Journal of Biometrics & Biostatistics

Gelmi et al., J Biomet Biostat 2011, 2:3
DOI: 10.4172/2155-6180.1000114

Research Article Open Access

Experimental Validation of a Probabilistic Framework for Microarray Data Analysis

Claudio A. Gelmi[1], Purusharth Prakash[2], Jeremy S. Edwards[3,4] and Babatunde A. Ogunnaike[2]*

[1]Department of Chemical and Bioprocess Engineering, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Santiago, Chile
[2]Department of Chemical Engineering, University of Delaware, Newark, DE 19716, USA
[3]Molecular Genetics and Microbiology, Cancer Research and Treatment Center, University of New Mexico Health Sciences Center, Albuquerque, NM, USA
[4]Chemical and Nuclear Engineering, University of New Mexico, Albuquerque, NM, USA

Abstract

With the primary objective of developing fundamental probability models that can be used for drawing rigorous statistical inference from microarray data, we have presented in a previous publication, theoretical results for characterizing the entire microarray data set as an ensemble. Specifically, we established, from first principles, that under reasonable assumptions, the distribution of microarray intensities follows the gamma model, and consequently that the underlying theoretical distribution for the entire set of fractional intensities is a mixture of beta densities. This probabilistic framework was then used to develop a rigorous statistical inference methodology whose outcome, for each gene, is an ordered triplet: a raw computed fractional (or relative) change in expression level; an associated probability that this number indicates lower, higher, or no differential expression; and a measure of confidence associated with the stated result.

In this paper we validate the probabilistic framework and associated statistical inference methodology through confirmatory experimental studies of gene expression in *Saccharomyces cerevisiae* using Affymetrix Genechips®. The array data were analyzed using the probabilistic framework, and 9 genes—with indeterminate expression status according to the standard 2-fold change criteria, but for which our probabilistic method indicated high expression status probabilities—were selected for higher precision characterization. In particular, for genes CGR1, GOS1, ICS2, PCL5 and PLB1, the high probabilities of being differentially expressed (up or down) were found to be in excellent agreement with the expression status determined by the independent, high precision confirmatory experiments. These confirmatory experiments, using the high precision, medium throughput polonies technique, confirmed that the probabilistic framework performs quite well in correctly identifying the expression status of genes in general, but especially differentially expressed genes that would otherwise not have been identifiable using the standard 2-fold change criteria.

Keywords: Mixture models; Beta distributions; Gene expression; Polonies

Introduction

One of the primary goals of functional genomics is to provide a quantitative (as opposed to qualitative) understanding of the functions of genes, how they influence and are influenced by proteins and the environment, and how they regulate the function of complex living organisms from the cellular level all the way to the physiological level. Because the structure, function, and behaviour of a cell are determined by gene expression patterns, it is therefore no surprise that considerable research effort has been devoted to the development of techniques for measuring the expression level of all the genes in the cell. And with the advent of the extremely high throughput microarray technology, researchers finally have the means to collect expression data on every gene in a cell simultaneously. However, these vast data sets provide too much unstructured information to be analyzed without computational tools. In response to these challenges, the microarray literature continues to grow exponentially, with the publication of many novel findings from gene expression studies as well as new and more sophisticated statistical methods for analyzing gene expression data. We believe that because the high-throughput microarray technology does not produce high-precision data, it is best used for "screening" thousands of genes to identify the subset that may be of potential interest. A complementary technology that is more precise (and more accurate) but not necessarily "high-throughput" (e.g., polony technology [1]) may subsequently be used to characterize this smaller subset more efficiently.

