Technical Paper ■

Some observations on experimental design of microarray experiments

Lara Lusa

Abstract. Gene-expression microarrays measure simultaneously the expression of thousands of genes and are nowadays widely used in genomic research. The aim of this paper is to give a brief overview of the objectives of microarray gene-expression experiments and to describe some statistical issues related to study design and data preprocessing. Quality control, normalization, replication, validation and use of pooling of independent samples will be discussed.

■ **Infor Med Slov**: 2006; 11(1): 16-24

Author's institution: Istituto Nazionale per lo Sudio e la Cura dei Tumori, Milano and Fondazione Istituto FIRC di Oncologia Molecolare (IFOM).

Contact person: Lara Lusa, Department of Experimental Oncology, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milano and Molecular Genetics of Cancer Group, Fondazione Istituto FIRC di Oncologia Molecolare (IFOM), Via Adamello 16, 20139 Milano, Italy. email: lara.lusa@ifom-ieo-campus.it.

Introduction

Gene-expression microarrays measure simultaneously the expression of thousands of genes and are nowadays widely used in biomedical research to pursue many different objectives.

Microarrays have been used since the end of the 90's.[1,2] Since then, it has become clear that adequate statistical methods play a crucial role in maximizing the potentials of this rapidly evolving field.

In early microarray studies statistics was often misused or not used at all, and this sometimes resulted in scientific contributions that presented results that were unreliable and not reproducible.[3-5]

At the same time, thanks to the extensive use of microarrays in biomedical research, novel statistical methods were developed while many old statistical methods and principles gained new popularity.

The aim of this paper is to give a brief overview of the objectives of microarray gene-expression experiments and to describe some statistical issues related to study design and data preprocessing. Quality control, normalization, replication, validation and use of pooling of independent samples will be disussed. Specific methods for data analysis have been discussed elsewhere[6,7] and will not be covered here.

Objectives and characteristics of gene expression microarray experiments

Most of the objectives of gene-expression microarray experiments can be categorized in three broad classes

- **class discovery objectives**: when the aim is to discover previously unknown subgroups of genes or subjects that are homogeneous in their gene expression (for example, Perou *et al.*[8] proposed a molecular classification of breast cancer identifying different subtypes with distinct gene expression profiles);

- **class comparison objectives**: when the aim is to compare two or more classes (phenotypes or experimental conditions) in terms of gene expression, identifying genes that are differentially expressed between them (for example, Hedenfalk *et al.*[9] compared expression profiles of sporadic breast cancers and of breast cancer tumors with mutations in BRCA1 and BRCA2 genes, and identified the genes that were differentially expressed between the three groups); class comparison can be seen as a special case of all those problems in which it is of interest to evaluate the association of gene expression with other variables such as, for example, expression levels of a biological marker, size of the tumor, survival time;

- **class prediction objectives**: when the aim is to develop classifiers based on expression profiles that predict an outcome (for example, van't Veer *et al.*[10] developed a predictor based on the expression of 70 genes to predict the relapse of breast cancer within 5 years after surgical treatment).

It is not uncommon for microarray experiments to pursue more than one of these aims at the same time.

From a statistical point of view, the most important peculiarity of microarray experiments is the large number of variables (genes) being measured for each subject. On the other hand, in most experiments the sample size (the number of subjects) is still very small. This is known as the "large p, small n" problem,[11] where p is the number of measured variables and n the sample size; due to this characteristic, the straightforward application of standard statistical procedures for data analysis of microarray experiments can be problematic.

The most dangerous consequence of the "large *p*, small *n*" problem in class comparison experiments is the so called *multiple testing problem*. Differentially expressed genes between the classes are generally identified performing hypothesis testing gene by gene. In order to control for false discoveries, several multiple testing procedures, which control in different ways for false discoveries, are available and should be used.[12-14,6,7]

Another consequence of the "large *p*, small *n*" problem is the possibility of easily overfitting data when constructing class predictors using gene expression data. Therefore, when independent data are not available for external validation, the performance of the predictor has to be properly evaluated using cross-validation or bootstrap techniques.[4,6,7]

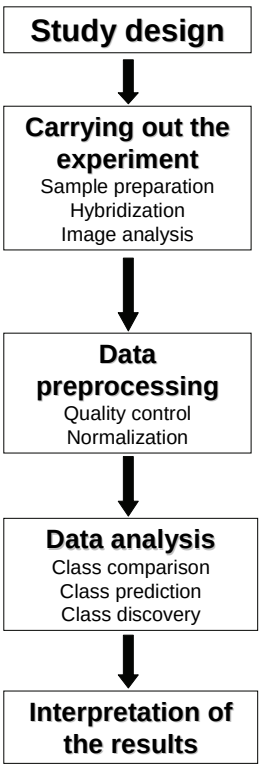

Figure 1 Schematic of the steps of a microarray experiment.

