Analysis and Interpretation of Microarray Data

Michael F. Miles, M.D., Ph.D. and Robnet Kerns, Ph.D.

Department of Pharmacology and Toxicology

Virginia Commonwealth University
Richmond, VA

Table of Contents

1) Introduction and overview of analysis pipeline for microarray experiments.

The advent of the "genomics era" has presented neuroscientists with the ability to have full-genome sequence information available for an ever-expanding number of species. The practical use of this enormous amount of sequence information is clearly evident in the recent development of various DNA microarray technologies that allow full-genome interrogation of gene expression at the level of mRNA abundance. (*1-3*). DNA arrays simply represent small two-dimensional surfaces to which thousands of DNA probes are attached in a defined order. They are most frequently used for massively parallel hybridization-based mRNA expression quantification. There are several types of DNA microarrays, including cDNA or oligonucleotide arrays (*4-8*) that can be synthesized in-house (http://cmgm.stanford.edu/pbrown/mguide/) or acquired commercially. Each type of microarray has it own unique analysis features. A major commercial source of oligonucleotide arrays (GeneChip™, Affymetrix, Inc., Santa Clara, CA) has a design distinct from spotted arrays as described below. Due to increasingly widespread utilization of GeneChip™ oligonucleotide microarrays in neuroscience and familiarity with this type of microarray in our own laboratory, this syllabus will focus on analysis of this microarray platform but many of the issues discussed are widely applicable to multiple microarray platforms.

Microarrays can be applied to problems of functional classification and molecular understanding of disease states (*9*) drug action and discovery (*10*), DNA sequencing (*11*), single nucleotide polymorphism detection (*12*), or expression profiling of experimental perturbations such as treatment with drugs or ligands. In expression profiling or functional genomics, a microarray experiment typically consists of the measurement of levels of gene expression from multiple samples to identify genes differentially expressed due to experimental conditions. Candidate genes for further analysis are selected based on statistical significance of observed expression differences between samples. The precision and sensitivity of microarray data analysis are a critical determinant on whether downstream analysis and interpretation will yield significant results. Because of the inherent complexity of the central nervous system, expression profiling in neurobiology is even more challenging.

By far, the most time consuming issue in performing microarray experiments concerns data analysis. The concept of performing a simple statistical analysis, identifying a few "significant" genes and using these to paint an elegant biological story is fanciful when one is studying 1-3 x 10^4 genes. Table 1 illustrates a typical multi-level analysis of microarray data. It is critical that each step of the process incorporates statistical rigor as much as possible. Ensuring that high-quality results are derived from low-level analysis is especially important. Obviously, doing complex, time-consuming bioinformatics analyses on "garbage" will only result in "garbage". However, the advent of genomic scale data, such as that generated by microarrays, has fueled the rapid expansion of interest in bioinformatics tools. This is illustrated by the largely parallel increases in publications containing the terms "microarray" or "bioninformatics" as shown in Figure 1.

This syllabus will focus on low-level analysis issues and application of bioinformatics tools for the functional interpretation of microarray data. Dr. Chesler's presentation will cover issues such as experimental design, statistical comparison of arrays, and multivariate analysis.

2) Experimental design

Dr. Chesler will give a more extensive discussion of statistical concerns in the experimental design of microarray experiments. In addition, numerous recent references give a thorough discussion of various aspects of microarray experimental design (*13*). Terry Speed's laboratory also maintains an excellent website with links to presentations, technical reports and papers regarding design and analysis of both spotted cDNA and Affymetrix microarrays (http://stat-www.berkeley.edu/users/terry/zarray/html/).

However, several points cannot be over-emphasized. In general, many early microarray studies suffered from lack of sufficient (or any) replicates. Three independent experimental replicates is usually a minimum (*14*) but gene-specific variation in error occurs. We find that running the same sample on duplicate microarrays is usually unnecessary due to the very high intra-sample replicability with Affymetrix microarrays (generally, $r>0.98$). However, with spotted cDNA microarrays, performing replicate slides with dye-reversal is essential due to systematic bias induced by the dye-labeling. A far greater statistical yield is obtained by performing repeated experiments with single microarrays used for each sample within an experiment.

Performing any quantitative analysis requires assessing and minimizing variance. Microarrays present a huge problem in this regard because the very large number of factors that can induce variance when measuring the expression of over 10,000 genes. As summarized in Table 2, multiple sources of possible variability must be taken into account during experimental design. One of the largest sources of variance for microarray data in neuroscience research can concern the use of animal tissue. In general, we prefer a design where tissue from multiple animals (~4) is pooled for each experimental sample. Replicate experiments are then performed. This increases the number of animals used, but greatly reduces many of the sources of variance listed in Table 2. In particular, when brain regions dissections are performed, pooling of tissue from multiple animals is highly recommended. Depending on the experiment, brain microdissections (*15*) or laser-capture microscopy are highly preferable to "whole brain" studies since the latter can totally obscure even large region-specific changes in the expression of a given gene (*16*).

The goal of microarray studies is to identify a gene or pattern of gene expression that is associated with the trait under analysis. As with many biological studies, proving causality is difficult. Microarray studies benefit from the detection of correlated expression changes in many genes, often those subserving the same biological function. Multivariate methods for detecting such patterns will be discussed by Dr. Chesler. However, a general principle of experimental design can greatly increase the yield of multivariate studies in microarray experiments. As Hughes et al. elegantly showed with studies in yeast, a compendium of expression results across multiple pharmacological treatments and genetic alterations can provide a rich source of functional associations between expression patterns (*10*). Thus, incorporating multiple genetic models (null mutations, transgenic over-expression, inbred lines, etc.) and pharmacological interventions all affecting a trait of interest is a robust design feature for microarray experiments. We term this combined approach "molecular triangulation".

