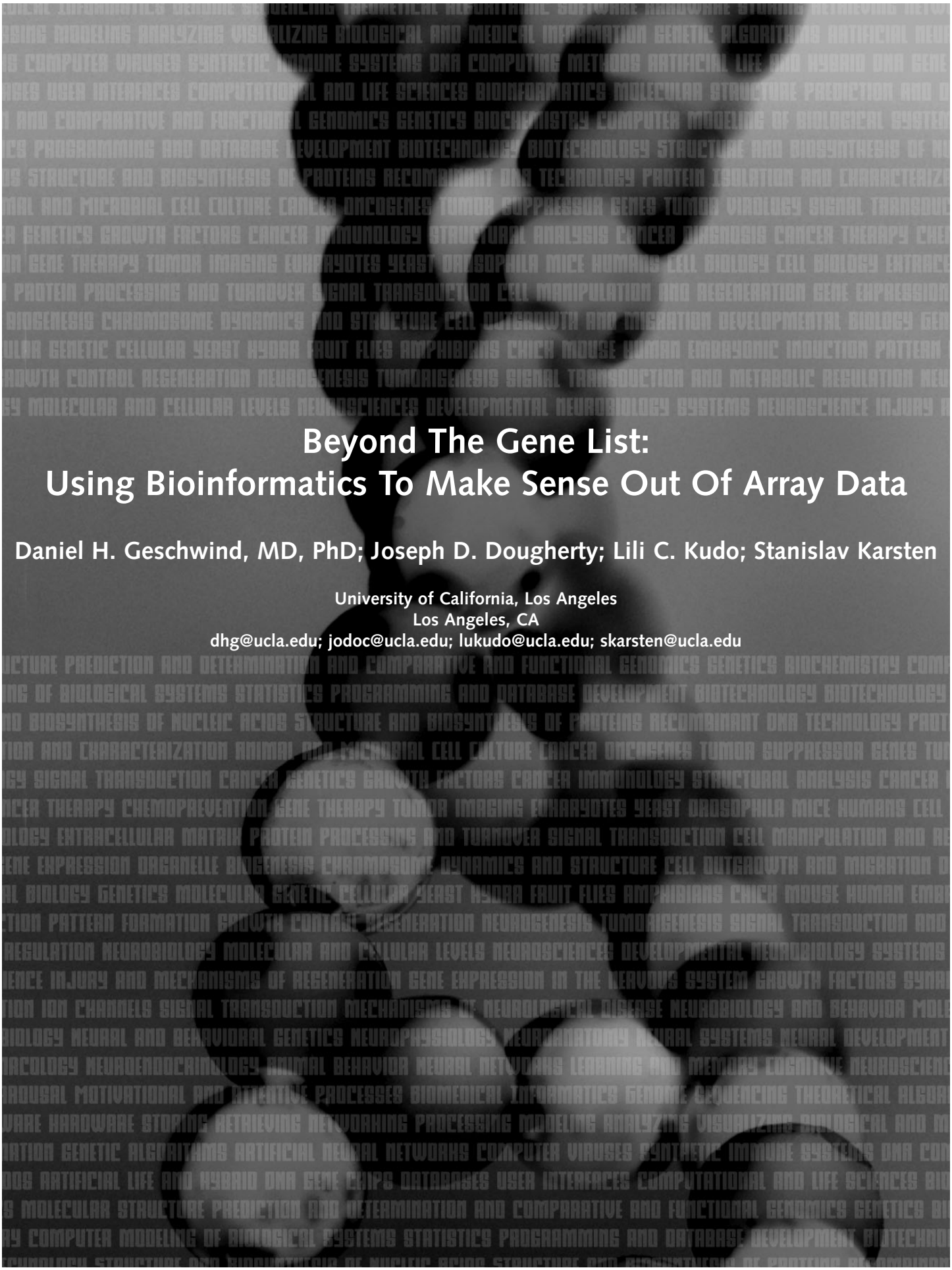


Beyond The Gene List: Using Bioinformatics To Make Sense Out Of Array Data

Daniel H. Geschwind, MD, PhD; Joseph D. Dougherty; Lili C. Kudo; Stanislav Karsten

University of California, Los Angeles
Los Angeles, CA
dhg@ucla.edu; jodoc@ucla.edu; lukudo@ucla.edu; skarsten@ucla.edu

Table of Contents

1. Introduction

Nervous system development and function in part depends on the careful orchestration of gene expression in the CNS. The magnitude of the problem of understanding this system from a molecular standpoint is underscored by the estimate that more than half of all genes are expressed in the nervous system and many of them are relatively specific to the brain. Currently, there are many different high throughput methods to examine gene expression in different cells, tissues or even pathological specimens under various conditions, including serial analysis of gene expression (SAGE)(Velculescu, Zhang et al. 1995) large-scale cDNA sequencing (Okubo, Hori et al. 1992), and DNA microarrays (Schena, Shalon et al. 1995). All of these approaches rely heavily on use of informatics for interpretation. In addition, there are now pure informatic approaches available now, which include expressed sequence tag (EST) database comparison (O'Dowd, Nguyen et al. 1998), and mining of SAGE or other databases. This syllabus focuses on DNA microarrays.

Oligonucleotide and cDNA arrays enable the monitoring of thousands of known and unknown genes in parallel in many samples in an efficient manner (Schena, Shalon et al. 1995; Lockhart, Dong et al. 1996; DeRisi, Iyer et al. 1997; Geschwind 2002). Due to the high throughput nature of this approach, large amounts of data are generated in a single experiment. Formerly, cost and availability of microarrays were limitations to performing experiments, but as these issues are resolved, the new limitation becomes interpretation and analysis. The ability to study gene expression on a genomic level using microarrays raises many important issues, most notably data sharing (Becker 2001; Geschwind 2001; Miles 2001; Mirnics 2001) and developing array bioinformatic tools for data mining.

Regarding data sharing, major steps have been taken towards the development and adoption of a uniform standard for array data management and storage. However, one of the factors limiting easy interaction between genome and array data is the lack of uniform database standards in use by the various protein and genome databases, and the standards continue to change, making easy connectivity between these resources and array data problematic even for the professional bioinformatician (Stein 2002).

