The International Journal of Biological Markers, Vol. 21 no. 2, pp. 67-73
© 2006 Wichtig Editore

Questioning the utility of pooling samples in microarray experiments with cell lines

L. Lusa[1,2], V. Cappelletti[1], M. Gariboldi[1,2], C. Ferrario[1,2], L. De Cecco[1,2], J.F. Reid[1,2], S. Toffanin[1], G. Gallus[3,4], L.M. McShane[5], M.G. Daidone[1], M.A. Pierotti[1,2]

[1]Department of Experimental Oncology, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milan
[2]Molecular Genetics of Cancer Group, Fondazione Istituto FIRC di Oncologia Molecolare (IFOM), Milan
[3]Institute of Medical Statistics and Biometry, Università degli Studi di Milano, Milan
[4]Medical Statistics and Biometry, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milan - Italy
[5]Biometric Research Branch, National Cancer Institute, Bethesda, MD - USA

ABSTRACT: We describe a microarray experiment using the MCF-7 breast cancer cell line in two different experimental conditions for which the same number of independent pools as the number of individual samples was hybridized on Affymetrix GeneChips. Unexpectedly, when using individual samples, the number of probe sets found to be differentially expressed between treated and untreated cells was about three times greater than that found using pools. These findings indicate that pooling samples in microarray experiments where the biological variability is expected to be small might not be helpful and could even decrease one's ability to identify differentially expressed genes. (Int J Biol Markers 2006; 21: 67-73)

Key words: Sample pooling, DNA microarrays, Affymetrix GeneChips, Breast cancer, MCF-7 cell line, Toremifene

INTRODUCTION

An important issue that arises when planning a microarray experiment is whether the RNA derived from the samples included in the study should be pooled or should be hybridized individually on the arrays. Although some kind of pooling is inevitable when the RNA quantity obtained from individual samples is insufficient to gain reliable array results (1) or linear amplification is not considered, many investigators wonder if pooling samples is advantageous even when RNA quantity is not limiting. Pooling is often perceived in preclinical studies as an effective way to decrease the number of expensive microarray hybridizations required by reducing biological variability (2, 3).

A frequent aim of microarray experiments is to identify genes that are differentially expressed between two or more prespecified classes, ie, phenotypes or experimental conditions (4). In this setting, pooling of samples can be useful, but in order to properly draw conclusions beyond the specific experiment to populations from which the samples were drawn, proper experimental design must be used, with multiple independent pools for each class, each composed of different units (4, 5).

Several recent papers have addressed statistical design considerations for pooling samples in microarray experiments. Kendziorski et al (6) and Shih et al (7) gave sample size formulae for pooled designs for class comparison experiments showing that a comparable precision or power to a non-pooled design can be obtained by a pooled design with fewer arrays, but increasing the number of individual samples. They also noted that pooling is more effective when the gene variability between samples (biological variability) is considerably larger than the variability across replicated arrays from the same sample (technical variability). Other recent studies have directly evaluated the impact of pooling on the identification of differentially expressed genes between two classes. In particular, Kendziorski and colleagues (8) used Affymetrix arrays in a large study measuring gene expression of both individual samples and independent pools from inbred rats in two different diet conditions, obtaining some technical replicates as well. Their results indicated that pooling reduces overall variability, even though, similarly to Han et al (9), they found that for many genes the technical variability was greater than the biological variability. Others have tried to address the problem of differences in inference with pools and individual samples by comparing sizes of gene lists (10, 11), but none of these experiments was properly designed for its aim. In this paper we focused on the utility of pooling in microarray experiments using cell lines to identify genes differentially expressed between two experimental conditions. We designed a microarray experiment using