In a previous publication [2] we proposed fundamental probability models for microarray data distributions. There, we presented first principles theoretical results that confirmed what had previously been speculated (*e.g.*, [3], or assumed for convenience [4]): that under very reasonable assumptions, the distribution of microarray intensities should follow the gamma (not lognormal) model. Furthermore, the biological interpretations of the model parameters emerge naturally from our derivation. We subsequently established that a polar coordinate transformation of raw intensity data provides the basis for a technique in which each microarray data set is represented as a mixture of beta densities, from which rigorous statistical inference may be drawn regarding differential gene expression. Specifically (see [2]), just as classical statistical inference is based on theoretical reference distributions (such as the Gaussian, *t*-, Chi-square, or the *F*-distributions) we developed a methodology for drawing statistical

***Corresponding author:** Babatunde A. Ogunnaike, Department of Chemical Engineering, University of Delaware, Newark, DE 19716, USA, Tel: +1-302-831-4504; Fax: +1-302-831-1048; E-mail: ogunnaike@udel.edu

Received April 26, 2011; **Accepted** August 01, 2011; **Published** September 25, 2011

Citation: Gelmi CA, Prakash P, Edwards JS, Ogunnaike BA (2011) Experimental Validation of a Probabilistic Framework for Microarray Data Analysis. J Biomet Biostat 2:114. doi:10.4172/2155-6180.1000114

inference using a mixture distribution of beta densities as its theoretical basis: it consists of (i) a probability statement that a gene belongs to a category (the "down-regulated", the "up-regulated", or the "no-change") and (ii) a degree of confidence associated with such probability statements, determined from the variability estimated from replicates, or else by propagation-of-error techniques when there are no replicates. The final outcome is an ordered triplet of results for each gene: a raw computed fractional (or relative) change in expression level, an associated probability that this number indicates lower, higher, or no differential expression (a category membership probability) and a measure of confidence associated with the stated result. Genes that clear user-specified threshold values for the probabilities, and for the confidence index, may then be selected by the researcher for further study. The principles and application of the technique was illustrated in [2] with a real experimental data set and also via simulation.

The objective in this work is to test and validate our probabilistic framework experimentally by analyzing gene expression in *Saccharomyces cerevisiae* using Affymetrix Genechips®. First, the raw microarray data is analyzed, according to the probabilistic framework, to produce the probability that each gene belongs to a category showing lower, higher, or no differential expression; and a measure of confidence associated with the stated probabilities. Next, independent follow-up studies are conducted on a subset of genes using the higher-precision, medium throughput polonies technique [1] to confirm or refute the expression status determined by the probabilistic analysis of the original microarray data. Specifically, these genes—nine in total—are selected from a group whose expression status would be indeterminable using the standard fold-change criteria.

Polony technology is a form of polymerase chain reaction (PCR) in which the reaction is immobilized in a thin polyacrylamide gel attached to a microscope slide. As the chain reaction proceeds, the PCR products diffuse radially within the gel from its immobilized template giving rise to a polymerase colony. When the gel is stained with SyberGreen I and scanned with a microarray scanner, the polymerase colony resembles a colony on an agar plate, hence its name. One major advantage of this technology is that each immobilized template gives rise to only a single polony. As a result, highly precise digital data can be obtained with this approach [1,5], especially in the context of mRNA expression profiling when first-strand cDNA is used as the PCR template [6-8]. The accuracy of this technique enables its use as a reliable confirmatory tool for gene expression studies.

Materials and Methods

Strain and media

Saccharomyces cerevisiae strain FY4 was used in this study. Fructose (control condition) and galactose were used as carbon sources and the minimal media used were prepared as described elsewhere [9]. All cultures were grown at 30ºC in a BioFlo 110 Benchtop Fermentor (New Brunswick Scientific). The pH was maintained at 5.5 and dissolved oxygen was maintained above 60%. The inoculum for the culture was grown in shake flasks until late log phase.

Parameters measured

Cell density was monitored by taking OD (600 nm) measurements at regular intervals. Samples were taken to measure substrate (carbon source) concentration during the course of the experiment. Cell dry weight and oxygen uptake rate were measured in early- to mid-log phase (OD_{600} 0.5-1.0). RNA samples were taken at mid-log phase (OD_{600} 0.8-1.0).

