

Getting Started in Gene Expression Microarray Analysis

Donna K. Slonim^{1,2*}, Itai Yanai^{3*}

1 Department of Computer Science, Tufts University, Medford, Massachusetts, United States of America, **2** Department of Pathology, Tufts University School of Medicine, Boston, Massachusetts, United States of America, **3** Department of Biology, Technion-Israel Institute of Technology, Technion City, Haifa, Israel

Gene expression microarrays provide a snapshot of all the transcriptional activity in a biological sample. Unlike most traditional molecular biology tools, which generally allow the study of a single gene or a small set of genes, microarrays facilitate the discovery of totally novel and unexpected functional roles of genes. The power of these tools has been applied to a range of applications, including discovering novel disease subtypes, developing new diagnostic tools, and identifying underlying mechanisms of disease or drug response. However, this technology necessarily produces a large amount of data, challenging us to interpret it by exploiting modern computational and statistical tools. In this brief review, we aim to indicate the major issues involved in microarray analysis and provide a useful starting point for new microarray users. Figure 1 outlines the steps in a typical expression microarray experiment and maps them to the different sections of this review.

Experimental Design

Careful experimental design is crucial for a successful microarray experiment [1,2], yet this important step is often shortchanged. Design issues depend in part on the exact array technology used, and indeed, choosing an array technology is often the first design choice. The main distinction is whether essentially full-length transcripts are printed onto slides (cDNA microarrays) or the desired—typically shorter—oligonucleotides are synthesized in situ (oligonucleotide arrays). While the former may be less expensive because they can be manufactured in the lab or at institutional core facilities, the latter may outperform the former in terms of number of spots per array and the spots' homogeneity [3,4].

Slightly different oligonucleotide array platforms are manufactured by companies such as Affymetrix, Agilent, and NimbleGen (see Text S1 and Table S1 for further discussion).

A major design question is whether to measure the expression levels from each sample on a different microarray (using single-color, or single-channel, arrays), or

instead to compare relative expression levels between a pair of samples on each microarray (two-color or two-channel arrays). There are tradeoffs between the two approaches. Single-color arrays allow for more flexibility in analysis, while two-color arrays can control for some technical issues by allowing a direct comparison in a single hybridization [5]. A recent comparison of single- and two-color methods on the same platforms found good overall agreement in the data produced by the two methods [6]. cDNA arrays typically involve two channels. Agilent and NimbleGen arrays can be run using either one or two channels. Affymetrix arrays are inherently single-channel, though some associated analysis tools facilitate pair-wise comparisons.

Design issues for two-color arrays are more complex [7]. Challenges include ensuring that all samples can be compared to the appropriate controls and avoiding any biases introduced by the different labeling. “Dye-swap” experiments, in which the same pairs of samples are compared twice with the labeling colors swapped, can permit the computational removal of such bias. Dye swapping imposes additional costs in both the number of arrays and the types of data analyses possible. However, clever design can somewhat reduce the required number of arrays [1].

As attractive as it might seem financially to run just one microarray for each “class” of samples (of the same phenotype, time-point, or tissue type) under consideration, replicates are essential for providing meaningful results [2]. Without replicates, no statistical analysis of the significance and reliability of the observed changes is

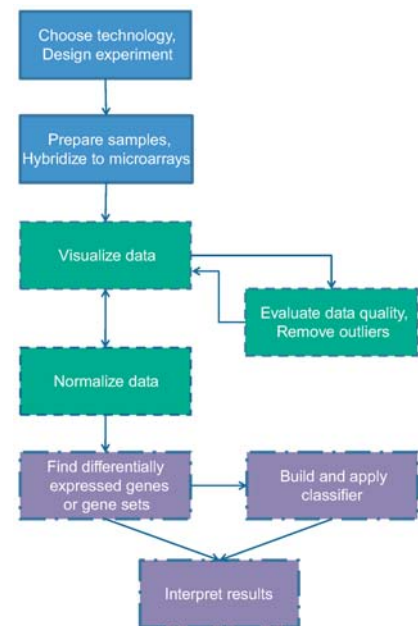


Figure 1. Overview of steps in a typical gene expression microarray experiment.

