AFFYMETRIX®

Technical Note

Sample Pooling for Microarray Analysis: A Statistical Assessment of Risks and Biases

Pooling of RNA samples isolated from tissue is a strategy that can be implemented in microarray experiments when the amount of sample RNA is limiting, or when variation across samples must be reduced. However, it is widely documented that important gene expression information is lost during pooling. This Technical Note qualifies and quantifies the effects of pooling on biological conclusions derived from gene expression data.

- Many transcripts identified as significantly changed in individual samples were not identified in the pooled samples. All information of small, but statistically significant, changes in expression was lost in the pooled samples.
- It is recommended that researchers use non-pooled (individual) samples in order to identify statistically significant changes in gene expression.

Introduction

Microarray samples are often pooled to reduce experimental costs, compensate for insufficient sample RNA, or to reduce sample variation. However, pooling results in an irreversible loss of information. This loss is severe in human clinical samples where there is variability in the genotype of individuals as well as other confounding variables, such as age, sex, disease progression, and treatment. Sample classification is based on subjective measures, such as clinical diagnosis or histological observation. For example, tumors that appear to be similar may have different origins or disease mechanisms. Once RNA samples are mixed, it is impossible to identify outliers or misclassified samples. In light of this issue, it is difficult to justify pooling clinical samples aside from technical reasons, such as limited availability of RNA.

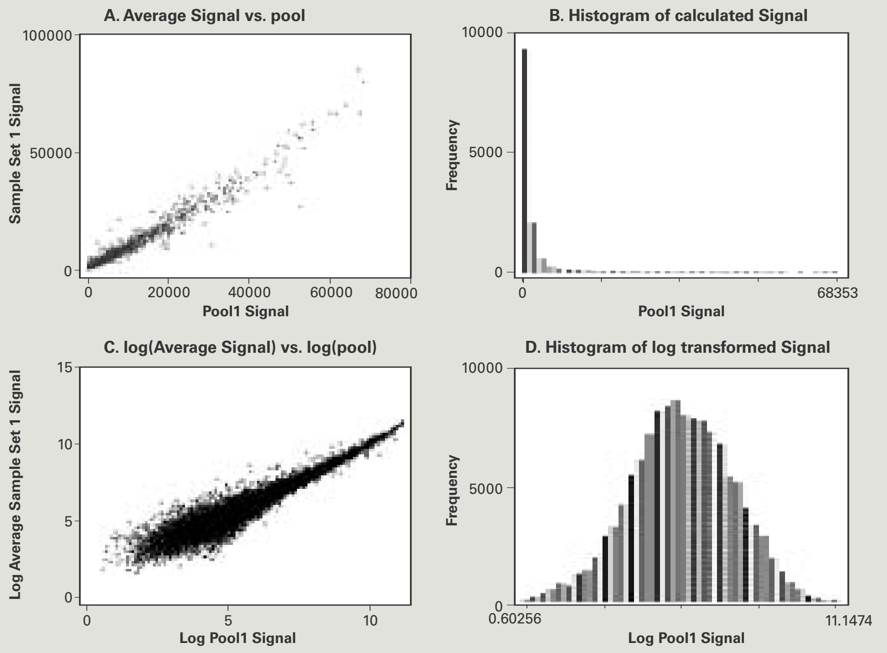
Experimental Design

This study was conducted to evaluate the expression changes between normal and diseased liver samples (Drs. S. Tsutsumi and H. Aburatani, University of Tokyo, ongoing study). Samples were obtained from healthy individuals and from patients with liver cancer. RNA was isolated and prepared from these samples and either hybridized individually to GeneChip® Human Genome U95Av2 (HG-U95Av2) arrays or pooled first and then hybridized to the arrays (Table 1). Data generated from the pooled samples were compared to data from the individual replicates within the same sample set. The term "signal" refers to the transcript abundance as calculated by MAS 5.0. The statistical calculations throughout this paper were computed using STATA7 Special Edition (Stata Corporation, College Station, TX).

Table 1. Sample sets and pooling strategy. Samples fall into two groups: liver cancer, and normal liver samples. Eight individuals were represented in each group, and RNA from each individual was hybridized to a single array. Pools 1 and 2 consisted of RNA from the entire group hybridized to a single array. Pooled samples were created by adding an equivalent amount of RNA from each individual sample. All samples were labeled and hybridized to arrays according to the GeneChip® Expression Analysis Technical Manual. Five g of total RNA were used as the starting material for all array experiments.