Array bioinformatics is a potentially enormous field; The bioinformatics issues related to microarray data could easily be the subject of a week of seminars. This chapter can, by its nature, deal only briefly with a few of the interesting bioinformatic approaches that one can take with microarray data. We take a biological rather than a mathematical perspective. Therefore, the focus is not on complex analysis or the algorithms per se, but rather to give some examples of how to use databases or bioinformatic approaches to enrich your experimental design and interpretation so as to make biological sense out of the data.

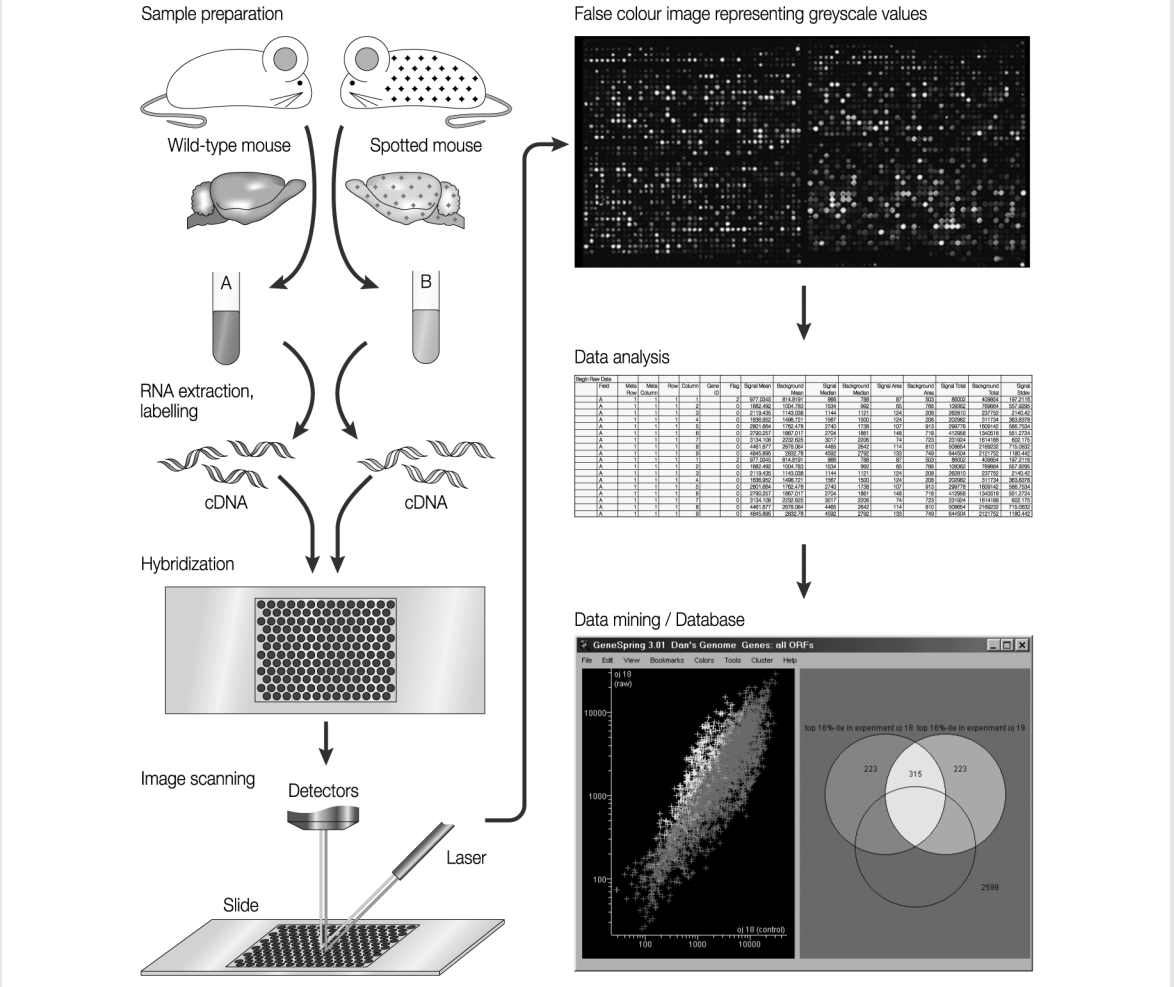
2. What Is A Microarray?

DNA microarrays, also known as "DNA chips", are ordered arrays of DNA grided andattached onto a rigid and non-porous surface such as glass or silicon. Arrays composed ofoligonucleotides synthesized by photolithography *in situ* are also available commercially andform the basis of GeneChip technology sold by Affymetrix. cDNA arrays ,originally developedby Pat Brown and colleagues at Stanford (Schena, Shalon et al. 1995), are another common formof microarrays and are usually comprised of PCR-amplified inserts from cDNA clonesrepresenting known genes and expressed sequence tags (ESTs) (Cirelli and Tononi 1999) (Wang,Gan et al. 1999). In addition, oligonucleotide arrays containing long oligonucleotides synthesized *in situ* using ink jet technology form the basis of the Agilent platform, and similar lengtholigonucleotides (50 to 70-mer) are being printed on glass slides in a manner similar to cDNAarrays by many laboratories. While the Agilent arrays clearly perform very reliably, whether thein-house custom printed oligonucleotide arrays perform as well as cDNA arrays remains to bedemonstrated. The relative technical merits of cDNA versus oligonucleotide arrays can be foundin several comprehensive reviews (Lockhart and Winzeler 2000) (Brown and Botstein 1999)(DeRisi and Iyer 1999) (Various 1999).

The basic steps of a microarray experiment are depicted in Figure 1 (from NatureReviews Neuroscience). Several critical issues that neuroscientists face are not as salient in otherdisciplines using microarrays (Geschwind, and Gregg 2002; Luo and Geschwind 2001). Forexample, because there are so many cell types and sub-types in many CNS tissues, one needs toconsider the need to enrich for the region or cells of interest so as to detect more subtledifferences in the expression of low-abundance genes. Microarrays have a detection limit andthus changes in low abundance genes in complex tissues may be missed (Geschwind 2000). Thisis illustrated graphically in Figure 2 (courtesy of Karoly Mirnics MD PhD) and must beconsidered during one?s interpretation of the data: In cases where detection of such rare speciesis necessary, microdissection or single cell capture methods can be used. Several signal or RNAamplification methods are commonly used in microarray experiments and have proven trackrecords; these methods may also increase detection of low abundance transcripts (Eberwine, Yehet al. 1992; Karsten, Van Deerlin et al. 2002; Karsten 2002; Luo and Geschwind 2001; Ginsberg2001).