Microarray reactions and pre-treatment

Total yeast RNA was prepared using MasterPure Yeast RNA Purification Kit (Epicentre) and samples were prepared according to the protocol described by HPCGG (http://www.hpcgg.org). RNA obtained from fructose minimal culture was used as control condition. The microarray analysis was done using Affymetrix Genechips® (Yeast Genome 2.0 Array) and was performed by HPCGG. In order to facilitate any posterior comparison of candidate genes, median signal intensities for all arrays were normalized using global LOWESS (smooth parameter = 0.33).

Polony slides

Polony reactions were conducted as previously described [6]. The following master mix recipe was used to cast 4 polony gels. In a microcentrifuge tube, a master mix was made containing 34.32 µL of molecular biology grade water, 7.55 µL of 10X JumpStart buffer (Sigma Aldrich), 3.03 µL dNTP (Ambion), 0.51 µL of 30% BSA (Sigma Aldrich), 0.76 µL of 10% Tween 20 (Fisher Scientific), and 18.85 µL of degassed, filter-sterilized acrylamide (Fisher Scientific), which were combined and vortexed briefly to mix. 63.47 µL of master mix was combined with 1.9 µL of appropriately diluted template cDNA, 1.5 µL each of a forward and a reverse primer (100 µM), 11 µL of JumpStart *Taq* (Sigma Aldrich), 1.5 µL of 5% (v/v) TEMED (Pharmacia), 1.5 µL of 5 % (w/v) ammonium persulfate (Pharmacia). 19 µL of this mixture was transferred to the well of a teflon coated single well bind-silane (Pharmacia) treated glass slide (Erie Scientific). The well was covered with a glass slide cover and the gel was allowed to polymerize for 10 min. The slide cover was then covered with a hybridization well (Grace Bio-Labs) and mineral oil was injected into the open cavity surrounding the glass slide cover and sealed. The slides were thermal cycled (24˚C 5 min., 94˚C 2 min., 54 cycles of 94˚C 15 sec., 61˚C 30 sec., 72˚C 30 sec. and a final 72˚C extension for 2 min) using a PTC200 thermal cycler adapted for use with glass slides (16/16 twin tower block, MJ Research). Note that the reverse primer has a 5' acrydite modification. The acrydite modification immobilizes the resulting polonies in the acrylamide gel. The hybridization wells were then removed, the slides were washed in hexane to remove the mineral oil and the glass slide covers were removed.

Polony slides were SYBR green (Molecular Probes) stained, imaged and the single base extension protocol was initiated. SYBR green staining was conducted by immersing polony slides in 2X SYBR green I (TBE, pH 8.0) for 10 min. followed by a 5 min. wash in TBE. Slides were imaged on a ScanArray Express (GSI Lumonics).

Single Base Extension (SBE) sequencing

Single base extension sequencing was used to quantify polonies arising from different genes in the mixed sample. Polony slides were initially incubated in a 70 % (v/v) formamide in 1X SSC solution for 15 min. at 70˚C to denature the double stranded polony DNA. Following the denaturing step, electrophoresis (42 % (w/v) urea in 0.5X TBE) was conducted to remove the DNA strand not covalently bound to the gel with the 5' acrydite. The electrophoresis was run for 2.5 h at 140 V in a standard DNA gel electrophoresis box. Following the electrophoresis, slides were washed 4x5 min. in Wash 1E (100 mM tris pH 7.5, 20 mM EDTA, 500 mM KCl). 150 µL of 1.33 µM sequencing primer in 6X SSPE and 0.1% (v/v) Triton X-100 (Acros Organics) was placed on the surface of the gel, covered with a hybridization well and allowed to anneal to the immobilized polony DNA (94˚C for 3 min. then 55˚C for 15 min). Slides were washed 2x4 min. in Wash 1E and equilibrated for 1 min. in Klenow Extension Buffer (50 mM tris pH 7.5, 5 mM $MgCl_2$, 0.01% (v/v) Triton X-100). Slides were then covered with 60 μL of an extension solution consisting of 60 μL Klenow Extension Buffer, 0.5 μL Klenow fragment (5000 U/mL, New England Biolabs), 0.5 μL Single Stranded DNA Binding Protein (1-5 μg/ μL, USB), 1.25 μL of either Cyanine-5-dATP, -dCTP or –dUTP (10 μM, PerkinElmer Life Sciences) and 1.25 μL of either Cyanine-3-dATP, -dCTP or –dUTP (10 μM, PerkinElmer Life Sciences) depending upon the desired extension. The extension reaction was allowed to proceed for 3 minutes at room temperature. Slides were then washed 2x4 min. in Wash 1E and imaged using a ScanArray Express scanner.