The use of appropriate statistical methods is important in all steps of a microarray experiment, starting from the experimental design. Figure 1 gives a schematic of the steps of a microarray experiment, which include: study design, carrying out of the experiment, quality control, normalization, data analysis and interpretation of the results. It can be noted that after the experiment has been carried out, an additional step is generally required before data analysis.

Preprocessing of the data

Quality control and normalization steps are sometimes referred to as *preprocessing of the data* and they differ substantially between one-channel (Affymetrix Gene Chips ®) and two-channel arrays (two-color cDNA arrays or long oligonucleotide arrays). The reason is related to the many differences between these arrays: the way in which gene expression is measured, the sources of systematic biases, the way image analysis is performed, with different algorithms and different outputs. [15]

Even within the same type of arrays usually there is no general agreement on which method should be used for the preprocessing of the data: many alternatives exist and some of these problems are topics of active statistical research[16-19]. Although the Affymetrix proprietary software includes a program for the preprocessing of GeneChips,[20] many *ad hoc* methods were independently developed for this aim, which generally perform better than the original method. [18]

With the quality control step, genes or samples that were not measured *reliably* are removed from the analysis. Since microarray data are normally very noisy, usually this step reduces greatly the number of genes that are eventually used in the analysis. A lot of useful information on the quality of the measurements can be derived from visual inspection of the images and from image analysis outputs. Generally, small spots, spots with relative large background, with weak or saturated signals are considered unreliable. However, determining, for example, how small a spot should be to be unreliable or, more in general, what a *reliable*

measurement is, is to a great extent arbitrary and not many commonly accepted rules or methods exist.

Within gene-expression microarray experimental procedures, there are many factors that are unrelated to the biological characteristics of the samples but that can influence the outcome of the experiment; the normalization step is aimed at removing these experimental artifacts which can produce systematic biases.

Such experimental artifacts can be due, among other technical reasons, to

- imbalances between RNA amounts,
- RNA amplification,
- RNA degradation,
- retro-transcription efficiency,
- efficiency in dye incorporation or dye fading,
- order in which arrays were hybridized,
- batch to which the slides belong,
- operator that performed the hybridizations or the RNA extraction,
- temperature, humidity or ozone level at the time of the hybridization,
- efficiency of the washing procedure.

Common choices for the normalization of the data include median centering the arrays and methods that adjust the gene expression depending on their intensity or location on the array;[21,17] sometimes just a subset of the genes, which expression is supposed not to change across arrays (*housekeeping genes*), is used for the normalization. However this approach can be difficult to apply since there is no general agreement on how to identify these genes.

Ideally, the normalization process should be incorporated in data analysis rather than separated from it and treated as a preprocessing step.[22] Some attempts in this sense have been made in the context of microarrays,[22-25] mostly using ANOVA models and considering the *nuisance* effects together with the effects of interest. This kind of approach has in principle many advantages, mostly because it does not suppose that no additional error is introduced by the normalization. However, modeling appropriately the nuisance effects can be difficult and computationally challenging and therefore this approach is seldom used in practice.

In general, there is no best method for normalizing data. The choice of a specific normalization method should depend on the characteristics of the data at hand and it should be made after a careful inspection of the data. When data characteristics allow it, the simplest methods should be used, so as to avoid making too many assumptions and limiting the overfitting of the data.

Experimental Design

Proper experimental design plays an essential role for correctly addressing the questions of interest and a clear definition of the objectives of the study is crucial for the correct planning of a microarray experiment.

While experimental design was largely neglected in the early stage of microarray use, its need is now becoming increasingly acknowledged, with a greater emphasis on the importance of replication.[26-28]

Replication

The use of replicates from independent subjects (*biological replicates*) cannot be avoided when the aim of the experiment is finding results that can be extended from the samples being analyzed to the populations to which they belong, *i.e.* drawing proper statistical inference. Multiple

measurements of the same subject (*technical replicates*) do suffice only in quality control studies, where just the evaluation of the reproducibility of the measurements and of the error associated with the array process is of interest.[27,28]

The distinction between biological and technical replicates has been a source of substantial confusion in early microarray experiments, where often only technical replicates were used.[27] Misuses are still frequent in experiments with cell lines or inbred animals, where biological variability is supposed to be negligible or very small.