3) Microarray fundamentals

The design and utilization of spotted cDNA arrays has been discussed in detail in numerous review articles (*17-19*). GeneChip™ oligonucleotide microarrays (hereafter referred to as "oligonucleotide arrays") are constructed in a considerably different fashion with each gene

represented by multiple (12-20) probe pairs (*7*) (Figure 2). A probe pair consists of a perfect match probe (PM) and a mismatch probe (MM). A probe is a 25-mer oligonucleotide, and in the case of a PM, its sequence is complementary to a segment of the target gene. A MM is identical to the corresponding PM except that the middle base of the MM is altered to no longer complement the target sequence. Thus, the MM was designed to assess the extent of non-specific hybridization. The probes are generally biased toward the 3'-end of the mRNA molecules and often there is considerable overlap between individual probes. All the probes on these arrays are synthesized in parallel by photochemistry (*20, 21*). Preparation of target molecules for hybridization to these oligonucleotide microarrays is done using a linear amplification process to increase the abundance of material and incorporate biotin molecules for later labeling with fluorescent dyes (see Figure 3). Following hybridization, washing and scanning of oligonucleotide microarrays, a scanning confocal microscope detects fluorescence from bound cRNA target molecules. Typically, oligonucleotide microarrays are scanned with a 3 μm pixel size, generating a .DAT file of >40 Mb. Affymetrix software (MAS, see below) is used to define position of oligonucleotides and calculate signal intensities of individual oligonucleotides. This converts the .DAT file to a .CEL file that still contains probe-level signal information but has now averaged the individual pixels for a given probe.

Although usually done following low-level analysis (see below), assessment of quality control is a crucial aspect of microarray experiments. Many approaches have been proposed for ensuring adequate quality control in microarray experiments (*22-25*). Using Affymetrix oligonucleotide microarrays, issues such as scaling factor, background noise, % of genes called "present", and the ratio of 3'- versus 5'-end probes are all used to assess performance and reproducibility of microarrays (see discussion in (*15*)). Although different laboratories may set different standards for such quality control measures, and different microarray types will vary in their performance, a general rule is that these measures should be highly consistent across all chips for a given experiment. A more deliberate assessment of probeset and chip performance with Affymetrix oligonucleotide arrays can be performed using the d-chip software (http://biosun1.harvard.edu/complab/dchip/) utilized by the MBEI analysis algorithm discussed below. This software will flag outliers using statistical assessment of variance from modeled performance.

4) Low level analysis issues

Following signal acquisition and calculation of individual oligonucleotide intensities, the first step in microarray analysis, is normalization of signal intensities (*26, 27*). Although normalizing across all probes on an array (whole chip) is often adequate, there can be significant problems with non-linearity, particularly with high-abundance genes (Figure 4). Figure 4 utilizes a useful graphical display (M vs. A plot) for checking reproducibility across microarrays. Non-linear behavior is readily detected with this display. Affymetrix has addressed the issue of non-linearity with high abundance genes, in part, by reducing the maximum photomultiplier tube output. However, statistical approaches such as employing iterative linear regression or quantile methods to providing more robust normalization have also been applied (*27*) and are readily available through the R statistical software consortium (http://cran.r-project.org; http://www.bioconductor.org). In addition, some approaches use a set of "invariant" genes, determined over a large series of microarrays, to normalize subsequent arrays (*28*).

In older probe designs, and with spotted cDNA arrays, there can be significant position-dependent differences in normalization due to inconsistencies in microarray manufacturing. This required additional position-dependent or pin-dependent normalization. Current designs of oligonucleotide microarrays utilizes a random dispersal of probe pairs for a given gene, thus reducing position dependent normalization issues.

Once normalized, individual oligonucleotide probe pairs are usually "reduced" to a single number representing the expression level for the given gene. Multiple algorithms for deriving this expression intensity have been developed (see Table 1). The major problem being twofold: 1) the MM probes designed to compensate for "non-specific" hybridization clearly also contain "specific" hybridization signal and not infrequently, actually hybridize stronger than the cognate PM probe; 2) there is very large variation in the hybridization performance of individual PM probes (see Fig. 2), thus making it clear that they cannot be treated as repeated equivalent measurements. Affymetrix originally devised a "trimmed mean" method for determining an "average difference" value (MAS 4) but this was prone to large fluctuations with lower abundance genes, even producing "negative" values (see large dispersal at low abundance range in Fig. 4) (*29*). A more recent version, MAS 5.0, uses a statistical expression algorithm to calculate the signal on the oligonucleotide array (*30*). In this case, MM values are replaced by a modeled "IM" value if MM>PM and a Tukey biweight method is used to calculate the mean of the PM-MM (IM) values. This process eliminates the occurrence of negative expression levels and produces a mean value less sensitive to outliers.

Another probe-based method, model based expression index (MBEI), was developed by Li and Wong (*31*) for oligonucleotide array analysis. In summary, MBEI is the weighted average of probe pair signals within a probe set; the weights are calculated from the pattern of probe pair signals observed from multiple samples. However, the MBEI modeling of probe pair signals requires large numbers of microarrays for optimal performance. Irizarry *et al*. recently performed a detailed comparison of the MAS 5, MAS 4 and MBEI methods of summarizing probe pair data for oligonucleotide arrays (*32*). They demonstrated that a new method, referred to as the log scale robust multi-array analysis (RMA) appeared to outperform both MAS 5 and MBEI in terms of providing reliable estimates of expression levels for given genes.