Polonies gene expression analysis

Gene expression analysis was conducted as described [6,8]. Briefly, 5 mL of each culture was harvested in mid-exponential phase (OD_{660} approximately 0.5-2.5) and was immediately chilled to approximately 0˚C. Media was removed by decanting following centrifugation at 9500 rpm and 4˚C for 10 min. Cells were stored at −80˚C prior to recovering RNA. Total RNA was isolated using the MasterPure Yeast RNA Purification Kit (Epicentre) according to the manufacturer's protocol. 5 μg of RNA from each culture was used as template for cDNA synthesis using the SuperScript First-Strand Synthesis System for RT-PCR (Invitrogen) according to the manufacturer's protocol. This first-strand cDNA was the template for polony slides. Polonies were prepared as described above.

Statistical analysis

The Affymetrix GeneChip® array data was analyzed as described in Ogunnaike et al. [2]. Briefly, for each gene we computed: (i) x_i, a raw fractional (or relative) change in expression level; (ii) ***P***, an associated probability vector that this number indicates lower (P_{down}), higher (P_{up}), or no differential expression (P_{non}) (note that, by definition, $P_{down} + P_{up} + P_{non}=1$); and (iii) c_i ($0 \le c_i \le 1$), a measure of confidence associated with the computed probabilities. A low value of c_i indicates a correspondingly low degree of confidence in the assertion about the true state of expression status of the gene *i* implied by the computed probabilities. Conversely, a high value of c_i corresponds to a higher degree of confidence in what the computed probabilities imply about the expression status of the gene *i*.

Results and Discussion

With the premise that confirmatory experiments constitute a natural next step in validating the performance of any statistical analysis method, we present in this section a discussion of how the entire probabilistic framework has been tested in practice. The two steps involved in the validation procedure are: i) analysis of Affymetrix GeneChip® array data and selection of some candidate genes (for higher precision characterization) based on high probabilities of expression status and high associated confidence indexes; and ii) independent characterization of the real expression status (up-regulated, down-regulated or not differentially expressed) of the selected genes, using a high-precision but not necessarily high-throughput technology. These two steps are not only relevant to this confirmatory study; we believe that they should be part of any gene expression study using microarrays. Because high-throughput microarray technology rarely produces high-precision data, it is best used for "screening" thousands of genes to identify a smaller subset that may be of potential interest. It is then more efficient to use a higher-precision technology to characterize this smaller subset more precisely as a follow up to such preliminary screening.

Of particular interest to us are genes that show subtle *but* biologically significant changes in gene expression because such changes are especially difficult, if not outright impossible, to identify as statistically significant using current techniques. As presented in Ogunnaike et al. [2], one of the distinguishing features of our probabilistic approach is that it provides a means for identifying significant changes in such genes. The ability to identify such changes in gene expression should therefore provide a most stringent test of the capabilities of this approach. Thus, this validation study focuses specifically on the category of difficult-to-identify genes for which the expression data show fold-changes (the ratio of measured "test" to "control" expression levels) that are lower than 2.0 for potentially up-regulated genes, or higher than 0.5 for potentially down-regulated ones, *i.e.*, genes that are entirely undetectable as differentially expressed strictly by the 2-fold change criterion.