Sample Set 1: 8 liver cancer samples (8 individuals)		**Sample Set 2: 8 normal liver samples (8 individuals)**	
Individuals		Individuals	
1	Pool 1	9	Pool 2
2		10	
3		11	
4		12	
5		13	
6		14	
7		15	
8		16	

Figure 1. Log transformation of microarray data. (**A**) The y axis represents the mean signal of individual samples and the x axis represents the pooled signal of Sample Set 1. (**B**) The distribution of the signal data in panel A. Most data points are concentrated at the low end of signal values, while a few outliers were distributed in the high end of the signal values. This distribution of data allows only non-parametric statistical tests which are less powerful than parametric tests. (**C** and **D**) The same data as before but log transformed. The data now approximate a normal distribution which allows more powerful statistical tests such as *t*-test and ANOVA.



Results and Interpretation

LOG TRANSFORMATION OF DATA

Log transformation of data prior to analysis has become the standard in microarray analysis. While looking at untransformed data may be useful for detecting global correlations or identifying significant changes in gene expression levels, log transformed data provide more reliable and valuable information because they approximate a normal distribution, which is necessary for statistical analyses such as *t*-tests and ANOVA (Figure 1).

When individual samples are available, log transformation allows a geometric mean to be calculated. In contrast, when only pooled samples are available, only the arithmetic mean can be calculated. The results produced are different, as illustrated in the following paragraphs.

The values for an arithmetic mean are inflated compared to geometric mean values. This is because the log transformation performed to derive the geometric mean lessens the effect of high signal value outliers, whereas arithmetic mean values remain susceptible to the effects of such outliers[1]. Figure 2 illustrates the differences between the arithmetic and the geometric mean in this data set. The arithmetic mean signal from the pooled samples was consistently larger than the geometric mean signal from the individual samples. All differences fall below the 45 degree identity line.

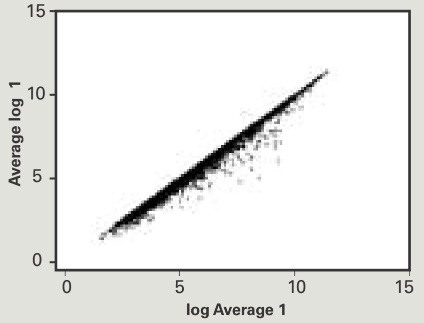
COMPARISON OF CHANGES IN GENE EXPRESSION USING THREE DIFFERENT METHODS OF STATISTICAL ANALYSIS ON POOLED AND INDIVIDUAL SAMPLES

The simplest comparison between pooled and individual samples is to look at the difference between the average signal values of the individual samples and the signal values of the pooled samples (Figure 3A). We call this measure Absolute Difference. For Sample Set 1, Absolute Difference increased with the abundance of the transcript and there were more outliers at the high end of the signal distribution.

A more informative measure is the Proportional Difference (or percentage change) expressed as the Absolute Difference between pooled and individual samples divided by the signal value (Figure 3B). Although the Absolute Difference increased with signal abundance in Figure 3A, Figure 3B shows that, when viewed relative to signal abundance, low to moderate signal values were more affected by the pooling than were high-end signal values.

An indication of the significance of the difference between pooled and individual samples can be found by plotting the Standardized Difference, which is Absolute Difference divided by the variance in the individual samples (Figure 3C). It is not possible to calculate the variance of the

Figure 2. Comparison of pooled and individual samples. The y axis represents the average of the log-transformed signal for the eight arrays in Sample Set 1. The x axis represents the pooled sample in Sample Set 1.



pooled samples, and this is one of the key pieces of information lost during the pooling process.

Figures 3B and 3C indicate that rare transcripts are more influenced by pooling than are abundant transcripts.

IDENTIFYING CANDIDATE GENES USING POOLED AND INDIVIDUAL DATA

The most significant difference between pooled and individual data is the basis on which transcripts of interest are selected. In pooled data, probe sets can only be selected based on magnitude of change. When individual replicates or replicate pools are available, genes may also be selected on the basis of the significance of changes in signal. This allows the selection of transcripts that have small, but statistically significant, changes in abundance. In effect, the biological picture represented by the data is much more sensitive and complete when samples are not pooled. The picture is also more accurate because changes in expression level that are large but not statistically significant can be eliminated. Eliminating such false positives reduces the cost of follow-up experiments and the risk of chasing false leads.

To investigate how different methods of analysis affect which genes are identified as changed, genes that demonstrated changes in expression levels between normal and liver cancer samples (Table 1) were identified by both the Proportional Difference and Standardized Difference methods. The top two percent of transcripts measured by the Proportional Difference method (pooled) and the Standardized Difference method (individual) were compared (Figure 4). Data points above the horizontal line mark the top two percentile of genes selected by the Standardized Difference method (sectors A, B, and C), to be the most reliably altered, while the dashed vertical lines represent the 1st and 99th percentiles of genes selected by the Proportional Difference method, based on the magnitude of the expression change (sectors A, D, C, and F). These selected genes are

Figure 3. Differences between results from pooled and individual arrays from Sample Set 1 (Table 1). Each panel shows log-transformed pooled signal on the x-axis. The y axis represents the following three measures as defined in the text: (**A**) Absolute Difference between the log-transformed signal from Pool 1 and the geometric mean of the individual samples from Sample Set 1. (**B**) Proportional Difference, which is greatest at the low end of abundance, as is (**C**) Standardized Difference.