0393-6155/067-7$15.00/0

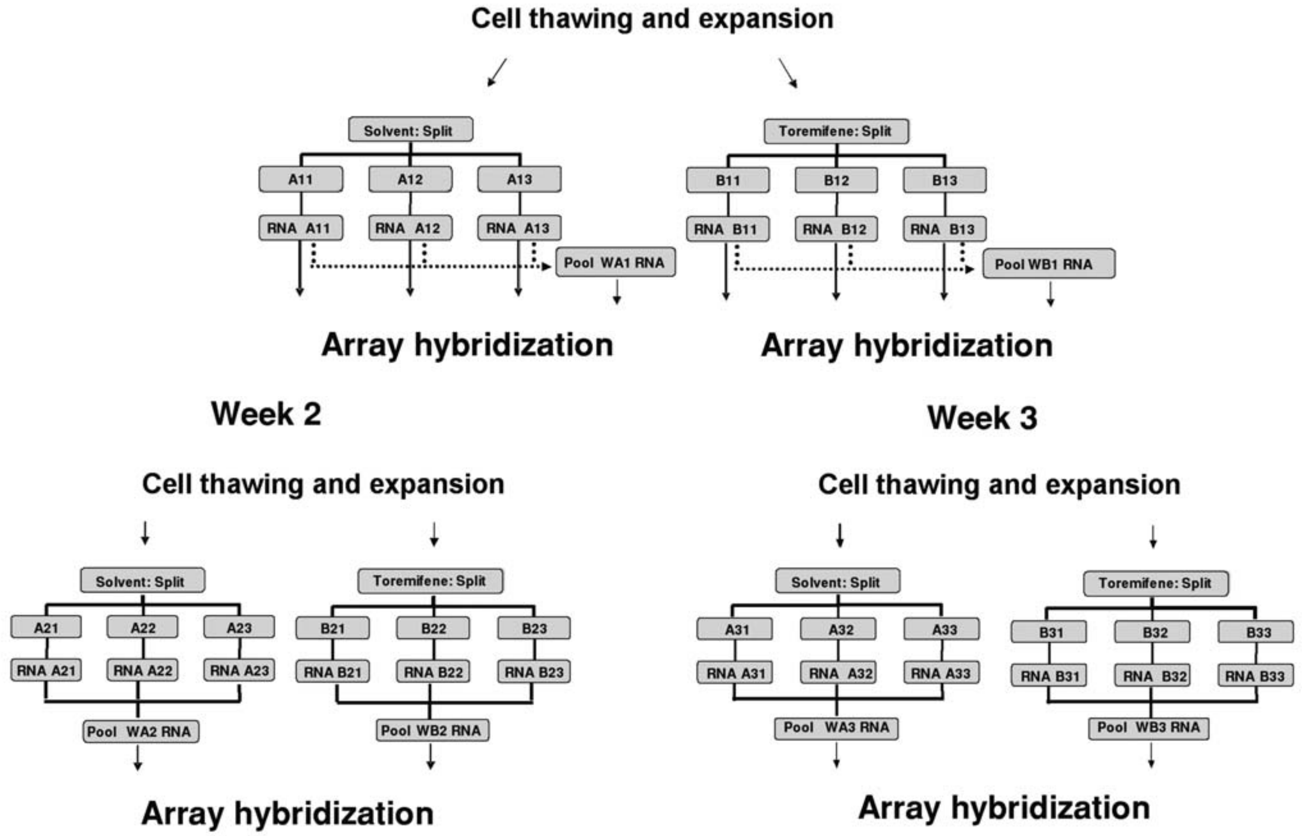

Fig. 1 - *Experimental design showing how pools and individual samples, treated with toremifene (B) or with its solvent (A, untreated samples), were obtained.*

independent individual and pooled samples of RNA hybridized on Affymetrix GeneChip Arrays (Affymetrix, Santa Clara, CA, USA) to assess the differences in gene expression following selective estrogen receptor modulator treatment of the MCF-7 breast cancer cell line. In this situation, the "biological" variability refers to replicate cultures of a single cell line. We used an equal number of arrays for independent and pooled samples, while three times as many individual samples were needed to obtain the pools. We performed a separate analysis of pooled and individual samples and focused on the differences in the lists of differentially expressed probe sets. We list possible reasons to explain our findings and discuss the utility of pooling in microarray experiments when using RNA extracted from repeated cultures of a cell line and in other settings in which the within-class biological variability is expected to be very small.

MATERIALS AND METHODS

Sample preparation

Separate aliquots from the same frozen batch of MCF-7 breast cancer cell line were thawed, amplified, and split into six cultures in three successive weeks. At each time point, cell line samples were treated in triplicate with 10^{-7}M of the antiestrogen 4-hydroxytoremifene (4-OH-TOR) or with its solvent. Cultures were grown independently for nine days and then harvested; RNA was extracted, quantified, checked for integrity and labeled independently for each of the cultures. In the first week individual samples and pools were hybridized on arrays. In the second and third weeks the experiment was independently repeated, but only the pools obtained from each of the three cell line cultures were hybridized. Figure 1 presents a schematic of the experimental design.