A summary of the experiments and the main highlights of the results now follow. *Saccharomyces cerevisiae* was grown under two different conditions and its gene expression was studied using Affymetrix GeneChips®. Four pairs of microarrays (biological replicates) were analyzed using the probabilistic framework. A sample of the results is shown in (Table 1). (The data sets that passed the initial quality control, along with the complete set of analysis results are available at http://www.che.udel.edu/systems/supplements/EVPartII.zip).

Nine candidate genes were selected from the above-mentioned "nebulous" fold-change region, using the following criteria: P_{down}, P_{non} or $P_{up} \ge 0.80$, and $c_i \ge 0.70$. In other words, we select for higher precision characterization, 9 genes whose expression status are indeterminate using the 2-fold change criterion, but for which (i) the probabilistic analysis indicates reasonably high probabilities (0.8 or higher) of being up- or down-regulated, or not differentially expressed, and (ii) there is also a reasonably high degree of confidence (0.7 or higher) in these estimated probabilities. The selected genes are highlighted in boldface type in Table 1. Finally, the expression characteristics of the selected genes were subsequently determined in a different set of independent experiments, using the higher-precision, medium throughput polonies technique [1]. The results of the polonies experiments were then compared with what was predicted by the probabilistic analysis of the original microarray data. Note: because of the chosen probability threshold of 0.8, the expectation is that there should be only one or two cases (20% of 9 genes)—definitely no more than 3 cases, if some allowance is made for the small sample size of 9—in which the probabilistic approach's predictions may not fully agree with the "true" expression status of the genes as determined by the higher-precision technique.

Table 2 shows the names of the 9 selected genes, the fold-changes computed from the Affymetrix data, the corresponding probabilities that the gene in question is up- or down-regulated, or not differentially expressed, and the confidence index associated with the probabilities. The last two columns show the fold-change computed from the confirmatory polonies experiments, along with the indicated expression status.

Some important points to note from this table are: (i) for genes *CGR1, GOS1, ICS2, PCL5* and *PLB1*, the indicated high probabilities of being differentially expressed (up or down) are in excellent agreement with the expression status determined by the independent, high precision confirmatory experiments. It is important to stress once again that none of these genes would have been identified as differentially expressed by the 2-fold change criterion. (ii) Even though most micro-