Topics in blue boxes with solid borders are addressed in the Experimental Design section, those in green boxes with dashed borders are covered in the section on data preparation, and those in purple boxes with dash-dotted borders are discussed in the Data Analysis section of this review.

doi:10.1371/journal.pcbi.1000543.g001

possible; the typical result is an increased number of both false-positive and false-negative errors in detecting differentially expressed genes [8]. However, we distinguish between technological and biological replicates. Technological replication—the same biological material hybridized independent times—is generally no longer

Citation: Slonim DK, Yanai I (2009) Getting Started in Gene Expression Microarray Analysis. *PLoS Comput Biol* 5(10): e1000543. doi:10.1371/journal.pcbi.1000543

Editor: Olga G. Troyanskaya, Princeton University, United States of America

Published: October 30, 2009

Copyright: © 2009 Slonim, Yanai. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: DKS is supported in part by NIH grants LM009411 and HD058880. IY is a Horev Fellow, supported by the Taub Foundations. The funders had no role in the preparation of the article.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Donna.Slonim@tufts.edu (DKS); yanai@technion.ac.il (IY)

performed, as analyses have shown that the results will be relatively consistent overall [4], although they may include consistent sources of bias [2]. Instead, different patients or animals from the same class can serve as biological replicates. To improve the ability to detect outliers and their effects, we do not recommend pooling samples unless necessary to obtain sufficient amounts of material for hybridization, and even then, replicates measuring different pools with the same phenotypes must be performed [7].

During the experimental design stage, it is important to identify all the variables to be compared and to ensure that the proposed design allows their measurement. Be aware of other variables, such as patient age or date of sample collection, that might confound the distinction between the compared classes. One option is to randomize confounding variables related to experimental conditions under your control.

Preparing Microarray Data for Analysis

The task of analyzing microarray data is often at least as much an art as a science, and it typically consumes considerably more time than the laboratory protocols required to generate the data. Part of the challenge is assessing the quality of the data and ensuring that all samples are comparable for further analysis.

Normalization of the raw data, which controls for technical variation between arrays within a study, is essential [7]. The challenge of normalization is to remove as much of the technical variation as possible while leaving the biological variation untouched. This is a big challenge, and here we only touch upon the main issues. First, visualization of the raw data is an essential part of assessing data quality, choosing a normalization method, and estimating the effectiveness of the normalization. Many methods for visualization, quality assessment, and data normalization have been developed (see [9] for a review, Text S1, and Figure S1). Related issues of background adjustment and data “summarization” (reducing multiple probes representing a single transcript to a single measurement of expression) for Affymetrix arrays are well introduced in chapter 2 of [10].

Clustering is a way of finding and visualizing patterns in the data. Many papers and indeed books have been written on this topic (see e.g., [11–13] and Text S1). Different methods highlight different patterns, so trying more than one

method can be worthwhile. Note that while clustering finds predominant patterns in the data, those patterns may not correspond to the phenotypic distinction of interest in the experiment. To identify gene expression patterns related to this distinction, more directed methods are appropriate.

Data Analysis

There are many commercial packages for microarray analyses, and we have by no means evaluated all of them. However, commercial tools can be expensive, and we find many that we have tried to have limited flexibility. Fortunately, in the past few years a number of Web-based tools and open-source software packages for microarray data analysis have become available (see below and Text S1), and we recommend taking advantage of them.

One common strategy is to create a custom data analysis pipeline using statistical analysis software packages such as Matlab or R. Both allow great flexibility, customized analysis, and access to many specialized packages designed for analyzing gene expression data. Not only is R freely available, but it also allows the use of BioConductor [14], a collection of R tools including many powerful current gene expression analysis methods written and tested by experts from the growing microarray community.

The fundamental goal of most microarray experiments is to identify biological processes or pathways that consistently display differential expression between groups of samples. While the exact approach depends in part on the design of the experiment, there are two broad approaches to detecting differential expression. The first examines each gene or transcript individually to find genes that, by themselves, have statistically significant differences in expression between samples with different phenotypes or characteristics. The set of genes thus identified is then examined for over-representation of specific functions or pathways [15]. A powerful alternative is to identify groups of functionally related genes ahead of time and to test whether these gene sets—as a group—show differential expression [16–18]. Both of these approaches can be effective, and sometimes the combination of the two is stronger than either alone [19].