Affymetrix GeneChip hybridization and image acquisition

Each target was prepared as described by Borrello et al (12) and hybridized on Affymetrix HG-U133 GeneChip Set (HG-U133A and HG-U133B). Images from the second scan, performed after antibody amplification of the signal at the end of the washing procedure, were used in all the subsequent elaborations.

Data preprocessing and normalization

For our primary analyses, intensity values for each probe

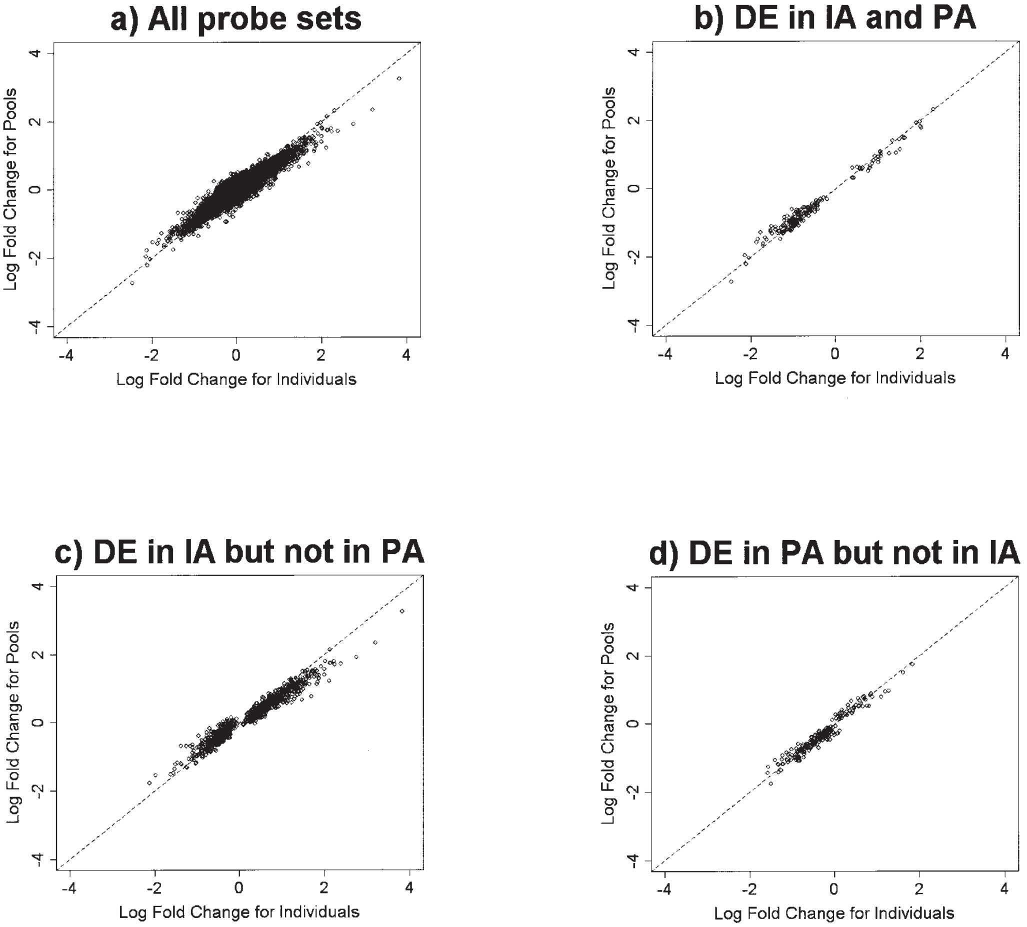

Fig. 2 - *Mean log fold changes for individuals and pools from RMA analysis. a) All probe sets. b) Probe sets differentially expressed in both IA and PA. c) Probe sets differentially expressed in IA but not in PA. d) Probe sets differentially expressed in PA but not in IA.*

set were summarized and normalized with robust multichip analysis (RMA) (13), using the default parameters of RMA function from the Bioconductor Affy package (14). Individual and pooled arrays from both experimental conditions were processed together, merging HG-U133A and HG-U133B arrays after normalization. Similar to the approach described by Kendziorski et al (8), all 44,928 probe sets included in the HG-U133 Set were used in the analysis.