FIGURE 1

Experimental steps in a typical cDNA microarray experiment. 1)Total RNA is extracted from the two comparison specimens (spotted and non-spotted mice) and cDNA are separately reversed transcribed and labeled with different fluors. 2)The two samples are co-hybridized onto the array, the dyes are excited by a laser and the scattered light is collected and analyzed by a scanner. (Another common experimental design includes the use of a reference sample labeled with one dye to which all experimental and control samples are compared by cohybridization. This method is advantageous for statistical analysis, but care has to be taken that the reference sample remains uniform.) 3)The images produced by each dye are registered and a false color overlay is produced. 4) Software is used to segment the raw image into spots and background and intensity values are obtained for each spot and stored in table format. 5) Ratios of differential expression are calculated and various statistically derived or empirically derived thresholds are used to classify spots as differentially expressed or not, by combining a number of replicate hybridizations. 6) This data can be combined with other array data obtained under different conditions or different time points, and various data visualization and data exploration algorithms such as clustering or principal components analysis are applied (Geschwind 2001). Figure reprinted by permission from Nature Reviews Neuroscience Vol 2 (6): 437 copyright (2001) Macmillan Magazines Ltd.



3. Experimental Design

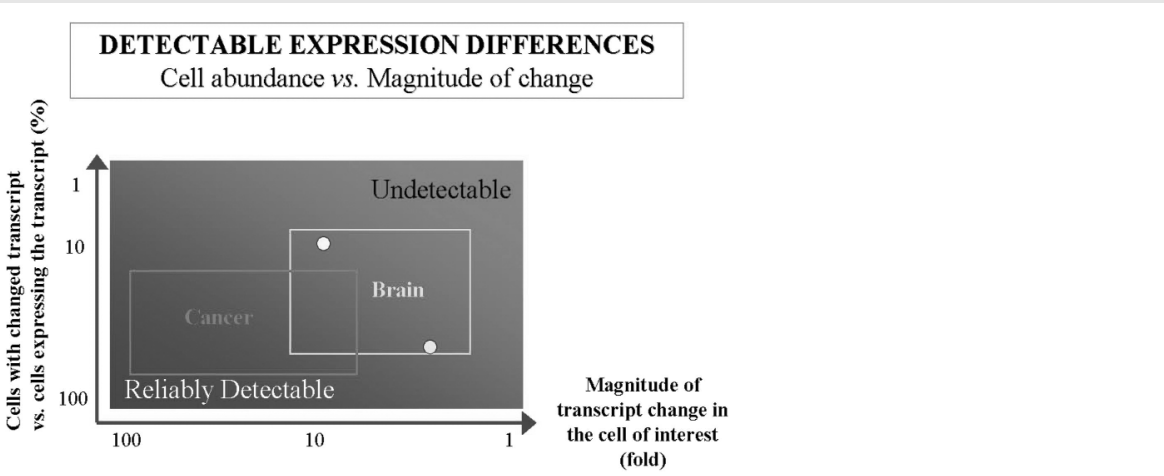
Once the hybridization has been performed, the analysis of microarray data has many potential steps, only a few of which are mandatory. Table 1 summarizes several basic "rules". Nowhere is the motto "Garbage in, Garbage out" more aptly applied. RNA quality and experimental design are the two most important factors in determining the quality of experimental outcome. Good resources for experimental design focused on the neuroscientist include Karoly Mirnic's chapter in last year's *DNA Microarrays* short course syllabus (Various 2001), Terry Speeds new book (Speed 2002), the microarray chapter in the Wiley current protocols series, *Current Protocols in Neuroscience* (Unit 4.28); (Karsten 2002), and the newly available multi-author book, *Microarrays for the Neurosciences* (Geschwind 2002).

Figure 5 provides a general schematic of the relationship of various experimental and analytic steps to each other, adopted from Kevin Becker's lecture in the 2001 SFN short course (Various 2001). Technical analysis to determine whether the experiment is of good quality is the usual first step once the hybridization has been performed. Visual inspection of the array for gross abnormalities such as missing spots, smears, poor blocking, etc. is always important. In addition, replicate arrays can also be clustered to identify possible outliers that should be discarded, augmenting the process of visual inspection. Assessment of hybridization signals versus background and the number of "hits" is also important. Similar experiments should have similar levels and hits. Commercial arrays, such as Affymetrix and Agilent often have a large number of sophisticated internal quality standards.

The second step involves statistical analysis to answer the question of which genes are differentially expressed at a significant level. There are several valid approaches to this. Many experimental confounds such as non- linearities due to dye incorporation effects and other factors, although present, may not have significant influence on the actual list of genes identified as differentially expressed. We have performed two quite different

Figure 2

If you don't see it, it doesn't mean that it isn't there. Illustrates that low abundance mRNAs detectable in a single cell may be below the detection limit in a complex tissue containing many cell types. In addition, expression changes in genes that are detectable, but near this threshold may not be detected as reliably as those with stronger signals.



analyses on the same data sent and had an approximately 85% overlap in gene lists, despite different expression thresholds and normalization methods. Reasonable quality data sets when analyzed with reasonable methods should yield robust results.

The final two steps described here are recommended but not required and can be done in parallel. One of these involves exploratory data analysis, such as clustering, and dimensionality reduction techniques, such as principal component analysis to perform higher order data analysis. These techniques can be done in a multitude of situations and the precise application is entirely dependent on the experimental design. The other can be thought of as biological analysis and involves using manual or bioinformatic approaches to identify protein families, transcriptional control factors or literature searches to provide a more solid biological context for the data and to generate new hypotheses for downstream experiments. Some general reviews for statistical analysis of spotted arrays include *Computational Analysis of Microarray Data* (Quakenbush, (Nadon and Shoemaker 2002; Sabatti, Karsten et al. 2002) Chiarra Sabatti?s chapter in the 2001 *Microarrays* short course syllabus (Various 2001) and many websites, most notably Terry Speed's website at Berkeley (http://stat-www.berkeley.edu/users/terry/zarray/Html/). Statistical analysis of affymetrix arrays is sufficiently different that we refer you to Li and Wong (2001), Tusher, Tibshirani et al. (2001), and www.Affymetrix.com instead.