Methods for the calculation of the number of independent biological replicates needed in a microarray experiment depend on the aim of the experiment, on the method used for data analysis and, in two-channel arrays, on the way samples are allocated to the arrays.

The main feature of two-channel arrays is that two samples are hybridized on the same array. Many alternative designs exist,[29] but the simplest way to allocate samples on the arrays is to use a *reference design*, in which an aliquot of a reference sample is labeled with the same label and hybridized on each array. This design has many advantages over its alternatives:[29] it allows the direct comparisons of any subset of arrays in class comparison problems, also across experiments if the same reference was used. Moreover, it is robust to the presence of bad quality arrays and data can be straightforwardly used also in class discovery and class prediction problems. Last but not least, it is easy to perform in the lab.

A disadvantage of the reference design is that half of the hybridizations are used for the reference sample, for which usually there is no biological interest. Other designs which allocate samples more efficiently or that have some *optimality* properties have been proposed, examples being the balanced block design,[28] the loop design[23] and the interwoven loop.[30] These methods for sample allocation are less flexible, less suited for class discovery problems, they require more complex methods for the analysis of data and are more sensitive to the presence of bad quality arrays.

For class comparison studies, methods based on power analysis and depending on the method in which samples are allocated on the arrays have been proposed by Dobbin and Simon,[29,31] which used classical statistical sample size reasoning, taking into account the multiple testing problem, and reviewed[31] previously proposed methods for sample size calculation for microarray experiments. These methods usually require some knowledge on gene variability in the population of interest and on the variability of the experimental error, both of which might be available from previous experiments or can be estimated by pilot studies.

Sample size estimation methods have been developed also for class prediction problems and are more complex, having to take into account the variability deriving from both the predictor construction and the choice of the genes to be included in the predictor (*feature selection*).[32-34]

Randomization and confounding

As mentioned in the section on the preprocessing of data, in microarray gene-expression experiments many factors that are unrelated to the characteristics of the samples can influence the outcome of an experiment. Normalization is generally used to correct for these effects. There are however some situations in which, due to bad design of the experiment, normalization cannot be effectively used for its purpose.

As an example, suppose that in an experiment all the samples of normal tissue are hybridized in one day, while all the tumor samples are processed on the following day, in which the level of the humidity unexpectedly and dramatically rises. When looking for the genes that are differentially expressed between normal and tumor tissue, it will be impossible to identify the genes that have a different expression in the two types of tissues from those which change was influenced by the humidity level.

Given the strong influence that experimental factors can have on the hybridization results, the usual principles of statistical study design should be applied also to microarray experiments, the most basic of which is the randomization of the samples to the levels of the known confounding factors.

Even though some methods for correcting for batch effects exist,[35] in most cases some care in the experimental design can avoid their use and make data analysis simpler and not dependent on additional assumptions.

Planning the validation of the results

In class comparison experiments a very common practice consists in validating the results obtained from the analysis of microarray data with a different technology, usually with real time – quantitative polymerase chain reaction (RT-QPCR).

Most of the times the validation is performed on the same samples that were used in the microarray experiment. In this case what is being validated is merely the validity of the microarray measurements and therefore this approach cannot be seen as a way to improve the confidence on the generalizability of the results. Moreover, the genes that are not identified as being differentially expressed from the microarray experiment are hardly ever validated, so this kind of validation can identify false positive but not false negative results.

When an independent set of samples is used to validate the findings of a microarray experiment, comparisons with the original findings are generally made comparing P-values rather than the sizes of the effects.

In class prediction problems the validations of the results is usually made evaluating the predictive accuracy of the model using cross-validation or bootstrap.[4,6,7]

When an independent set of samples is used to validate a predictive model developed from gene-expression microarray data, it is important that the predictive model is completely specified before applying it to the new data set.[36] This avoids running the risk of overfitting the model on the new data and obtaining biased estimates of the predictive accuracy. This problem is particularly interesting when RT-PCR or custom arrays are used to measure the genes that were included in the predictive model developed from original microarray data. In this case, since the measurements are made using different methods of measurements it seems legitimate to re-estimate the model on the independent data set. However, this cannot be claimed to be a completely independent validation and the model developed on the independent data set can still be prone to overfitting.

Pooling of samples

Whether to pool samples and hybridize them instead of individual samples on the arrays is another option in the design of gene-expression microarray experiments.