To check the consistency of the results under different preprocessing and data filtering methods, we also applied a filtering criterion in which intensity values with an Affymetrix "Absent" detection call (15) were treated as missing, and intensities were thresholded to a minimum value of 100. Screening of the genes based on minimum variation of log ratios across (*n*) arrays was also considered; any probe set for which the variance (σ^2) was not significantly higher than the median probe-set-specific variance of all probe sets (σ^2_{med}) at a 0.01 significance level was filtered out. (The statistic, $(n\text{-}1)\ \sigma^2/\sigma^2_{\text{med}}$, was assumed to have an approximate chi-square distribution with *n-1* degrees of freedom) (16).

We also obtained intensity signals from Affymetrix Microarray Suite Version 5.0 software (MAS5) (15), scaling all images to a value of 500 and applying the same filtering criteria as described above. All intensities were $\log_2$ transformed.

Statistical methods and data analysis

Probe sets differentially expressed between treated and untreated samples were identified separately for the individual-sample and pooled-sample arrays. Probe sets with a p value less than 0.001 from a univariate two-sample pooled variance *t* test were considered as differentially expressed.

Overall within-class variance can be estimated separately for each probe set using the unbiased estimator $\hat{\sigma}^2 = \sum_i \sum_j (X_{ij} - \bar{X}_j)^2 \,/\, (n_1 - n_2 - 2)$, where X_{ij} is the intensity measurement of the *i*th sample in the *j*th class for a specific probe set and $\bar{X}_j$ is the mean intensity for that probe set in the class *j*; n_1 and n_2 are the number of samples in each class. $\hat{\sigma}_i^2$ and $\hat{\sigma}_p^2$ are the separate estimates obtained

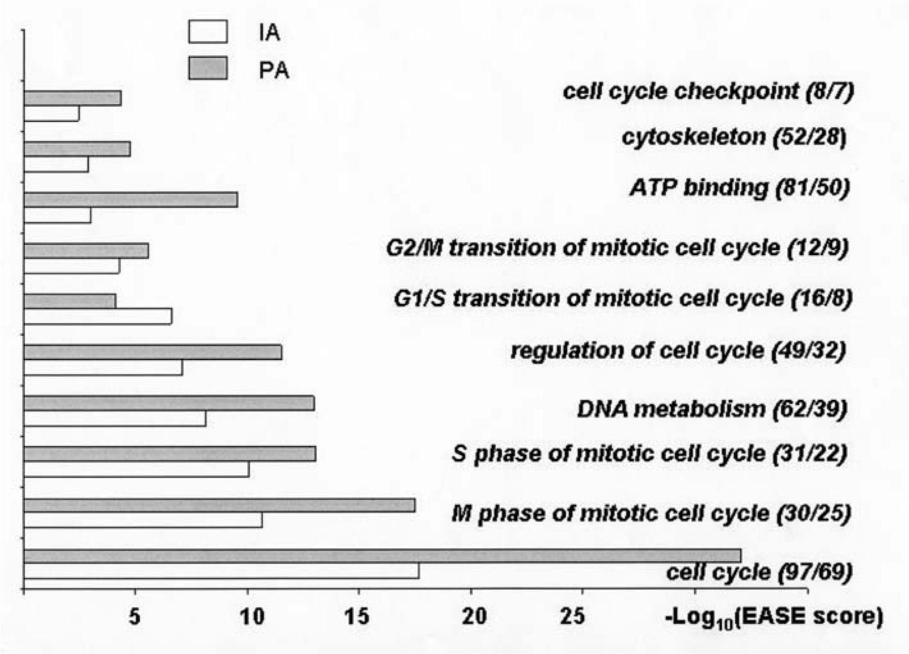

Fig. 3 - *Comparison of some of the significantly enriched biological themes common to IA and PA. For each category the –Log_{10} value of the EASE score is reported. EASE score is a variant of the Fisher exact test and is used to find within a gene list the most overrepresented gene categories. Numbers in brackets represent the number of DE genes in the IA and PA lists, respectively, for each category.*
In the figure only some of the significantly enriched categories common to the two analyses are reported. Roughly the same number of categories were identified as significantly enriched with IA and PA, with some categories exclusive to one of the two analyses. Those apparently relevant to the biological problem are present both in IA and PA, although for PA each category has a lower number of genes.

independently from individual and pool arrays.