4. Informatic Assessment of Array Data

Rather than focusing on specific databases, websites, etc themselves because these resources are in flux and improving, we now focus on potential uses of bioinformatics tools following the identification of an initial gene list is made (of differences, or critical genes) and give some examples.

Figure 5
Adapted from Kevin Becker, PhD.

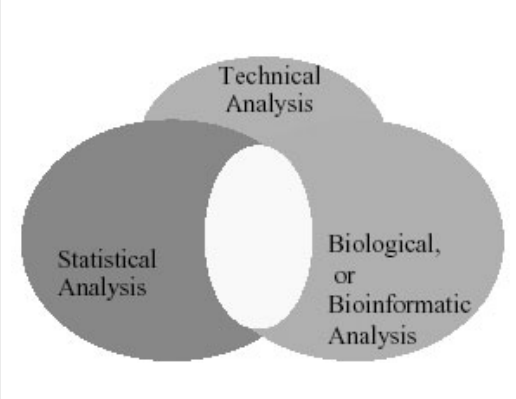


TABLE 1

Fundamentals of experimental design.

1) Perform an adequate number of independent replicates to identify differentially expressed genes of interest. Three replicates is a minimum for most typical experiments (Lee et al. 2001). RNA quality check is mandatory.

2) Assess array quality and biological variability by inspection and image analysis. Homotypic hybridizations (same sample compared to itself; See Sabatti et al., 2001), and array versus array plots (independent replicates of the same experiment; see Figure 3) are often helpful in this regard.

3) Perform appropriate normalization. Data is typically log transformed. Standard techniques include global normalization, normalization to standards on the array such as housekeeping genes, and normalization to doped in controls or a reference sample. Typically, non-linear normalization is used since ratios will deviate as a function of signal strength for a variety of reasons (Figure 4, "A vs. M" plot).

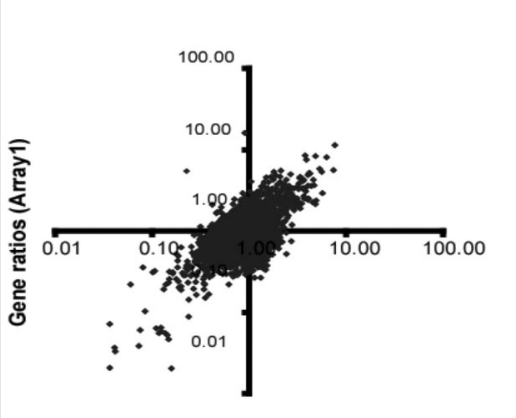


FIGURE 3

Comparing independent "replicate"arrays. Gene expression ratios between 2 experiments conducted using independent samples, but making the same basic comparison onto two different arrays. Tyramide signal amplification was used to amplify 1 ug of total RNA from each sample for hybridizaton onto a mouse ten thousand spot array. This not only depicts variation due to the hybridization and arrays, but that due to biological variability between culture samples. Correlation between replicate experiments is usually over 95% when direct labeling methods are used, but falls into the 80% range when TSA amplificaton is used, necessitating more replicates to account for the variance.

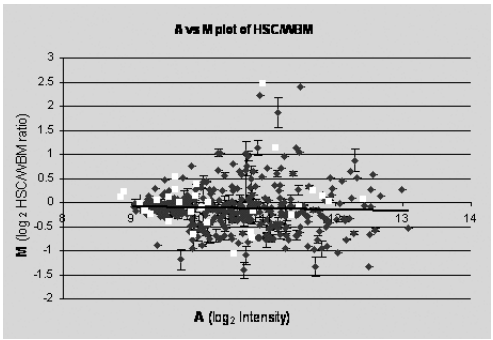


FIGURE 4

"A vs. M" plot. The A (signal intensity) vs M (Log2 ratios of differential expression) plot provides a good check of normalization of data, because relationship between ratio and intensity can be easily visualized (Yang 2001). In this case the data has been subjected to a non-linear normalization, so that there is no real skew in the data. For most arrays, ratios should be evenly distributed without regard to intensity if normalization has been properly performed, and the center of the distribution along the M axis should be around zero.