Pooling the RNA of independent samples is a necessary choice in microarray experiments where the amount of available RNA from each sample is not sufficient for obtaining a good quality array[37] and RNA amplification is not considered.

Even when the RNA quantity is not a concern, investigators often consider pooling of samples as a choice when designing their class comparison studies.[38,39] Pooling is seen as an effective way to cut the costs of the expensive microarray experiments, while reducing the biological variability, at the cost of loosing the individual information.

Many investigators have been misusing pooling, typically obtaining one pool per condition and then hybridizing one or multiple aliquots of the pools on the arrays. As noted above, independent biological replicates are needed in order to make inference on the populations to which samples

belong. Therefore, multiple independent pools for each class, each composed by different units, must be used in order to be able to extend the experimental findings from the sample to the classes to which the pools belong.[27]

Recently some papers addressed the issue of sample size requirements for microarray class comparison experiments with pooled samples and compared pooled and individual samples designs. It was shown that increasing the number of independent subjects included in the study, comparable precision or power to a non-pooled design can be obtained by a pooled design with fewer arrays.[40,41]

However, there is some experimental evidence that the major assumption underlying pooling, namely that the gene expression of the pool equals the average expression of the individual samples in that pool, may not hold. Shih *et al.*[41] showed that gene-expression of the pool can significantly differ from the average expression of the individual samples, especially for high signals and more markedly for Affymetrix data. Moreover, experimental data showed that the expected reduction of overall within-class variability in pooled samples can be observed for only a part of the genes[42-44](from 70 per cent to 40 per cent).

Kendziorski *et al.*[42] after a comprehensive experimental comparison of pooled and non-pooled experimental results, recommend that "pooling be done when fewer than 3 arrays are used in each condition". However, more than 2 independent replicates per condition should be used in each microarray experiment in order to apply statistical methods for the analysis of data, therefore the utility of pooling seems limited.

Especially when the biological variability of the samples is expected to be small compared to technical variability, pooling is not likely to be beneficial and a large number of individual samples is required in order to be able to reduce the number of arrays without loosing power.

Conclusions

The focus of this paper was mainly on experimental design, which constitutes a fundamental but often neglected step in each microarray experiment. This aspect, together with thoughtful validation of the results is crucial for transferring the discoveries of this powerful tool into clinical applications.

Acknowledgements

This work was partially supported by an Italy-U.S.A. Fellowship of the Istituto Superiore di Sanità on Oncological Pharmacogenomics-Seroproteomics. I would like to thank James F. Reid for helpful discussion.

References

1. Schena M, Shalon D, Davis RW, et al.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; 270 (5235): 467-470.
2. Lockhart DJ, Dong H, Byrne MC, et al.: Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol. 1996;14(13):1675-80.
3. Lee M-LT, Kuo FC, Whitmore GA, et al.: Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. Proceedings of the National Academy of Sciences U S A 2000; 97(18):9834-9839.
4. Simon R, Radmacher MD, Dobbin K, et al.: Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute 2003; 95(1): 14-18.
5. Ioannidis JPA: Microarrays and molecular research: noise discovery? Lancet 2005; 365(9458): 454-455.
6. Simon RM, Korn EL, McShane LM, et al.: Design and analysis of DNA microarray investigations. New York (NY) 2004; Springer-Verlag.
7. Speed T, editor: Statistical analysis of gene expression microarray data. Boca Raton (FL) 2003; Chapman & Hall/CRC.
8. Perou CM, Sorlie T, Eisen MB, et al.: Molecular portraits of human breast tumours. Nature 2000; 406: 747-52.