Gene name	Fractional intensity (x)			P_{down}			P_{non}			P_{up}			Confidence index (c_i)
	Min	Aver.	Max	Min	Aver.	Max	Min	Aver.	Max	Min	Aver.	Max	
NUP120	0.48	0.49	0.50	0.00	0.02	0.05	0.91	0.96	0.98	0.00	0.02	0.04	0.94
PCL5	**0.37**	**0.39**	**0.40**	**0.76**	**0.86**	**0.91**	**0.08**	**0.14**	**0.24**	**0.00**	**0.00**	**0.00**	**0.88**
HDA2	0.46	0.49	0.51	0.01	0.03	0.06	0.92	0.97	0.99	0.00	0.01	0.02	0.95
PAU7	0.48	0.52	0.57	0.00	0.01	0.02	0.95	0.97	0.99	0.00	0.02	0.05	0.97
MIP1	0.52	0.53	0.54	0.00	0.01	0.02	0.87	0.92	0.97	0.01	0.07	0.13	0.91
SIN4	0.19	0.24	0.28	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
GOS1	**0.60**	**0.61**	**0.62**	**0.00**	**0.00**	**0.00**	**0.05**	**0.19**	**0.40**	**0.60**	**0.81**	**0.95**	**0.71**
MET31	0.45	0.47	0.49	0.02	0.07	0.20	0.80	0.93	0.98	0.00	0.00	0.00	0.85
TOR1	0.53	0.53	0.54	0.00	0.01	0.02	0.81	0.90	0.97	0.00	0.09	0.19	0.86
BRE2	0.45	0.50	0.54	0.01	0.03	0.06	0.89	0.95	0.98	0.00	0.02	0.08	0.92
GAL11	**0.49**	**0.52**	**0.56**	**0.00**	**0.02**	**0.05**	**0.77**	**0.88**	**0.95**	**0.00**	**0.10**	**0.23**	**0.83**
FYV7	0.33	0.38	0.42	0.21	0.66	0.92	0.08	0.34	0.79	0.00	0.00	0.00	0.42
ELP6	**0.48**	**0.50**	**0.52**	**0.00**	**0.02**	**0.03**	**0.91**	**0.95**	**0.98**	**0.00**	**0.03**	**0.05**	**0.95**
BNA1	0.51	0.55	0.58	0.00	0.01	0.03	0.81	0.93	0.99	0.00	0.06	0.16	0.86
MCM3	0.46	0.48	0.50	0.00	0.04	0.14	0.86	0.93	0.99	0.00	0.03	0.07	0.89
KIP3	**0.47**	**0.49**	**0.50**	**0.00**	**0.02**	**0.04**	**0.96**	**0.97**	**0.98**	**0.00**	**0.01**	**0.02**	**0.97**
HXT4	0.08	0.09	0.11	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
ENP1	0.45	0.46	0.47	0.01	0.09	0.18	0.81	0.90	0.99	0.00	0.01	0.02	0.86
NAF1	0.38	0.40	0.42	0.23	0.63	0.92	0.08	0.36	0.77	0.00	0.00	0.00	0.43
CGR1	**0.38**	**0.38**	**0.39**	**0.76**	**0.85**	**0.95**	**0.05**	**0.15**	**0.24**	**0.00**	**0.00**	**0.00**	**0.85**
IRA1	0.30	0.33	0.37	0.89	0.95	0.99	0.01	0.05	0.11	0.00	0.00	0.00	0.92
PLB1	**0.62**	**0.65**	**0.66**	**0.00**	**0.00**	**0.00**	**0.00**	**0.03**	**0.09**	**0.91**	**0.97**	**1.00**	**0.93**
THI4	0.47	0.50	0.53	0.00	0.04	0.12	0.87	0.96	0.99	0.00	0.01	0.01	0.90
DID2	0.49	0.50	0.53	0.00	0.02	0.06	0.84	0.93	0.98	0.00	0.05	0.15	0.87
ICS2	**0.38**	**0.39**	**0.40**	**0.79**	**0.84**	**0.94**	**0.06**	**0.16**	**0.21**	**0.00**	**0.00**	**0.00**	**0.87**
MID1	0.50	0.51	0.52	0.00	0.01	0.03	0.86	0.94	0.98	0.00	0.05	0.13	0.90
BRE2	0.45	0.50	0.54	0.01	0.03	0.06	0.89	0.95	0.98	0.00	0.02	0.08	0.92
YIP5	**0.48**	**0.50**	**0.52**	**0.00**	**0.01**	**0.03**	**0.91**	**0.96**	**0.98**	**0.00**	**0.03**	**0.06**	**0.95**
PBS2	0.49	0.51	0.53	0.00	0.01	0.03	0.81	0.93	0.98	0.00	0.06	0.18	0.86

Table 1: Sample of genes from the Affymetrix data analysis. Highlighted genes in bold were chosen for further confirmatory studies.