One crucial issue for all microarray analysis methods is adjusting for multiple testing [20]. Each statistical test reports the probability of seeing the observed test score by chance under the null hypothesis

that there is no difference in expression related to the phenotype being studied. Even if this reported “*p*-value” is low, say 0.001, one might expect to see 20 of these one-in-a-thousand events when performing 20,000 independent tests (a reasonable number of genes on a microarray). A range of methods to adjust for multiple testing are available (see [21] for an overview). The preferred approach for microarray analysis is to control the “false-discovery rate” (FDR): the probability that any particular significant finding is a false positive [22].

Once a list of differentially expressed genes has been assembled, some functional analysis is essential for interpreting the results. There are many tools available to identify pathways or biological functions that are over-represented in a given gene list. Again, adjustment for multiple testing may be desirable, although complex dependencies between pathways make finding an appropriate adjustment method controversial [23]. A good review of the earlier tools that discusses many of the statistical issues is [15].

An alternative to the individual-gene analysis workflow is to consider entire gene sets or pathways together when looking for differential expression. There are many approaches that do this (e.g., [16,24–26]), but a fundamental and widely used version is the Gene Set Enrichment Analysis (GSEA) software from the Broad Institute [17]. Gene set analysis can be advantageous because it can detect subtle changes in gene expression that individual gene analyses can miss, and because it combines identification of differential expression and functional interpretation into a single step.

The disadvantage of this method is that appropriate gene sets need to be known ahead of time. When studying a biological process that is still poorly understood, an individual gene method may be more appropriate, as it allows for the opportunity of implicating hitherto unexpected genes and gene sets. Given that gene set analysis is more sensitive and therefore potentially more powerful, a greater effort in defining the pathways needed to support this approach is warranted. Toward this end, GSEA’s gene set database incorporates some computationally derived gene sets, including expression neighbors of known cancer genes [17] and network modules mined from a large collection of expression data [27]. Related work has used conserved coexpression [28] or differential coexpression [29] to discover new functional modules.

Much has also been written about sample classification using microarray data

(see review [13]) but, with a few exceptions [30,31], microarrays themselves have not been embraced as diagnostic tools. Rather, they have been used to identify smaller sets of predictive genes or pathways that might, when assessed by other technologies, aid in diagnosis or stratification of samples. A huge range of machine learning methods [11,12] can be applied to the related classification problems. Most people intent on doing this write their own code (but see Text S1 for an alternative). We note that simpler classification tools often perform as well as, and generalize better than, more complex ones [32].

Outlook

It has been our goal in this brief review to demonstrate that it is currently feasible for researchers with no previous experience to incorporate microarray analyses in their studies. The field is now reasonably mature, with available software and tools to make data analysis manageable by nonexperts. That said, newcomers to the field should be aware that the data analysis will require a dedicated commitment of time and effort that generally substantially exceeds that of data generation. We strongly recommend that researchers do the work to familiarize themselves with the

relevant analytical literature before beginning, or even designing, the experiment.

It has been speculated that microarray technology will soon be superseded by next-generation sequencing, in which the transcripts are directly sequenced by low-cost, high-throughput sequencing technologies [33]. However, currently, next-generation whole-transcriptome sequencing is still quite expensive and in its relative infancy. Its cost scales proportionally with its ability to assess low-abundance transcripts, as sufficient depth of sequencing must be performed. Further, analytic tools specific to this data source have not yet been developed for mass consumption. Recent studies have shown that the two transcriptomics technologies are expected to give very similar results [34,35], although for rare transcripts there is considerably less correlation between the methods [35]. Thus, until sequencing-based methods have become cost-effective and easily used, microarrays will remain a desirable alternative for many practitioners. We expect that, as RNA sequencing methods mature, many microarray analysis methods will come to be viewed as general analysis tools that can be applied or modified to fit any forthcoming transcriptomics technologies [36].