5. Initial annotation of gene list based on published literature (what is known?)

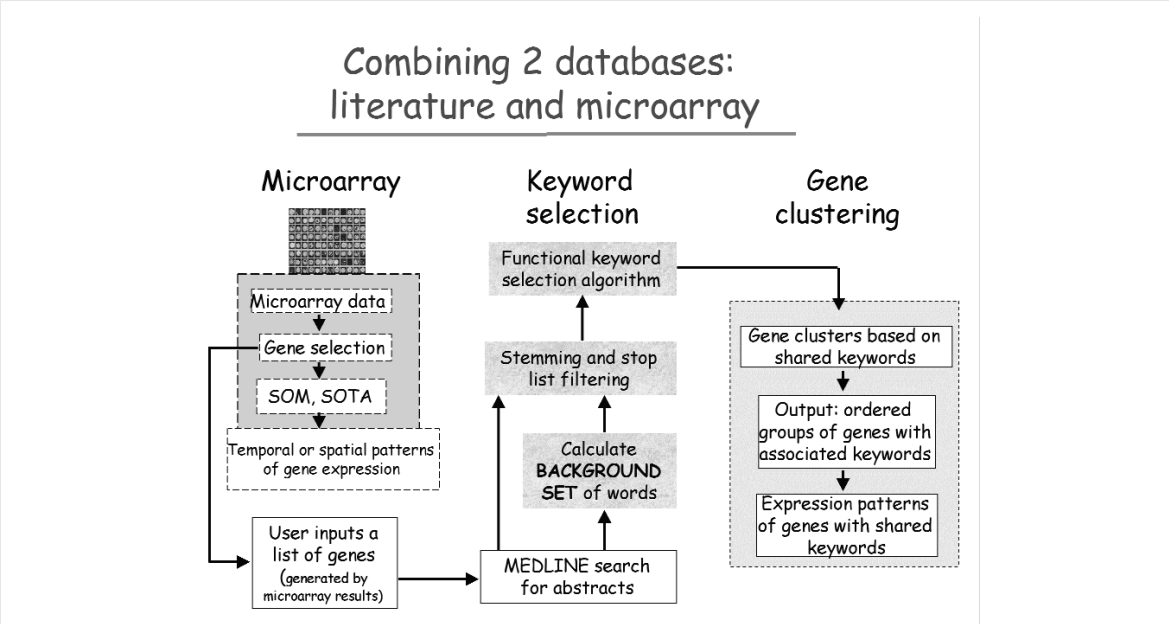
The first step in processing the set of differentially expressed genes is to find out what is currently known about them. Most of us usually rely on our own ever-shrinking fund of knowledge and manual literature searching, making this a tedious task, especially when hundreds of genes may be involved. Currently, there is no substitute for this approach of hitting the books (or keyboard). However, a variety of tools have recently been created to help in the task of placing microarray results into context of the previously existing literature. This is an important and yet daunting task as few scientists will have even a passing familiarity with every one of the thousands of genes that may be on a single array. To help reduce the amount of information that a scientist must sift through, and to help aid in the eventual interpretation of the results, many genes may be placed into pathways that visually describe their physiological roles and relationships to other genes.

GenMAPP

Tools to "map" the results of gene expression studies onto known pathways are now coming online. An example of such a tool is GenMAPP (www.genmapp.org) (Dahlquist, Salomonis et al. 2002) GenMAPP is a relatively simple program that allows the investigator to display their gene expression data on anyone of roughly a thousand pathways from the literature, or on a pathway that they themselves generate. The program simply color-codes genes on the pathway according to gene expression, following criteria set by the user. For example, genes with a 2-fold greater expresion in cancerous tissues may be color coded red, and those with a 2 fold greater expression in normal tissues be color-coded blue. This would allow one to quickly browse pathways to try and find ones that seem particularly relevant to cancer biology. The software runs in a windows environment, is free to the scientific community, and is relatively easy to set up and use. The most difficult portion is formatting

Figure 6

Linking array data to the literature. Starting with a list of genes derived from a microarray experiment, the program uses a highly developed keyword hierarchy to cluster genes by shared keywords, developing a kind of literature networks of genes. This tool is still in development. Courtesy of Ray Dingeldine.



your data and gene IDs to be read by the program. However, the program has a relatively well-developed online help and tutorial. Products such as this may help investigators to more rapidly uncover pathways of interest in their system without an exhaustive literature search of thousands of genes. Other new informatic tools can also help add to data garnered from manual searching. For example Ray Dingeldine at Emory has been working to adapt a literature?array linking tool so as to make it a more broadly useful for neuroscientists (Figure 6), and Jenssen et al have developed PubGene (figure 7), www.pubgene.org; (Jenssen, Laegreid et al. 2001).

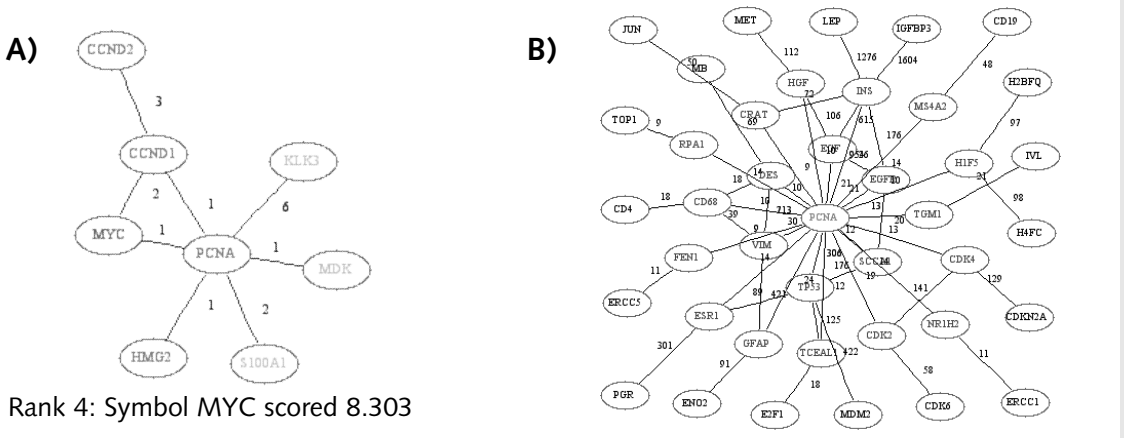
Pubgene

Pubgene is a "gene-to-gene co-citation network for 13,712 named human genes." (Jenssen, Laegreid et al. 2001) To develop the initial pubgene database, Jenssen et al. compiled text and abstracts from millions of medline records and gene identifiers were found (Official Gene symbols in Locuslink were used as the identifiers). Co-occurences were identified and a network of relationships based on the frequency of co-citation were compiled and databased. This database now can be searched with an input gene list and expression values; Up-regulated genes are marked as positive and down-regulated genes are marked as negative in this list. Relationships between genes identified on the array are depicted along with coding of the direction of the gene?s regulation in the input data set. Other investigators have developed similar tools based on Genbank accession numbers (www.array.ucsd.edu). This database uses medline keywords to help place genes into functional categories and identify relationships.

Unfortunately, these tools identify about 40 to 50 percent of gene co-occurences that are known at best. For the example shown in Figure 7, intensive

FIGURE 7

An example of a Pubgene output: We performed a Pubgene analysis on a short list of differentially expressed genes from an experiment comparing two neural progenitor-containing cultures. We submitted 136 official gene names (Locus Link) that were differentially expressed between the two conditions and their ratios together in one file at the Pubgene website (www.pubgene.org). Twenty five of the 136 genes had literature identified neighbors, which were depicted in 10 images displaying the literature network and direction of interaction. One such image is presented here (A). When any of the nodes on this network are clicked, an expanded network is displayed (B).