Major assumptions underlying pooling include that the gene expression of the pool equals the average expression of the individual samples comprising the pool and that the biological variability in the pool is reduced by a factor equal to the number of samples included in the pool (*r*), while the technical variability is unaffected.

Let σ_b^2 be the biological variance, ie, the variance between true expression of independent individual samples and σ_t^2 be the technical variance, ie, the variance between replicate arrays of the same sample. Overall within-class gene-specific variance is $\sigma_i^2=\sigma_b^2 + \sigma_t^2$ for individual samples, while it is $\sigma_p^2=\sigma_b^2 /r+ \sigma_t^2$ for pools.

Biological and technical variance can then be estimated with the method of moments, according to the formulae $\hat{\sigma}_b^2 = r / (r - 1) (\hat{\sigma}_i^2 - \hat{\sigma}_p^2)$ and $\hat{\sigma}_t^2 = r / (r - 1) \hat{\sigma}_p^2 - 1/(r - 1)\hat{\sigma}_i^2$. With estimates of σ_i^2 and σ_p^2, power can be calculated using the formulae proposed by Shih et al (7).

For a fixed number of arrays and a given significance level, we would expect a pooled design to have greater power because the “biological” variability is reduced. Higher power translates into longer lists of differentially expressed genes.

All analyses were carried out using R statistical language (17) and BRB Array Tools (16).

RESULTS

Quality control of the experiment

Standard metrics for quality control of Affymetrix arrays (18) indicated that the array hybridizations were of good quality. Background signal, row noise score, percentage of Absent Calls and spike-in behavior were consistent across arrays. Concordance between biological replicates was assessed using both graphical and analytical assessments.

Differentially expressed probe sets

The number of probe sets found differentially expressed was 1257 using individual samples (individual analysis, denoted as IA) and 413 using pools (pooled analysis, denoted as PA). One hundred eighty-nine of the significant probe sets were shared by the two analyses (46% of those from PA and 15% from IA). The aforementioned results were based on RMA non-filtered intensities. Similar results, both in terms of the number of differentially expressed probe sets and in terms of the percentage of differentially expressed overlapping probe sets between the two analyses, were found using MAS5 intensities and different preprocessing methods as described in “Materials and Methods”. Better agreement between gene lists from PA and IA was found when we screened out low-variance probe sets (increasing pool overlap to 56% with RMA and to 68% with MAS5 intensities, where about half as many genes were found to be differentially expressed compared to the RMA analysis). The large discrepancy in list length was not a function of the method used to generate the lists. Similar results were obtained when identifying differentially expressed genes by the methods described in (19) and (20).

Average intensities of differentially expressed probe sets from PA and IA analyses were very concordant, both in treated and untreated classes. This was also true for the non-overlapping differentially expressed probe sets from the two analyses. Fold changes between treated and untreated samples showed a good concordance, as shown in Figure 2a, where slight shrinkage of the fold changes of the pools can be observed. Non-overlapping probe sets maintained this characteristic, but the shrinkage of the fold changes was more evident for the probe sets unique to IA, where probes with high and concordant fold changes are missed by the pooled analysis (Fig. 2c-d). Overlapping probe sets did not present small fold changes (Fig. 2b). In general, the PA gene list contained mostly probe sets down-modulated in untreated samples, while up- and down-modulated probe sets found in the IA gene list were more balanced.

EASE (21) was used to identify enriched biological themes within gene lists. Comparison of biological categories revealed a similar ranking of significantly represented themes within the lists obtained from IA and PA. In the IA analysis, 86 of the 604 categories represented in the lists of differentially expressed genes were considered significantly enriched (Fisher exact test, p value less than 0.01), while for PA 81 categories were significantly enriched among the 281 represented. All but 8 of the categories found with PA were present also in the IA lists, while the overlap of significantly enriched categories was limited to 53. For most of the common categories (46/53) the number of differentially expressed genes in each group was greater in IA than in PA. Each biological category can contain from just a few to hundreds of genes; exclusive categories tended to contain fewer genes compared to categories shared by IA and PA and they represented functions not obviously known to be related to selective estrogen receptor modulator treatment.