References

- Kerr MK, Churchill GA (2001) Experimental design for gene expression microarrays. *Biostatistics* 2: 183–201.
- Yang YH, Speed T (2002) Design issues for cDNA microarray experiments. *Nat Rev Genet* 3: 579–588.
- Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, et al. (2005) Multiple-laboratory comparison of microarray platforms. *Nat Methods* 2: 345–350.
- Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, et al. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 24: 1151–1161.
- Altman NS, Hua J (2006) Extending the loop design for two-channel microarray experiments. *Genet Res* 88: 153–163.
- Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu TM, et al. (2006) Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nat Biotechnol* 24: 1140–1150.
- Simon RM, McShane LM, Korn EL, Radmacher MD (2003) *Design and Analysis of DNA Microarray Investigations* Springer.
- Lee ML, Kuo FC, Whitmore GA, Sklar J (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A* 97: 9834–9839.
- Quackenbush J (2002) Microarray data normalization and transformation. *Nat Genet* 32 Suppl: 496–501.
- Gentleman RC, Carey VJ, Huber W, Irizarry R, Dudoit S (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer.
- Duda RO, Hart PE, Stork DG (2001) *Pattern Classification*. New York: John Wiley & Sons.
- Mitra S, Datta S, Perkins T, Michailidis G (2008) *Introduction to Machine Learning and Bioinformatics*. London: CRC Press.
- Slonim DK (2002) From patterns to pathways: gene expression data analysis comes of age. *Nat Genet* 32 Suppl: 502–508.
- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5: R80.
- Khatri P, Draghici S (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21: 3587–3595.
- Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20: 93–99.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545–15550.
- Tomfohr J, Lu J, Kepler TB (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 6: 225.
- Slonim DK, Koide K, Johnson KL, Tantravahi U, Cowan JM, et al. (2009) Functional genomic analysis of amniotic fluid cell-free mRNA suggests that oxidative stress is significant in Down syndrome fetuses. *Proc Natl Acad Sci U S A* 106: 9425–9429.
- Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science* 18: 71–103.
- Westfall PH, Young SS (1993) *Resampling-based multiple testing: examples and methods for P-value adjustment*. New York: Wiley. pp xvii, 340.
- Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19: 368–375.
- Gentleman R (2004) Using GO for statistical analyses COMPSTAT. pp 171–180.
- Efron B, Tibshirani R (2007) *On Testing the Significance of Sets of Genes*. *Journal of Computational and Theoretical Nanoscience* 1: 107–129.
- Kim SY, Volsky DJ (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 6: 144.
- Pavlidis P, Lewis DP, Noble WS (2002) Exploring gene expression data with class scores. *Pac Symp Biocomput*. pp 474–485.
- Segal E, Friedman N, Koller D, Regev A (2004) A module map showing conditional activity of expression modules in cancer. *Nat Genet* 36: 1090–1098.
- Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302: 249–255.
- Kostka D, Spang R (2004) Finding disease specific alterations in the co-expression of genes. *Bioinformatics* 20 Suppl 1: i194–199.
- Glas AM, Floore A, Delahaye LJ, Witteveen AT, Pover RC, et al. (2006) Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics* 7: 278.
- Dankbar DM, Dawson ED, Mehlmann M, Moore CL, Smagala JA, et al. (2007) Diagnostic

Supporting Information

Figure S1 Three common normalization methods. The left plots show pairs of distributions of microarray intensities to be normalized (right plots). A) If the distributions are of the same overall shape, they can simply be scaled to the same mean. B) Quantile normalization imposes the same distribution on all samples. C) A known quantity of RNA is spiked-in to each sample (vertical line) and is then used as a scaling factor.

Found at: doi:10.1371/journal.pcbi.1000543.s001 (1.57 MB EPS)

Text S1 In this section we further discuss some of the issues raised in the main text. Found at: doi:10.1371/journal.pcbi.1000543.s002 (0.23 MB RTF)

Table S1 Comparison of commercial microarray manufacturers. Found at: doi:10.1371/journal.pcbi.1000543.s003 (0.05 MB RTF)

Acknowledgments

We thank the anonymous reviewers for helpful suggestions and comments.

The content is solely the responsibility of the authors and does not necessarily reflect the official views of any of the funding agencies.

- microarray for influenza B viruses. *Anal Chem* 79: 2084–2090.
32. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, et al. (2000) Tissue classification with gene expression profiles. *J Comput Biol* 7: 559–583.
 33. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10: 57–63.
 34. Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470–476.
 35. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18: 1509–1517.
 36. Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, et al. (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS Genet* 5: e1000569.