Rank 4: Symbol MYC scored 8.303

manual literature searching, as expected, identified a much richer list of genes involved in cell cycle and DNA repair that were regulated. Although this manual approach took several weeks, it provided a depth greater than Pubgene in its current incarnation, but at much greater time cost. For ESTs or less studied genes, this tool will be even less useful. Even so, these tools can identify relationships, and even disease associations that were previously not well appreciated in known genes by creating complex networks. Other widely used tools, such as Locuslink and Genecards databases also are useful aids for annotating gene expression lists on a gene-by-gene basis. However, how does one go about annotating and exploring the myriad ESTs and less studied genes?

One can start by "virtual northern blotting" — searching EST hits in sequenced libraries at NCBI such as Unigene and compiling them across tissues, i.e. BLASTing EST databases to identify in as semi quantitative manner the tissue distribution of expression of a given novel transcript. Data from library sequencing projects involving SAGE is growing and also can be used to identify tissues that a gene is expressed in. But SAGE libraries currently comprise only a small subset of tissues and conditions (http://www.ncbi.nlm.nih.gov/geo/).

Table 2

Table 1 | **Sharing paradigm advantages and disadvantages**

	Raw image	Raw spot and background data	Normalized, processed data (ratio)	Gene list
Advantages				
Overall flexibility	+++	++	+	
Least biased	+++	++	+	
Allows re-analysis	+++	++	+	
Comparison of analytical methods	+++	++	++	
Image segmentation	+++			
Image analysis	+++			
Statistical flexibility	+++	+++	++	
Data mining flexibility	+++	+++	++	
Easy to use — least analytical demands		+	++	+++
Disadvantages				
Requires large storage	+++	+		
Requires image analysis time and software	+++			
Requires data analysis time	+++	++	+	
Potential cross-platform incompatability	+++	+		
Bias due to normalization			+++	+++
Not useable for some statistical methods			++	+++
Less useful for re-analysis of data			+/–	+++
Less useful for combining of data sets			+	++
Not quantitative				+++

6. Databases

While the need to find a suitable storage system may not be evident when running only a few arrays, once a number of experiments have been run, the need for a database solution becomes more serious. Several commercial (Affymetrix, Axon Instruments, Biodiscovery, Silicon Genetics), are just a few of the vendors with widely used products that have some databasing functions and shareware (eg. BASE; http://base.thep.lu.se/,(Saal 2002, in press) options are available for local storage and manipulation of microarray data. We are currently implementing A) B) Rank 4: Symbol MYC scored 8.303 BioArray Software Environment (BASE) on our laboratory Linux server for databasing of our own local experiments. One can also access Internet-based databases where large number of experiments are stored in a free environment, such as GEO, the Stanford Microarray Database, and Dragon. Many laboratories find that Excel or locally developed databases suffice for their individual needs, although as discussed below, the public databases have large advantages for the field. With great foresight, the MGEM group has developed standards for the minimal information about a microarray experiment and developed a format and language for describing microarray experiments that has been widely accepted. This will greatly facilitate data sharing with large public databases.

The need for large public databases for gene expression annotation and analysis.
It is clear that large public databases compiling already performed microarray experiments could be a huge advance in this area. For example, at this point, thousands of microarray hybridizations have been done on the Affymetrix platform, which uses a standard set of non-custom arrays. Suc h databases have been proposed and standards have been adopted, but resistance against such sharing still remains. For many reasons, we and others believe that data sharing will have huge benefits and should be mandated (Brazma, Hingamp et al. 2001; Geschwind 2001). One simple advantage for local data sharing of gene expression data on commonly used arrays is the availability of information on unknown genes across a large number of tissues, stages and experimental conditions that could be useful in assigning functional classes to unknown genes.

Any database system designed for arrays should take into account the common language and standards that has been developed for this purpose. The minimal information about a microarray experiment (MIAME) developed by the Microarray Gene Expression Group (MGED) has been agreed upon as a standard that should be used by all groups. The actual form that the shared data could take depends upon its use, and each form has advantages and disadvantages (Table 2). At least ratios, but optimally raw expression signal and background values should be made available. The benefits of such sharing to the community and individual investigator are clear (Brazma, Hingamp et al. 2001; Geschwind 2001). A real world example of use of unrelated experiments in different tissues to interpret a microarray experiment is presented below, along with other examples of sophisticated analytic and informatic methods.

7. Data Mining And Annotation By Chromosome

A recent example of data reduction, clustering, mining of local array databases and annotation by chromosomal location and confirmation of informatic-based hypothesis.

Paul Mischel, Stan Nelson, Tim Cloughsey and colleagues at UCLA wanted to determine whether certain brain tumors have a unique molecular signature. It was already thought that certain growth factor receptor pathways were altered in certain brain tumors. They designed an experiment using Affymetrix arrays with 12,600 genes to determine whether expression profiling could be used to identify subclasses of GFR and non-GFR over-expressing tumors (Mischel et al. 2002).

Data reduction. They first narrowed down the potential gene list by removing genes that don?t show much variation between arrays. Such genes are unlikely to contribute to the ability distinguish between groups within the samples since they don?t vary much. This identified about 4000 genes with co-efficient of variation over 0.5. This remaining list of genes was used to perform unsupervised hierarchical clustering from which 3 groups were clearly identified. One of these groups was GFR+ and the other two, GFR-. Multidimensional scaling led to 3 distinct clusters of balls, demonstrating the same groups. By further constraining the list by setting a differential expression threshold of 1.5 fold, 90 genes were identified that could define 3 clusters of brain tumors, 2 clusters containing tumors that were growth factor receptor negative, and one cluster that was postitive by immunocytochemistry. Leave-one-out cross validation analysis, coupled with weighted gene voting algorithm, such as applied in Pomeroy, Tamayo et al. (2002) was applied to 13 new samples, which were correctly classified by their pattern of gene expression into the 3 clusters, two of which were GFR-. The next question was, what was the difference between the two groups of GFR- clusters?