Candidate genes	Fold change*	x	P_{down}	P_{non}	P_{up}	Confidence index (c_i)	Polony fold- change ± 95% CI	Expression status
CGR1	0.62	0.38	**0.85**	0.15	0.00	0.85	0.56 ± 0.03	D
ELP6	1.01	0.50	0.02	**0.95**	0.03	0.95	0.98 ± 0.08	N
GAL11	1.11	0.53	0.02	**0.88**	0.10	0.83	1.10 ± 0.06	N/U
GOS1	1.55	0.61	0.00	0.19	**0.81**	0.71	1.21 ± 0.04	U
ICS2	0.63	0.39	**0.84**	0.16	0.00	0.87	0.79 ± 0.05	D
KIP3	1.01	0.50	0.02	**0.97**	0.01	0.97	1.10 ± 0.06	N/U
PCL5	0.63	0.39	**0.86**	0.14	0.00	0.88	0.62 ± 0.08	D
PLB1	1.84	0.65	0.00	0.03	**0.97**	0.93	1.71 ± 0.13	U
YIP5	1.00	0.50	0.01	**0.96**	0.03	0.95	1.07 ± 0.04	N/U

D = down-regulated, N = not differentially expressed and U = up-regulated.
* Fold-change computed from the Affymetrix microarray data.

Table 2: Summary of the most important confirmatory experiment results: i) fold-change computed from the Affymetrix data and the corresponding fractional intensity (x), ii) the probabilities of expression status and corresponding confidence indexes, iii) the fold-change computed from polony technology data ±95% confidence intervals (CI), iv) expression status of the genes according to the confirmatory experiments.

array studies are geared towards identifying "interesting" (*i.e.*, differentially expressed) genes, it is also important to be able to avoid being fooled by non-differentially expressed genes. It is in this regard that we selected the genes *ELP6, GAL11, KIP3* and *YIP5* for inclusion in this validation study. The non-differentially expressed status of *ELP6* predicted by the probabilistic approach is unconditionally confirmed by the polonies experiment. With *GAL11, KIP3* and *YIP5,* the confirmatory experiments show a tantalizingly small possibility that they may be ever so slightly up-regulated, as opposed to not differentially expressed. However, in each case, the proximity of the lower bound of the indicated fold-change confidence intervals to 1.00 (*i.e.*, 1.03-1.04), and the fact that the polonies technique, while of higher precision is nonetheless not entirely error-free, combine to make it difficult to argue against the initial prediction that these genes are not differentially expressed.

Overall therefore, these results appear to confirm, with a reasonable degree of certainty, that the probabilistic framework performs quite well in correctly identifying the expression status of genes in general; and that it is especially effective in identifying differentially expressed genes, even in the nebulous region of subtle changes in gene expression where the fold-change criterion would have been too coarse to be of any use.

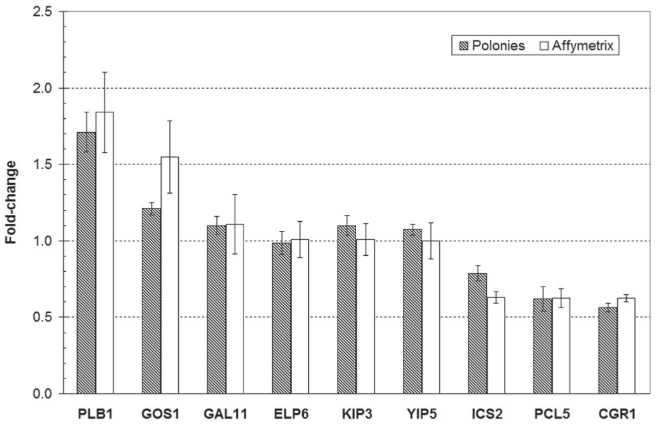

Figure 1: Comparison of fold-change values for 9 genes computed from data obtained using polony technology and Affymetrix GeneChip® (with 95% confidence intervals).