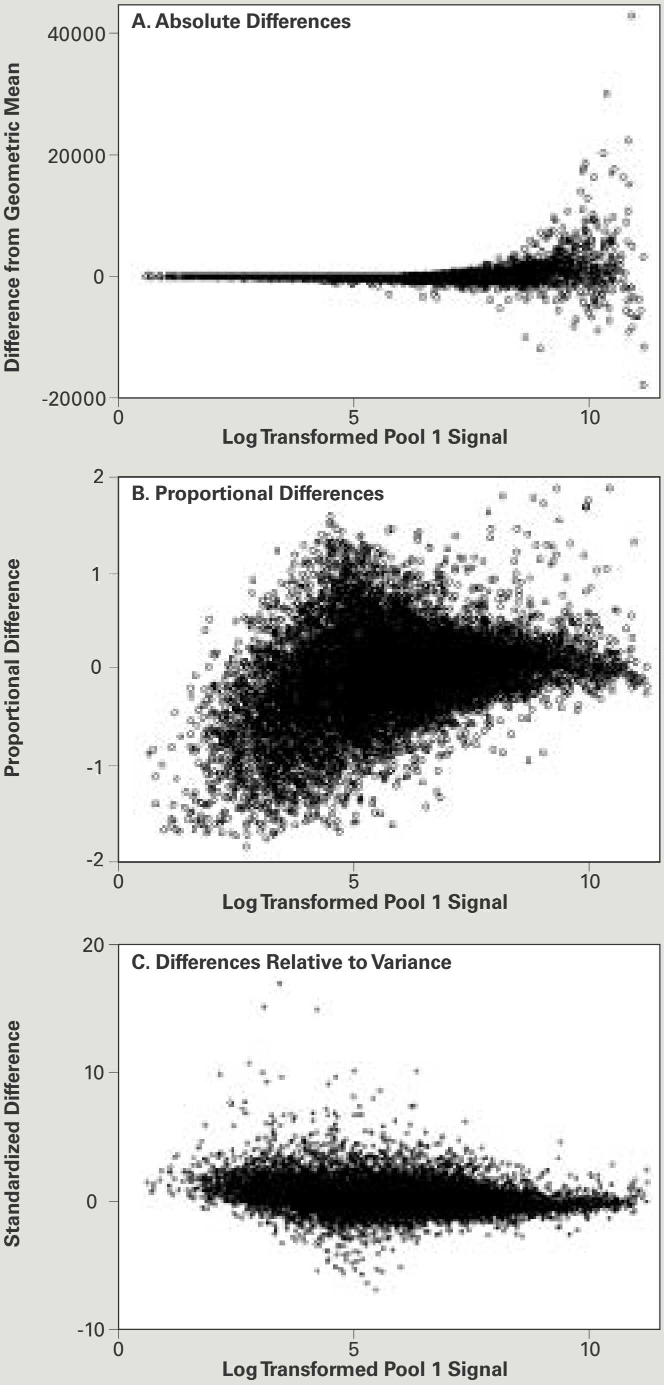
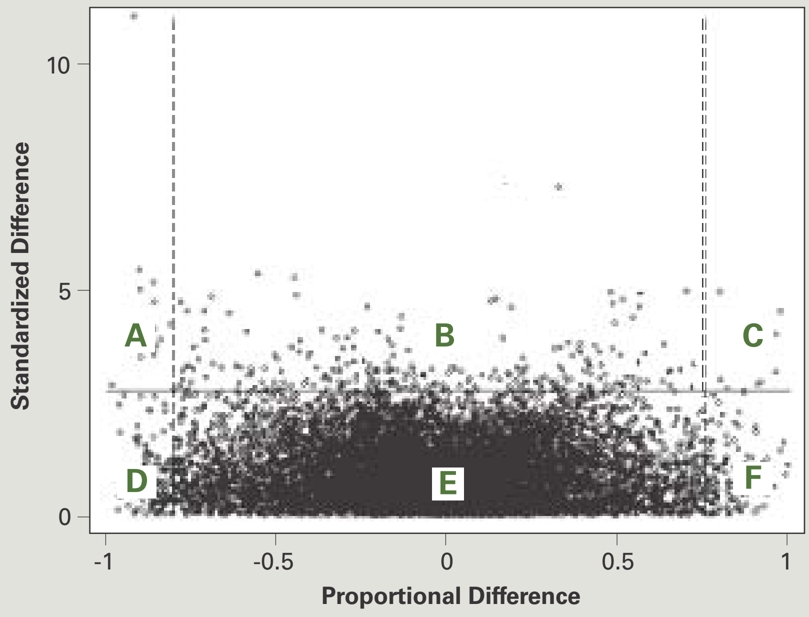