Assessing chromosomal location using genome databases. Bioinformatics had been used to annotate the array to map each gene to its precise chromosomal location. Simple resources at NCBI such as Locus Link provide such mapping data in human and homologues in other species, where known. Another new resource that may be useful for such searches is "The Gene Resource Locator", which has assembled gene maps, expression profiling data, and splicing data that allows the user to search and view information via a dynamic web viewer (http://grl.gi.k.u-tokyo.ac.jp).

The chromosome mapping demonstrated that some genes in one cluster of GFR- tumors mapped to a contiguous locus in the human genome. This suggested a region of genomic amplification or other means of co-regulation of this class of genes. This was further supported by the finding that all genes in that region present on the array were up-regulated in that cluster and that duplications in this region associated with the same category of brain tumors had been previously reported, suggesting this pattern of gene expression was associated with amplification of contiguous genes. Could this just be a co-incidence — are they always co-regulated and this has nothing to do with tumor specific transcriptional regulation?

Using databases of other array experiments to annotate gene functions. Because a local database of all Affymetrix experiments performed through the UCLA array facility was available, these investigators were able to ask if these genes were typically co-regulated. They assessed 368 experiments present in the array core database to ask if this set of nearby ge nes? expressions were correlated in any of the previous experiments in the database. No correlation was found, supporting the hypothesis that their co-regulation was specific to the underlying tumor biology being studied.

Validate informatic hypotheses prospectively. To validate the predictive value of these clusters and their reflection of underlying different biology of brain tumor classes and the finding of possible chromosomal duplication, 16 more cases from various sources were successfully used to predict GFR status based on gene expression with high accuracy.

8. Using Microarrays to study underlying regulatory mechanisms

Regulation of gene transcription is arguably the most important step in regulation of proteins expressed in a cell. It is very complex, highly regulated and usually depends on promoter sequences to which transcription factors bind that are typically located upstream of the transcription start site. Enhancer elements are also important and have a more variable location, and can be intronic, or more than 1 kb upstream of the start site. Complex patterns of gene regulation in Yeast and eukaryotes analyzed by microarrays has allowed identification of promoter regions in these organisms (Werner 2000; Pilpel, Sudarsanam et al. 2001; Werner 2001; Werner 2001; Brazma, Jonassen et al. 1998; Brazma, Jonassen et al. 1998).

Identify co-regulated genes. We performed a microarray comparison of neural progenitor cultures enriched in neural stem cells (NS) to more differentiated cultures to identify genes enriched in neural stem cells (eg. Geschwind et al. 2001). A large number of clones enriched in the NS condition were identified. Northern blotting and *in situ* hybridization confirmed differential expression. Unfortunately, we did not have microarray expressio n data on a large number of conditions or time points during development that would permit identifying genes that appeared co-regulated, such has been done in Yeast and prokaryotes. But, we were able to use *in situ* hybridization to identify genes enriched in embryonic germinal zones.

Collect 5kb upstream of transcription start site. We inferred that the shared temporal and spatial expression pattern may allow the identification of promoter regions responsible for the germinal zone pattern. Using NCBI, we found sequences up to 5kb upstream of ATG for the human homologues of 8 genes with strong VZ pattern. *Tracer*, a transcript sequence retrieval software from Stanford with various sequence retrieval capabilities, streamlined this process by allowing the user to find the upstream sequence to a given transcript sequence for a gene given its LocusLink ID number. We used the *Repeatmasker* Web Server to mask the repeat regions in the promoter regions of interest (http://repeatmasker.genome.washington.edu/cgi bin/ RepeatMasker) and performed the next stages with and without masking. Initially, the sequences were split into segments of up to 500 bp, with 100 bp overlap where appropriate, i.e. the segments were not separated by masked regions.

Perform similarity search to identify shared sequences. The 5?regions were then sent to MEME (http://meme.sdsc.edu/meme/website/meme-intro.html), which searches for regions of shared similarity among the sequences submitted. Initially one had the inconvenience of fragmenting the large submissions, but the MEME server has been upgraded to allow much larger submissions, simplifying this process (Grundy, Bailey et al. 1997). Sequences that are identified through MEME submission were then sent to MAST, which searches in other databases (eg. nr at NCBI; http://meme.sdsc.edu/meme/website/mast-intro.html) to find other genes that share the motif in their upstream sequences. The eukaryotic promotor dabase (EPD), which provides a list of sequences 500 bp upstream of the beginning of transcription site ("promoter regions") for over 1400 genes is another one of the databases that MAST can search (Praz, Perier et al. 2002). Unfortunately, the database is still small; however, it provides insight into the types of genes that may share similar motifs, as those being investigated. A flow chart summarizing the entire process is shown in Figure 8. Other tools such as SCANace and AlignAce can be used to identify putative shared regulatory regions (Roth, Hughes et al. 1998), http://arep.med.harvard.edu/mrnadata/mrnasoft.html.

