ* Oligonucleotide Microarray (8.4K)
G2507A	25-mer Custom *in situ* Oligonucleotide Microarray (22K)
G2508A	60-mer Custom *in situ* Oligonucleotide Microarray (8.4K)
G2509A	60-mer Custom *in situ* Oligonucleotide Microarray (22K)
G4100A	Human 1 cDNA Microarray Kit
G4101A	Human 2 cDNA Microarray Kit
G4104A	Mouse cDNA Microarray Kit
G4105A	Rat cDNA Microarray Kit
G4135A	Arabidopsis 1 Microarray Kit
Run it!	**Microarray Processing Tools**
G2554A	Fluorescent Linear Amplification Kit
G2556A	Fluorescent Linear Amplification Kit with Hyb'n Reagent
G2559A	*in situ* Hybridization Reagent Kit
G2557A	Fluorescent Direct Label Kit
G2555A	Fluorescent Direct Label Kit with Hybridization Reagent
G2558A	Deposition Hybridization Reagent Kit
G4145A	Large Volume Deposition Hybridization Kit
G2530A	Microarray Hybridization Chamber (8.4K configuration)
G2530-60002	Hybridization (8.4K format) Septa, Backings & Gasket
G2533A	Microarray Hybridization Chamber (16.2K configuration)
G2533-60002	Hybridization (16.2K format) Septa, Backings & Gasket
G2531A	Microarray Hybridization Chamber (22K configuration)
G2531-60002	Hybridization (22K format) Septa, Backings & Gasket
G2940BA	2100 Bioanalyzer Instrument System Bundle
5065-4476	RNA 6000 Nano LabChip Kit (messenger & total RNA)
5064-8230	DNA 7500 LabChip Kit (100 - 7500 bp)
5064-8231	DNA 12000 LabChip Kit (100 - 12000 bp)
5064-8284	DNA 500 LabChip Kit (25 - 500 bp)
5065-4449	DNA 1000 LabChip Kit (25 - 1000 bp)
G2565AA	48-slide, Dual Laser DNA Microarray Scanner
(# varies)	Microarray Technology Transfer, Services and Support Packages
Discover it!	**Microarray Analysis**
G2567AA	Feature Extraction Software License
(# varies)	Rosetta Resolver Software

Ordering Information

www.agilent.com/chem/dna
u.s. and canada 1 800 227 9770
japan +0120 477 111
europe: marcom_center@agilent.com
global: dna_microarray@agilent.com

Research Use Only

LifeSeq is a trademark or registered trademark of Incyte Genomics, Inc. in the U.S. and other countries. Rosetta Resolver is a U.S. registered trademark of Rosetta Inpharmatics. Information, descriptions and specifications in this publication are subject to change without notice. Printed in the U.S.A.

Printed in the U.S.A.
June 1, 2002
5988-6932EN

Agilent Technologies