Table 2. Different methods of data analysis affect the interpretation significantly. When the data set was analyzed with the Proportional Difference and Standardized Difference methods, 252 genes were identified as changed by each method. Only 21 of these 252 genes were identified as changed by both methods.

Category	Number of Genes Selected
98th percentile based on Standardized Difference	252
99th or 1st based on Proportional Difference	252
Genes in common	21

Figure 4. The x axis in Figure 4 represents a selection of genes that show a change in expression between normal and cancer samples, based on pools alone, by the Proportional Difference method. Proportional Difference, defined in the text, ranges from -1 to 1. The vertical lines denote the 1st and 99th percentiles. The genes selected by this magnitude-only method fall outside these lines. The y axis represents a selection of genes by the Standardized Difference method, based on individual samples in Sample Set 1 and 2. The horizontal line represents the 98th percentile, or roughly the top 2 percent of genes. Six sectors are designated A, B, C, D, E, F on the diagram as follows: (**A**,**C**) Genes selected by both methods. (**B**) Genes selected only by the Standardized Difference method. (**D**, **F**) Genes selected only by the Proportional Difference method. (**E**) Genes selected by neither method.



noted by their location to the left of the first dashed vertical line and to the right of the second dashed vertical line. The overlap between the genes selected by both methods was low (sectors A and C) (Table 2).

Sectors A and C represent transcripts that were determined to be the most altered by both the Standard and Proportional Difference methods, while sector E represents transcripts that were not determined to be significant by either method.

The key to understanding Figure 4 lies in sectors B, D, and F. These areas of the figure show that, when magnitude-only measurements (Proportional Difference) are used to determine changes in gene expression, transcripts that exhibit small but statistically significant changes are missed (sector B). These may have biological importance, but would not be detected if samples were pooled.

Additionally, according to the Proportional Difference method—which is what researchers must rely on with pooled data—transcripts in sectors D and F appear to be important. However, according to the Standardized Difference method, the transcripts in sectors D and F show changes that are high only in magnitude, but are not statistically significant, meaning that researchers who conduct follow-up studies with these transcripts may be chasing false leads.

Conclusion

In conclusion, informative gene expression data may be obtained from pooled samples, but the limitations of these data should be considered. Pooling results in a loss of sensitivity and an increase in false positives that can significantly alter the biological interpretation of the results.

REFERENCE

[1] J.H. Zar, *Biostatistical Analysis*. 4th edition, Prentice-Hall, N.J., p28.

AFFYMETRIX, INC.

3380 Central Expressway
Santa Clara, CA 95051 USA
Tel: 1-888-DNA-CHIP (1-888-362-2447)
Fax: 1-408-731-5441
sales@affymetrix.com
support@affymetrix.com

AFFYMETRIX UK Ltd

Voyager, Mercury Park,
Wycombe Lane, Wooburn Green,
High Wycombe HP10 0HH
United Kingdom
Tel: +44 (0) 1628 552550
Fax: +44 (0) 1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com

AFFYMETRIX JAPAN K.K.

Mita NN Bldg., 16 F
4-1-23 Shiba, Minato-ku,
Tokyo 108-0014 Japan
Tel: +81-(0)3-5730-8200
Fax: +81-(0)3-5730-8201
salesjapan@affymetrix.com
supportjapan@affymetrix.com

www.affymetrix.com Please visit our web site for international distributor contact information.

For research use only.
Not for use in diagnostic procedures.

Part No. 701494 Rev. 2

©2004 Affymetrix, Inc. All rights reserved. Affymetrix®, GeneChip®, ®, ®, ®, HuSNP®, Jaguar™, EASI™, MicroDB™, GenFlex®, CustomExpress™, CustomSeq™, NetAffx™, 'Tools to take you as far as your vision™', and 'The Way Ahead™' are trademarks owned or used by Affymetrix, Inc. Array products may be covered by one or more of the following patents and/or sold under license from Oxford Gene Technology: U.S. Patent Nos. 5,445,934; 5,744,305; 6,261,776; 6,291,183; 5,700,637; 5,945,334; 6,346,413; 6,399,365; and 6,610,482; and EP 619 321; 373 203 and other U.S. or foreign patents.