Recently, Dr. Li Zhang (M.D. Anderson Cancer Center) and our laboratory have developed an entirely different approach to modeling probe performance on Affymetrix oligonucleotide arrays (*33*). This method, termed "position-dependent nearest neighbor" (PDNN) actually predicts the disparate hybridization intensities of individual probes by taking into account the contribution of inter-base stacking energies in the probes themselves as affecting the stability of the probe-target hybridization. Thus, individual probes can be "corrected" and the actual abundance of the target cRNA be more accurately calculated.

In most cases, investigators are not interested in the absolute expression level of a genes but rather, whether expression changes between two conditions (chips). Fold-changes can be calculated as logarithm ratios of expression intensities for a given gene to compare baseline and experiment arrays, similar to the method used to analysis cDNA arrays (*34*). The logarithm ratio is not, however, appropriate for representing the significance of change when expression levels are close to background noise, thus reducing its usefulness. MAS 5.0 will also make a comparison analysis between experiment and baseline arrays based upon a non-parametric statistical analysis of probe pair data (PM-MM) on two arrays. This analysis produces a "Change p-value" and a "Change Call" (increase, marginal increase, no change, marginal decrease, decrease).

As an effort to improve the reliability of comparisons between microarrays, our laboratory developed the S-score algorithm (*35*) for comparison of oligonucleotide microarrays. The S-score method is based upon a simple error model that is used to estimate the variances for probe pair signals. This model includes both additive and multiplicative (intensity dependent) error terms that more accurately depict the behavior of microarray data. Similar approaches have been used for spotted cDNA microarrays (*10*). The method also defines a relative change of probe pair signals, which effectively converts probe pair signals into multiple measurements with equalized errors. The relative changes in individual probe pairs are then combined to determine a single variant measure of the significance of change for a targeted gene. The S-score algorithm has been applied to oligonucleotide microarray data from brain tissue or neural cells (*35-37*) and software is available at http://www.brainchip.vcu.edu/mm_analysis.html.

5) Defining functional relationships between expression profiles (genes)

Obviously, determining differences in expression across experimental samples is the goal of microarray studies. The above discussion concerned the generation of high quality primary microarray data. Dr. Chesler will discuss use of various statistical analyses, including multivariate studies, to define "gene lists" or "clusters" of coordinately regulated genes. Excellent overviews of statistical approaches to microarray data is provided in several recent reviews (*38-41*).

Despite increasingly elegant statistical approaches to analyzing microarray data, defining the "meaning" of microarray results remains a conundrum. Placing the entire "expressome" on a chip does not imply that we understand the multiple functions of each gene and the complex interactions between them. Thus, the functional analysis of microarray patterns is likely a science in its infancy. Most efforts to date have been directed at providing increasingly high-throughput, up-to-date and detailed annotations of genes on a list derived from statistical filtering or cluster analysis. Web sites such as the Stanford SOURCE database (http://genome-www5.stanford.edu/cgi-bin/SMD/source/sourceSearch) provide easy links to multiple other databases including compilation of SwissProt, UniGene, LocusLink, and Gene Ontology information about a given gene. SOURCE also provides links to existing microarray cluster data which can help establish the tissue distribution of a transcript and perhaps confirm cluster "neighbors" of an investigator's own results. Other useful compilations of annotation data are listed in Table 3.

A potentially productive, and highly active, bioinformatics approach for analysis of microarray data concerns superimposing existing gene-gene functional relationships upon microarray data. Thus, a cluster or list of genes with statistically significant expression changes can be interrogated for the occurrence of various functional annotations. These annotations might include biochemical pathways (KEGG), gene ontology classifications, protein-protein interaction databases, or defined signaling pathways. By applying various statistical approaches, generally including some type of permutation analysis, an investigator can determine whether the gene lists from microarray experiments has an over-representation of certain functional categories. Mirnics and colleagues, for example, used such an approach to identify functional patterns amongst gene expression changes seen in brain tissue of schizophrenics (*42*).

A variety of web-based or available software downloads are now available for performing functional group analysis on microarray data. Table 3 lists several of these resources. Perhaps two of the most popular include the GenMAPP/MAPPFinder and DAVID/Ease tools. Both of

these will assess the relative frequency of genes from a list in various functional groups such as GO categories. The Ease software provides a more detailed statistical analysis of the possible significance of such functional groupings. However, the GenMAPP program has capabilities for incorporated detailed biochemical pathway data, complete with links to primary microarray data or background information regarding a given gene. Both of these approaches allow a fairly high throughput analysis of possible functional overlays present within a complex set of microarray data.

The drawback to these approaches comes from the rather unsophisticated nature of the annotation data currently available for individual genes and the meager state of database information regarding known functional interactions between genes. Thus, finding that a GO category of "nuclear proteins" is over-represented in a particular group of microarray results might not be extremely helpful. However, as annotation or interaction databases improve in complexity, their merging with microarray results will become increasingly productive. An example of a more complex effort to merge large databases containing protein-protein or protein-DNA interaction data and microarray data can be seen with the Cytoscape software algorithm (http://www.cytoscape.org) (*43*).