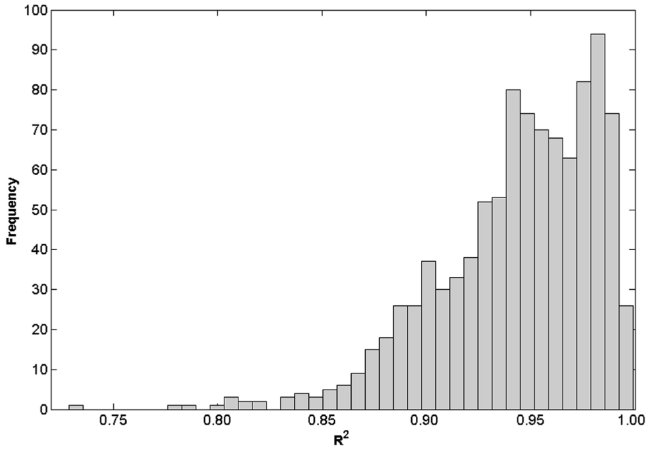

Figure 2: Histogram of 1,000 bootstrap samples of the coefficient of determination (R^2) between fold-change values computed from polonies data and those computed from Affymetrix data.

As an added benefit of the validation study, we were also able to explore how the fold-change values determined from the polony technology data compares with the one computed from Affymetrix data. Figure 1 shows, for each microarray technology, a plot of the fold-change values (and corresponding confidence intervals) computed for the 9 genes involved in the validation study. Observe that most of the confidence intervals overlap, indicating good agreement between the different technologies in quantifying the changes in the expression of the genes in question. A second comparison was carried out using bootstrap samples of the average fold-change values obtained with each technology. The bootstrap method is a procedure that involves choosing random samples with replacement from a data set and analyzing each sample using the same procedure [10]. We resampled each fold-change vector 1,000 times and computed the coefficient of determination (R^2) between the resulting fold-change "data" vectors using the *bootstrp* function of MATLAB®. The resulting histogram of the bootstrap analysis in Figure 2 shows that nearly all the estimates of R^2 lie in the interval [0.9, 1.0], with an average R^2 value of 0.94. This is strong quantitative evidence that the fold-change values computed using the two different gene expression quantification technologies are very strongly positively correlated. And given that the polony technology is quite accurate *and* precise [1,5], these high R^2 values along with the results in Figure 1 indicate that the high throughput Affymetrix GeneChip® technology produces results that are comparable –at least in accuracy (if not in precision) –to those produced by the lower throughput but more precise polony technology. It is therefore not surprising that the Affymetrix GeneChip® technology has now become a popular choice for massive gene expression studies.

Acknowledgements

CAG and BAO acknowledge funding from the Delaware Biotechnology Institute; JSE acknowledges a subcontract grant from Genomatica Inc.

References

1. Mitra R, Church G (1999) In situ localized amplification and contact replication of many individual DNA molecules. Nucleic Acids Res 27: 34-39.
2. Ogunnaike BA, Gelmi CA, Edwards JS (2010) A probabilistic framework for microarray data analysis: Fundamental probability models and statistical inference. Journal of Theoretical Biology 264: 211-222.
3. Sebastiani P, Gussoni E, Kohane IS, Ramoni MF (2003) Statistical challenges in functional genomics. Statistical Science 18: 33-60.
4. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol 8: 37-52.
5. Vogelstein B, Kinzler KW (1999) Digital PCR. Proc Natl Acad Sci 96: 9236-9241.
6. Merritt J, DiTonno JR, Mitra RD, Church GM, Edwards JS (2003) Parallel competition analysis of *Saccharomyces cerevisiae* strains differing by a single base using polymerase colonies. Nucleic Acids Research 31: 84.
7. Butz JA, Yan H, Mikkilineni V, Edwards JS (2004) Detection of allelic variations of human gene expression by polymerase colonies. BMC Genetics 5.
8. Mikkilineni V, Mitra RD, Merritt J, DiTonno JR, Church GM, et al. (2004) Digital quantitative measurements of gene expression. Biotechnology and Bioengineering 86: 117-124.
9. Rose MD, Winston F, Hieter P (1990) Methods in Yeast Genetics: A Laboratory Course Manual. Plainview NY: Cold Spring Harbor Lab, Press.
10. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. New York: Chapman & Hall 18:168-174.