9. Hedenfalk I, Duggan D, Chen Y, et al.: Gene-expression profiles in hereditary breast cancer. N Engl J Med. 2001; 344(8): 539-48.
10. van't Veer LJ, Dai H, van de Vijver MJ, et al.: Gene expression profi ling predicts clinical outcome of breast cancer. Nature 2002; 415: 530-6.
11. West M: Bayesian factor regression models in the "large p, small n" paradigm. Bernardo JM, Bayarri M, Berger J, Dawid A, Heckerman D, Smith A,West M (Eds.), Bayesian Statistics 7, Oxford University Press, 2003, pp. 723-732.
12. Benjamini Y, Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Statist. Soc. Ser. B 1995; 57: 289-300.
13. Korn EL, Troendle JF, McShane LM, et al.: Controlling the number of false discoveries: application to high-dimensional genomic data. Journal of Statistical Planning and Inference 2004;124(2):379-398.
14. Storey JD, Tibshirani R: Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100(16):9440-5.
15. Holloway AJ, van Laar RK, Tothill RW, et al.: Options available--from start to finish--for obtaining data from DNA microarrays II. Nat Genet. 2002;32 Suppl:481-9.
16. Li C, Wong W: Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proceedings of the National Academy of Science U S A 2000; 98: 31-36.
17. Irizarry RA, Hobbs B, Collin F, et al.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4:249-264.
18. Cope LM, Irizarry RA, Jaffee HA, et al.: A benchmark for Affymetrix GeneChip expression measures. Bioinformatics 2004;20(3):323-31.
19. Wu Z, Irizarry R, Gentlemen R, et al.: A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association 2004; 99(468): 909-917.
20. Affymetrix: Statistical algorithms reference guide:Technical report, 2001; Affymetrix.
21. Yang YH, Dudoit S, Luu P, et al.: Normalization of cDNA microarray data: a rubust composite method addressing single and multiple slide systematic variation. Nucleic Acid Research 2002; 30:4:e15.
22. Wu Z, Irizarry RA: A Statistical Framework for the Analysis of Microarray Probe-Level Data. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 73 2005; http://www.bepress.com/jhubiostat/paper73.
23. Kerr M, Afshari C, Bennett L, et al.: Statistical analysis of a gene expression microarray experiment with replication. Statistica Sinica 2002; (12): 203-217.
24. Kerr M, Martin M, Churchill GA: Analysis of variance for gene expression microarray data. Journal of Computational Biology 2000; 7: 819-837.
25. Wolfinger R, Gibson G, Wolfinger E, et al.: Assessing gene significance from cDNA microarray expression data via mixed models. Journal of Computational Biology 2001; 8(6): 625-637.
26. Yang YH, Speed T: Design issues for cDNA microarray experiments. Nat Rev Genet. 2002; 3(8): 579-88.
27. Churchill GA: Fundamentals of experimental design for cDNA microarrays. Nat Genet 2002; 32 Suppl: 490-5.
28. Simon R, Radmacher MD, Dobbin K: Design of studies using DNA microarrays. Genet Epidemiol. 2002; 23: 21-36.
29. Dobbin K, Simon R: Comparison of microarray designs for class comparison and class discovery. Bioinformatics. 2002; 18(11): 1438-45.
30. Wit E, McClure JD: Statistics for Microarrays; Design, Analysis and Inference, Chichester 2004; John Wiley & Sons.
31. Dobbin K, Simon R: Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 2005; 6: 27-38.
32. Dobbin K, Simon R: Sample size planning for developing classifiers using high dimensional DNA microarray data. Biostatistics 2006; In Press. doi:10.1093/biostatistics/kxj036
33. Fu WJ, Dougherty ER, Mallick B, et al.: How many samples are needed to build a classifier: a general sequential approach. Bioinformatics 2005; 21: 63-70.
34. Mukherjee S, Tamayo P, Rogers S, et al.: Estimating dataset size requirements for classifying DNA microarray data. Journal of Computational Biology 2003; 10: 119-142.
35. Johnson WE, Rabinovic A, Li C: Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics 2006; In Press. doi:10.1093/biostatistics/kxj036

36. Simon R: Development and Validation of Therapeutically Relevant Multi-Gene Biomarker Classifiers. J Natl Cancer Inst 2005; 97: 866-7.
37. Jin W, Riley RM, Wolfinger RD, et al.: The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nature Genetics 2001; 29: 389-395.
38. Agrawal D, Chen T, Irby R, et al.: Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. Journal of the National Cancer Institute 2002; 94: 513-21.
39. Enard W, Khaitovich P, Klose J, et al.: Intra- and interspecific variation in primate gene expression patterns. Science 2002; 296: 340-343.
40. Kendziorski CM, Zhang Y, Lan H, et al.: The efficiency of pooling mRNA in microarray experiments. Biostatistics 2004; 4: 465-477.
41. Shih JH, Michalowska AM, Dobbin K, et al.: Effects of pooling mRNA in microarray class comparisons. Bioinformatics 2004; 20: 3318-3325.
42. Kendziorski C, Irizarry RA, Chen K-S, et al.: On the utility of pooling biological samples in microarray experiments. Proceedings of the National Academy of Sciences U S A 2005; 102: 4252-4257.
43. Han ES, Wu Y, McCarter R, et al.: 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: B306-315.
44. Lusa L, Cappelletti V, Gariboldi M, et al.: Caution regarding the utility of pooling samples in microarray experiments with cell lines. International Journal of the Biological Markers 2006; In Press.