Another approach to superimposing biological relationship data upon microarray results comes with efforts to find associations between genes in the biomedical literature. Thus, genes in a given cluster might be interrogated for having appeared together in the same article in the biomedical literature. The PubGene application (http://www.pubgene.org/) performs such a task and can also superimpose microarray data upon associations from the literature (see Fig. 5). Thus, if a group of genes respond similarly on microarray studies and have repeated associations in the biomedical literature, their grouping will ranked higher in the PubGene output.

Motif mapping represents yet another effort at identifying biological relationships between genes within expression profiles. Genes showing highly correlated expression profiles across multiple experimental conditions could be expected to have some common promoter motifs underlying their coordinate regulation. Numerous examples of exactly such relationships have now been identified (*10, 44*). Databases and search algorithms for identifying over-represented known promoter motifs or novel conserved motifs are now becoming increasingly available. The MEME/MAST system (http://meme.sdsc.edu/meme/website/intro.html) and TRANSFAC database (http://transfac.gbf.de/TRANSFAC/) are publicly available examples of such resources. Unfortunately, the redundancy of eukaryotic promoter motifs makes such algorithms often produce large numbers of "conserved" motifs.

Finally, as mentioned under "experimental design", combining microarray studies with genetic and pharmacological interventions allows increased selectivity in identifying functionally relevant expression profiles. An elegant demonstration of this principle is the recent application of expression profiling as a "trait" to be integrated with genetic and phenotypic studies across panels of genetically diverse organisms (*45*). The WebQTL site (http://webqtl.org/) provides an extremely powerful first look at the utility of integrating genetics and genomics. Drs. Williams and Chesler will discuss this exciting approach in detail.

6) Conclusions

Functional genomics (and proteomics) studies have already produced data sets of enormous size and complexity in less than a decade. The organization and analysis of this data is an

ongoing process that is starting to produce increasingly sophisticated results. It seems clear that the optimal approach for deriving complex biological information from such complex quantitative data resides in combining careful experimental design, rigorous technical standards and statistical analysis, and novel bioinformatics tools. The latter must be capable of identifying new relationships by merging large datasets of diverse information. As such tools develop, and the complexity of biological databases expands, then the promise of using genomic or proteomic approaches to advance our understanding of complex biology may be realized.

References

1. D. Gerhold, T. Rushmore, C. T. Caskey, *Trends Biochem Sci* **24**, 168-73 (May, 1999).
2. D. J. Lockhart, E. A. Winzeler, *Nature* **405**, 827-36 (Jun 15, 2000).
3. R. A. Young, *Cell* **102**, 9-15 (Jul 7, 2000).
4. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467-70 (Oct 20, 1995).
5. M. Schena *et al.*, *Proc Natl Acad Sci U S A* **93**, 10614-9 (Oct 1, 1996).
6. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res* **6**, 639-45 (Jul, 1996).
7. D. J. Lockhart *et al.*, *Nat Biotechnol* **14**, 1675-80 (Dec, 1996).
8. J. DeRisi *et al.*, *Nat Genet* **14**, 457-60 (Dec, 1996).
9. T. R. Golub *et al.*, *Science* **286**, 531-7 (1999).
10. T. R. Hughes *et al.*, *Cell* **102**, 109-26. (2000).
11. A. C. Pease *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **91**, 5022-6 (1994).
12. J. G. Hacia *et al.*, *Nat Genet* **22**, 164-7 (Jun, 1999).
13. Y. H. Yang, T. Speed, *Nat Rev Genet* **3**, 579-88 (Aug, 2002).
14. M. L. Lee, F. C. Kuo, G. A. Whitmore, J. Sklar, *Proc Natl Acad Sci U S A* **97**, 9834-9 (2000).
15. R. Sandberg *et al.*, *Proc Natl Acad Sci U S A* **97**, 11038-43 (2000).
16. E. Wurmbach *et al.*, *Neurochem Res* **27**, 1027-33 (Oct, 2002).
17. M. Schena, *DNA microarrays : a practical approach*, The practical approach series ; 205. (Oxford University Press, Oxford ; New York, 2000).
18. D. Shalon, S. J. Smith, P. O. Brown, *Genome Research* **6**, 639-45 (1996).
19. G. Ramsay, *Nature Biotechnology* **16**, 40-4 (1998).
20. S. P. Fodor *et al.*, *Science* **251**, 767-73 (1991).
21. R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, D. J. Lockhart, *Nature Genetics* **21**, 20-4 (1999).
22. G. C. Tseng, M. K. Oh, L. Rohlin, J. C. Liao, W. H. Wong, *Nucleic Acids Res* **29**, 2549-57 (Jun 15, 2001).
23. S. Draghici, A. Kuklin, B. Hoff, S. Shams, *Curr Opin Drug Discov Devel* **4**, 332-7 (May, 2001).
24. D. Finkelstei *et al.*, *Plant Mol Biol* **48**, 119-31 (Jan, 2002).
25. W. Raffelsberger, D. Dembele, M. G. Neubauer, M. M. Gottardis, H. Gronemeyer, *Genomics* **80**, 385-94 (Oct, 2002).
26. Y. H. Yang *et al.*, *Nucleic Acids Res* **30**, e15 (Feb 15, 2002).
27. B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, *Bioinformatics* **19**, 185-93 (Jan 22, 2003).