Figure 3 reports the EASE scores for some of the significantly overexpressed functional categories common to the two gene lists. The list of the differentially expressed genes from IA and PA and the complete EASE outputs are available in the “Supplementary Material”.

Estimation of individual and pool variability and of technical and biological variability

Sixty-two percent of the probe sets showed higher within-class variance estimates when based on individual-sample arrays compared to pooled-sample arrays. The estimated biological variance was negative for this 62% of the probe sets while the estimated technical variance was negative for 10% of the probe sets. (Method-of-moments estimates of variance may be negative when the true variance is near zero). The estimated technical variability was higher than the biological variability for 75% of probe sets, but this percentage was increased to 94% for probe sets identified as differentially expressed in IA and reduced to 58% for genes identified as differentially expressed in PA. Similar results were observed using other preprocessing methods.

Roughly similar proportions of probe sets for which individual samples had a higher variability than pools were obtained when variances were estimated separately for treated and untreated samples, but the identities of the probe sets having higher variance in individuals differed between the two treatment classes (with an overlap of less than half of the probe sets).

DISCUSSION

In this paper we focused on the utility of pooling in microarray experiments using the MCF-7 breast cancer cell line to identify genes differentially expressed following selective estrogen receptor modulator treatment. We designed an experiment in which both individual and pooled samples were measured, using the same number of Affymetrix arrays.

Irrespective of low-level analysis options, we found the number of differentially expressed probe sets from PA to be smaller than from IA (usually less than a third). Even if the number of the arrays was equal in the two experiments, more work and a greater number of individual samples were required to create the pools. In our experimental setting, in which the number of arrays and the significance level were fixed, theoretical arguments would indicate that the PA should have had higher power and provided a larger gene list because of the reduction in biological variance due to pooling. Nevertheless, in our data the expected reduction in overall variance was not observed for more than 60% of the probe sets. Similar observations of increased variability in pools compared to individuals were noted by Kendziorski et al (8) and Han et al (9), who observed 30% and 50% of genes, respectively, that did not benefit from pooling. These two studies analyzed gene expression data from tissues of inbred animals while our experiment used replicate cultures of a cell line, for which an even smaller biological variability would be expected.

There are several possible explanations for our results. First, the variance estimates are based on few observations for each gene (six for the pools and six for the individuals in our experiment); therefore, the estimates may not be reliable and the expected ordering of the variances might be reversed by chance. However, the number of probe sets for which we observed this inversion was higher than we might expect by chance. Assuming no biological variability for all of the genes, we would have expected to obtain a negative estimate of biological variability for 50% of the genes (±0.5% if genes were assumed to be independent). Another suspect might be the pool averaging assumption. Shih et al (7) presented two interesting examples based on real cDNA and Affymetrix data, showing that this assumption may not hold, especially for high signals and more markedly for Affymetrix data. Also, the results of Kendziorski et al (8) indicate that the pooling averaging might not be supported for genes with low variability. Our results showed some departures from the averaging assumption, the nature and magnitude of which depended on data preprocessing and normalization. Unlike Kendziorski et al (8) we did not find that pools were more similar to averages of the samples comprising them than to other samples. A further potential explanation for the failure in variance reduction in pools for most genes might be related to an increase in technical variability due to the additional experimental step of pooling samples. A possible limitation of our study was that pools were obtained in different weeks while individuals were all from the first week. However, when comparing individual samples with the pools containing them, we did not find the individuals to be more similar to their pools than to the pools from different weeks. Therefore, we do not believe that a week effect played an important role.

One interesting difference that we found between IA and PA was that probe sets identified in IA are modulated in both directions (up and down), while PA mainly contains down-modulated probe sets. In a separate microarray experiment performed by us, which involved paired samples from 11 women treated with toremifene, preoperative core-needle biopsies of pretreatment tumor tissue were compared to posttreatment tumor tissue obtained at surgery, and almost the same number of up- and down-modulated genes were observed (data not shown, manuscript in preparation). If the findings of this small, 11-patient study were to hold in general, this would suggest that IA results might resemble the toremifene-induced gene modulation better than PA results.