28. C. Li, W. H. Wong, *Proc Natl Acad Sci U S A* **98**, 31-6. (2001).
29. Affymetrix, *Affymetrix, http://www.affymetrix.com/support/technical/manuals.affx* (1999).
30. Affymetrix, *Affymetrix, http://www.affymetrix.com/support/technical/manuals.affx* (2001).
31. C. Li, W. H. Wong, *Proc Natl Acad Sci U S A* **98**, 31-6 (Jan 2, 2001).
32. R. A. Irizarry *et al.*, *Nucl. Acids. Res.* **31**, e15 (2003).
33. L. Zhang, M. F. Miles, K. D. Aldape, *Nat Biotechnol* **21**, 818-21 (Jul, 2003).
34. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc Natl Acad Sci U S A* **95**, 14863-8 (Dec 8, 1998).
35. L. Zhang, L. Wang, A. Ravindranathan, M. F. Miles, *J Mol Biol* **317**, 225-235 (2002).
36. R. C. Elliott, M. F. Miles, D. H. Lowenstein, *J Neurosci* **23**, 2218-27 (Mar 15, 2003).
37. S. Hassan, B. Duong, K. S. Kim, M. F. Miles, *J Biol Chem* (Jul 3, 2003).
38. S. Dudoit, Y. H. Yang, M. J. Callow, T. J. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments" *Tech. Report No. 578* (Stanford University School of Medicine, 2000).
39. J. Quackenbush, *Nat Rev Genet* **2**, 418-427 (2001).
40. G. Sherlock, *Curr Opin Immunol* **12**, 201-5. (2000).
41. J. Wittes, H. P. Friedman, *Journal of the National Cancer Institute* **91**, 400-1 (1999).
42. K. Mirnics, F. A. Middleton, A. Marquez, D. A. Lewis, P. Levitt, *Neuron* **28**, 53-67. (2000).
43. T. Ideker, O. Ozier, B. Schwikowski, A. F. Siegel, *Bioinformatics* **18 Suppl 1**, S233-40 (Jul, 2002).
44. J. L. DeRisi, V. R. Iyer, P. O. Brown, *Science* **278**, 680-686 (1997).
45. E. E. Schadt *et al.*, *Nature* **422**, 297-302 (Mar 20, 2003).
46. S. Dudoit, J. Shaffer, J. Boldrick, *U.C. Berkeley Division of Biostatistics Working Paper Series; http://www.bepress.com/ucbbiostat/paper110* (2002).
47. V. G. Tusher, R. Tibshirani, G. Chu, *Proc Natl Acad Sci U S A* **98**, 5116-21 (Apr 24, 2001).
48. L. Zhang, L. Wang, A. Ravindranathan, M. F. Miles, *J Mol Biol* **317**, 225-35 (Mar 22, 2002).
49. J. Quackenbush, *Nature Reviews Genetics* **2**, 418-427 (June, 2001).
50. M. Diehn *et al.*, *Nucleic Acids Res* **31**, 219-23 (Jan 1, 2003).
51. T. K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, *Nat Genet* **28**, 21-28 (2001).
52. K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, B. R. Conklin, *Nat Genet* **31**, 19-20 (May, 2002).
53. S. W. Doniger *et al.*, *Genome Biol* **4**, R7 (2003).

Figure Legends

Figure 1: Parallel growth in publications concerning microarray data and bioinformatics. The number of publications listed in PubMed for the indicated years was determined using searches for the words "microarray" or "bioinformatics".

Figure 2: Design and performance of oligonucleotide "probeset". The schematic design of a typical Affymetrix oligonucleotide probeset is shown. In current designs, the different probe pairs (PM and MM) are distributed randomly across the chip. The lower diagram shows a plot of the PM-MM values for each probe pair and the approximate calculation of a trimmed mean as done with the MAS 4 analysis software.

Figure 3: Synthesis of cRNA and hybridization of oligonucleotide microarrays. The various biochemical reactions for preparation of biotinylated cRNA from starting total RNA is diagrammed.

Figure 4: M vs. A plot of replicate hybridizations analyzed with MAS 4. Following hybridization of two biological replicate samples to two mouse oligonucleotide microarrays (U74Av2), data was analyzed with MAS 4 analysis software (Affymetrix) and expression values for two chips compared. Data was filtered to eliminate absent genes and "negative" average-difference values. Ratio of expression levels for individual genes (y-axis) is plotted versus a measure of abundance of the transcripts (x-axis). Figure depicts mild non-linear behavior at high-abundance classes (circled) and increased scatter of data for low abundance transcripts.

Figure 5: Relationships between genes in the biomedical literature. The gene for serum and glucocorticoid regulated kinase (SGK) was use to probe the PubGene (http://www.pubgene.org) analysis program for possible relationships with other genes in the biomedical literature.

Table 1: Summary of Analysis Stages and Methods for Oligonucleotide Microarrays.

Analysis Stage	Description	Examples of Methods
Normalization	Equalizes overall signal across arrays to be compared, ensures linearity of response across abundance classes	Whole chip(*26*) Quantile(*27*)
Probe reduction	Combines signals from multiple probes or probe pairs to define "expression level". Identifies genes with invalid or hyper-variable expression levels.	Weighted average (MAS 4)(*29*) Tukey bi-weight (MAS 5)(*30*) Model-based (MBEI)(*31*) Log scale linear additive (RMA)(*32*) Position-dependent stacking energy modeling (PDNN) (*33*)
Comparative	Compares expression of a gene across two or more arrays to determine significant changes in expression	t-test rank order (MAS 5) (*30*) permutation (SAM) (*46, 47*) S-score (*48*)
Multivariate studies	Identifies significant correlations in expression data across experiments/conditions	hierarchical clustering k-means clustering self-organizing maps principle components analysis & many more(*34, 49*)
Biological overlay	Identify functions for given genes, clusters of genes; hypothesis generation	Multiple database access (Source)(*50*) PubMed correlations (PubGene)(*51*) Gene Ontology rankings (GenMAPP, MAPPFinder, DAVID/EASE)(*52, 53*)

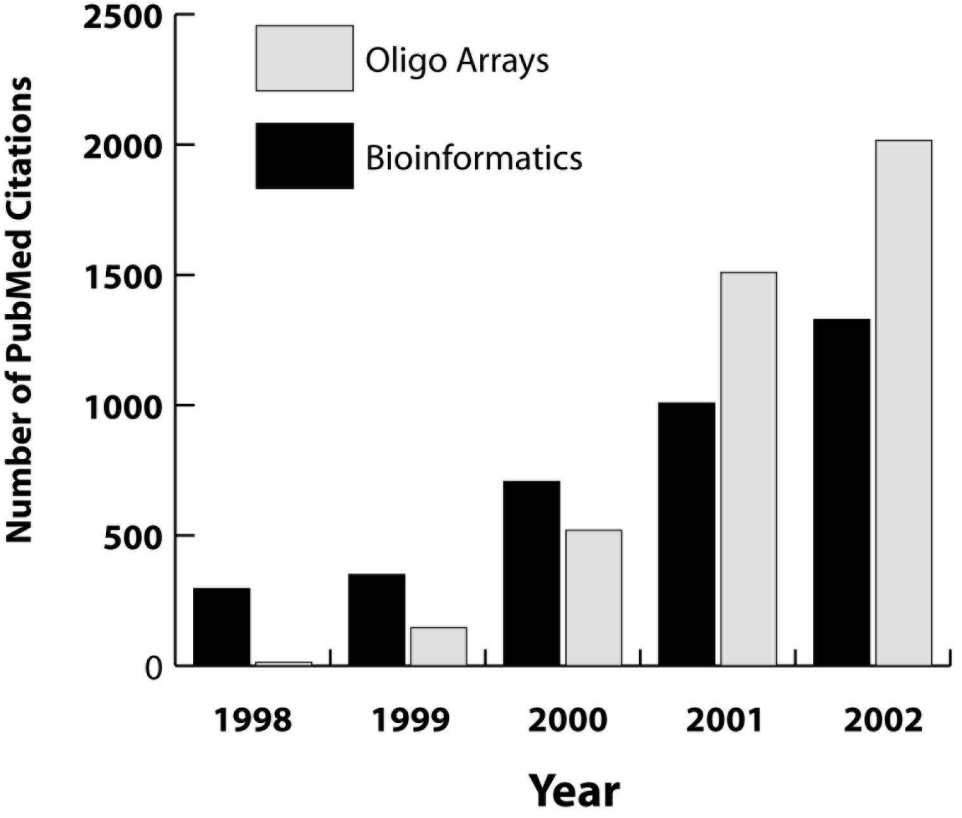
Table 2: Sources of Variance in Microarray Experiments.

Type of Variance	Factors
Biological	Animal-animal differences (intra/inter cage, supplier)
	Genotype
	Circadian rhythms
	Stress
Technical	Sample treatment/harvesting (dissections, injections)
	Target preparation (enzyme lots, mRNA quality)
	Lot-to-lot chip variation
	Chip processing (scanning order)
Environmental	Temperature
	Handling
	Noise/odors

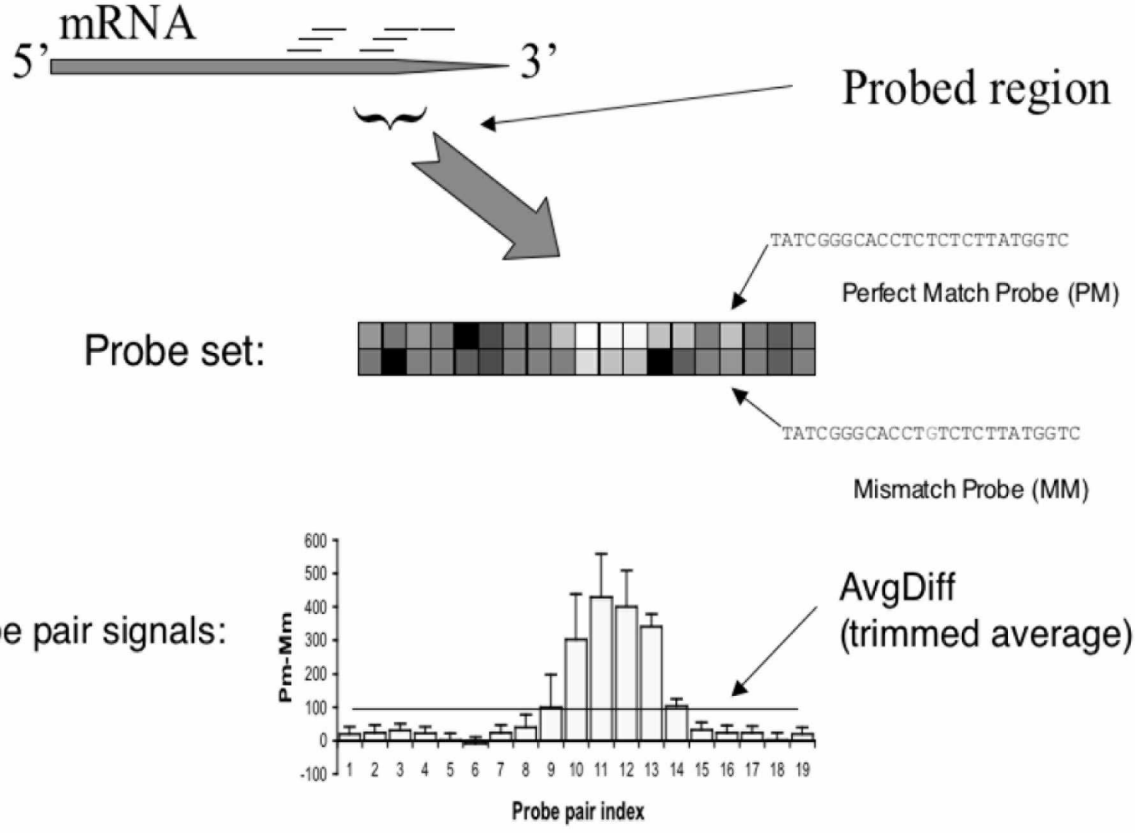
Table 3. Examples of Bioinformatics Resources for Microarray Experiments