Analyzing data from our microarray experiment with pooled and individual samples, we found unexpected results, which question the utility of pooling in experiments with small biological variability. Further experiments are needed to isolate possible biases or sources of additional variability in microarray experiments using pooling. If our findings in this paper hold up in further independent studies, our conclusion would be that pooling samples in microarray experiments where the biological variability is expected to be small is not likely to be very helpful, and could even diminish one's ability to identify differentially expressed genes while requiring more experimental time and samples.

ACKNOWLEDGMENTS

We thank the Affymetrix Microarray Unit of IFOM. This study was supported in part by P.S. Oncologia CNR-MIUR and by individual grants from AIRC (Associazione Italiana per la Ricerca sul Cancro) to M. Gariboldi and M.A. Pierotti; further support was received from the Ministero dell'Istruzione, dell'Università e della Ricerca (grant PRIN 2003133820) and from an Italy-USA Fellowship of the Istituto Superiore di Sanità on Oncological Pharmacogenomics-Seroproteomics.

SUPPLEMENTARY MATERIAL (ON REQUEST TO THE EDITOR)

Supplementary file 1: File in Excel format, containing the list of differentially expressed Affymetrix probe sets for IA and PA, and the corresponding EASE outputs.

Address for correspondence:
Lara Lusa
Fondazione Istituto FIRC di Oncologia Molecolare (IFOM)
Via Adamello, 16
20139 Milan, Italy
e-mail: lara.lusa@ifom-ieo-campus.it

REFERENCES

1. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G: The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nat Genet 2001; 29: 389-95.
2. Agrawal D, Chen T, Irby R, et al. Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. J Natl Cancer Inst 2002; 94: 513-21.
3. Enard W, Khaitovich P, Klose J, et al. Intra- and interspecific variation in primate gene expression patterns. Science 2002; 296: 340-3.
4. Simon R, Radmacher MD, Dobbin K. Design of studies using DNA microarrays. Genet Epidemiol 2002; 23: 21-36.
5. Churchill GA. Fundamentals of experimental design for cDNA microarrays. Nat Genet 2002; 32 (Suppl): 490-5.
6. Kendziorski CM, Zhang Y, Lan H, Attie AD. The efficiency of pooling mRNA in microarray experiments. Biostatistics 2002; 4: 465-77.
7. Shih JH, Michalowska AM, Dobbin K, Ye Y, Qiu TH, Green JE. Effects of pooling mRNA in microarray class comparisons. Bioinformatics 2004; 20: 3318-25.
8. Kendziorski C, Irizarry RA, Chen K-S, Haag JD, Gould MN: On the utility of pooling biological samples in microarray experiments. PNAS 2005; 102: 4252-7.

9. Han E-S, Wu Y, McCarter R, Nelson JF, Richardson A, Hilsenbeck SG. Reproducibility, sources of variability, pooling, and sample size: important considerations for the design of high-density oligonucleotide array experiments. J Gerontol A Biol Sci Med Sci 2004; 59: B306-15.
10. Jolly RA, Goldstein KM, Wei T, et al. Pooling samples within microarray studies: a comparative analysis of rat liver transcription response to prototypical toxicants. Physiol Genomics 2005; 22: 46-55.
11. Affymetrix. Sample pooling for microarray analysis – Technical note. San Diego: Affymetrix, 2004.
12. Borrello MG, Alberti L, Fischer A, et al. Induction of a proinflammatory program in normal human thyrocytes by the RET/PTC1 oncogene. Proc Natl Acad Sci USA 2005; 102:14825-30.
13. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4: 249-64.
14. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology 2004; 5: R80.
15. Affymetrix: Statistical algorithms reference guide – Technical note. San Diego: Affymetrix, 2004.
16. Simon R, Lam A: BRB-ArrayTools User Guide, version 3.2. Biometric Research Branch, National Cancer Institute. http://linus.nci.nih.gov/brb.
17. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2005; URL: http://www.R-project.org.
18. Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. Design and analysis of DNA microarray investigations. New York: Springer Verlag, 2003; chapter 5.
19. Wright GW, Simon RM. A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics 2003; 19: 2448-55.
20. Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. J Stat Plann Infer 2004; 124: 379-98.
21. Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol 2003; 4: R70.

Received: February 17, 2006
Accepted: April 28, 2006