Name	Description	Link
SOURCE	Human, rat, mouse gene compilation from multiple databases; allows batch submissions for annotation	http://source.stanford.edu/cgi-bin/sourceSearch
GeneLynx	Human, mouse gene compilation; multiple database links regarding gene/protein structure and function	http://www.genelynx.org/
DAVID/Ease	Mines gene list for frequency of GO categories; annotation of gene list; statistical analysis of biological themes in gene list (EASE)	http://apps1.niaid.nih.gov/David/upload.asp
GenMAPP/MAPPFinder	Superimposes array data on biological pathways; statistical ranking of functional groups	http://www.genmapp.org/
FatiGO	Mines gene list for occurrence of GO terms; statistical comparison of two lists for over-representation	http://fatigo.bioinfo.cnio.es/
PubGene	Finds associations between genes in biomedical literature; superimposes array data on literature links; commercial version available	http://www.pubgene.org/
MEME	Search promoter regions of genes in list/cluster for conserved motifs	http://meme.sdsc.edu/meme/website/intro.html

Miles Fig 1

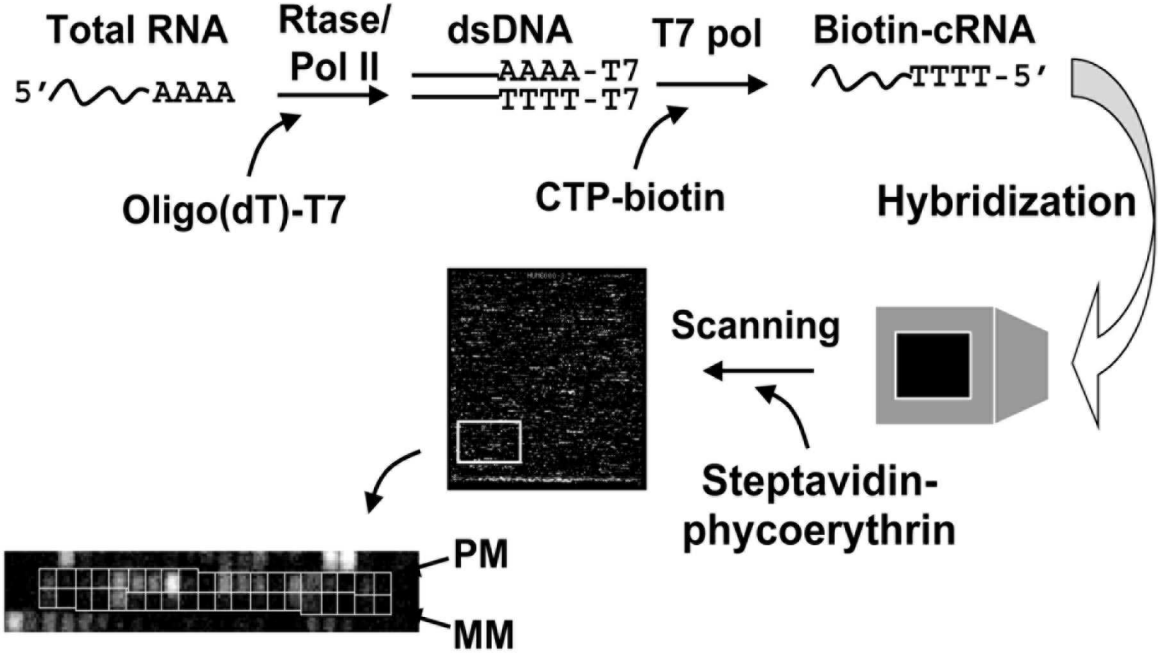

Number of PubMed Citations
2500
2000
1500
1000
500
0
Oligo Arrays
Bioinformatics
1998
1999
2000
2001
2002
Year

Miles Fig 2

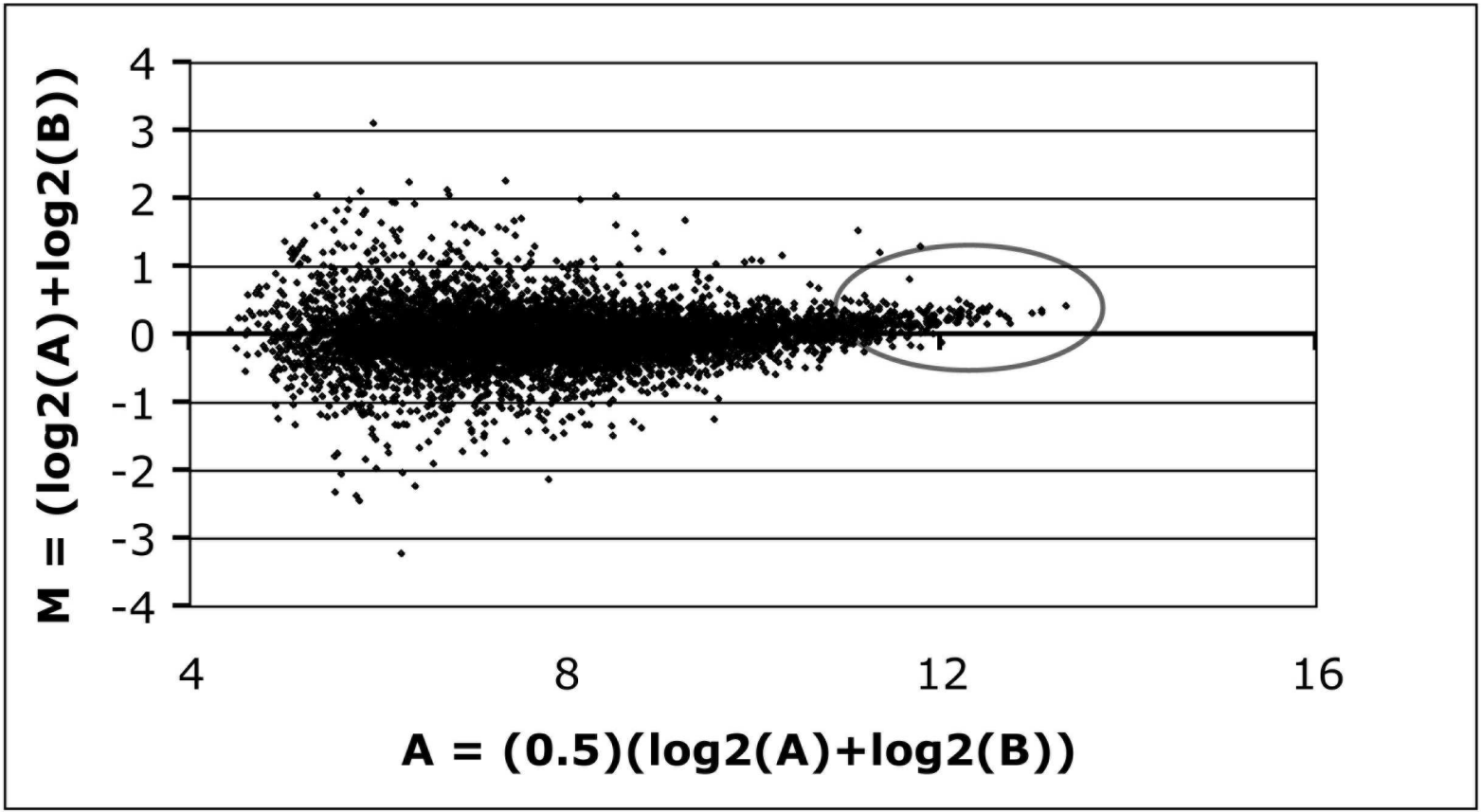

5'
mRNA
3'
Probed region
TATCGGGCACCTCTCTCTTATGGTC
Perfect Match Probe (PM)
Probe set:
TATCGGGCACCTGTCTCTTATGGTC
Mismatch Probe (MM)
Probe pair signals:
Pm-Mm
600
500
400
300
200
100
0
-100
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Probe pair index
AvgDiff
(trimmed average)

Miles Fig. 3

Oligonucleotide Array Analysis



Miles Fig. 4

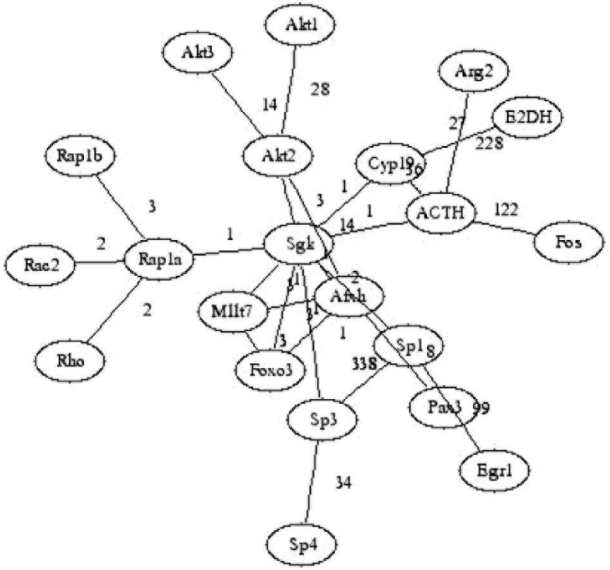


Sgk

Found in 14 articles with 9 neighbours

Visit Neighbour

- Akt2 3 times
- Sp1 2 times
- Rap1a 1 times
- Afxh 1 times
- Mllt7 1 times
- ACTH 1 times
- Sp3 1 times
- Foxo3 1 times
- Cyp19 1 times



Subset tool menu

Subset Network | Submit

- ☑ Sgk
- ☑ Akt2
- ☑ Mllt7
- ☑ Cyp19
- ☑ ACTH
- ☑ Sp1
- ☑ Sp3
- ☑ Afxh
- ☑ Foxo3
- ☑ Rap1a
- ☑ Akt1
- ☑ Akt3
- ☑ E2DH
- ☑ Fos
- ☑ Arg2
- ☑ Egr1
- ☑ Sp4
- ☑ Pax3
- ☑ Rac2
- ☑ Rho
- ☑ Rap1b