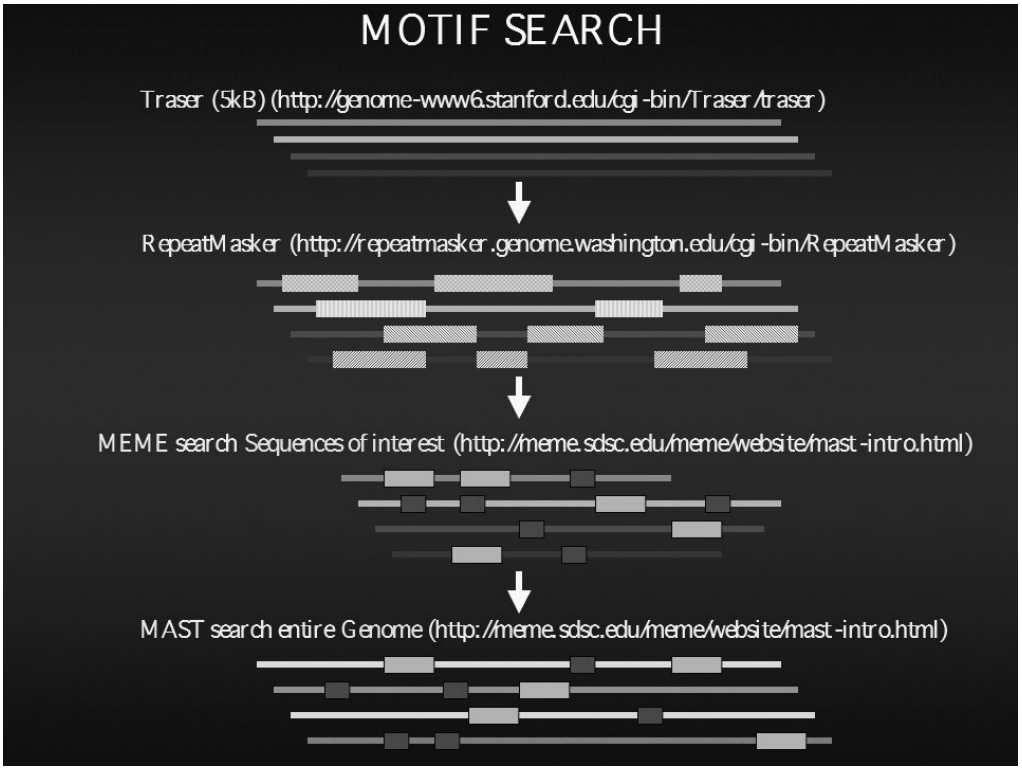
Follow-up interesting shared sequences to identify putative transcription factor binding sites. We identified several interesting putative regulatory regions, including a 50 bp sequence, which was subsequently submitted to Tfsitescan at IFTI (http://www.ifti.org) to determine whether any known transcription factor binding sites lay

within this 50 bp sequence. Four genes having this motif were identified in the human genome, two of which were identified in our microarray study, and another of which was previously known to be enriched in the mouse embryonic ventricular zone, providing some confirmation that this motif might be meaningful.

Pitfalls/Suggestions. This kind of analysis identified a large number of motifs and the problem is identifying which motifs are important. Since transcription factor binding sites are small, such motifs occur frequently by chance in the genome. Therefore, statistical validation that the sequence is over-represented in the genes in question is critical and focusing on rare motifs, or several sites occurring together in a particular pattern increases the likelihood of identifying functionally relevant regulatory regions (Hughes, Estep et al. 2000; Werner 2000) We have also simply searched the 5?sequences for known regulatory motifs, using software such` as Transfac (Quandt, Frech et al. 1995). Again, this results in a large number of hits, so it helps to filter the data by focusing on patterns and groupings, rather than individual motifs. Sophisticated commercial applications are available for this purpose http://www.genomatix.de/ (Werner 2001), as well as non-commercial shareware (Roth, Hughes et al. 1998).

FIGURE 8

Promotor/Enhancer Motif Search: This provides an example of the steps that one might take to identify common regulatory elements. This area of informatics and computational biology is in its infancy and it will be valuable to spend some time upfront investigating other resources prior to embarking on this path. For example Other tools such as SCANace can be used to identify putative shared regulatory regions



9. Summary

Bioinformatic analysis of arrays is in a very early stage. Even so, several options for combining array databases or gene lists with other relevant databases exist, some of which have been discussed here. These approaches clearly highlight the need for large publicly accessible databases of gene expression data culled from microarray experiments using similar arrays. In addition, other databases of genome-wide expression, such as GENSAT, which aims to offer access to *in situ* expression data (in mouse) on all genes in the CNS, offer the potential for unprecedented annotation of the genome. As these resources develop, the idea that microarray experiments will really begin after the list of differentially expressed genes is generated will be realized.

Acknowledgements

We thank Karoly Mirnics MD PhD, Kevin Becker PhD and Ray Dingeldine PhD for sharing their schematic figures with us and Paul Mischel, MD and colleagues for sharing their data and figures with us. We are also grateful to the UCLA array facility directed by Stan Nelson MD, for its collaboration and support and our other collaborators including Harley Kornblum MD PhD and members of the Geschwind lab for their contributions to work discussed in this syllabus. Our microarray work is supported by grants from (PI DHG: NIMH grant #MH60233, NINDS grant #NS40752, NIA AG16570; DHG as Co-PI: NIMH grant # MH65756, NIMH grant #MH62800 and NINDS NS43562; JDD: Howard Hughes Medical Institute Pre-doctoral Fellowship).

10. Resources

Open source code: (http://www.open-bio.org)

Arrayit.com: (http://www.arrayit.com)

Stanford page (brown): (http://cmgm.stanford.edu/pbrown/)

UCLA web page: (http://www.genetics.ucla.edu/microarray/)

TIGR:(http://www.tigr.org/)

DeRisi: (http://derisilab.ucsf.edu/)

NCGR: (http://www.ncgr.org/)

Stanford Microarray database http://genome-www5.stanford.edu/MicroArray/SMD/

Listing Microarray databases: http://www.Biologie.ens.fr/en/genetiqu/puces/bddeng.html

European Bioinformatics Institute: http://www.ebi.ac.uk/microarray/index.html

Telechem electronic library on microarray topics http://arrayit.com/e- library/

Papers and Links: http://linkage.rockefeller.edu/wli/microarray/

Microarray analysis R program http://www.stat.uni-muenchen.de/~strimmer/rexpress.html

YF Leung: Database software reviews

http://ihome.cuhk.edu.hk/~b400559/arraysoft.html#Database%20Software

Nucleic Acids Research 2002 Database issue http://nar.oupjournals.org/content/vol30/issue1/

Harvard-Lipper Center for computational genomics http://arep.med.harvard.edu/

REFERENCES

Becker, K. G. (2001). "The sharing of cDNA microarray data." Nat Rev Neurosci 2(6): 438-40.

Brazma, A., P. Hingamp, et al. (2001). "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." Nat Genet 29(4): 365-71.

Brazma, A., I. Jonassen, et al. (1998). "Approaches to the automatic discovery of patterns in biosequences." J Comput Biol 5(2): 279-305.

Brazma, A., I. Jonassen, et al. (1998). "Predicting gene regulatory elements in silico on a genomic scale." Genome Res 8(11): 1202-15.

Brown, P. O. and D. Botstein (1999). "Exploring the new world of the genome with DNA microarrays." Nat Genet 21(1 Suppl): 33-7.

Cirelli, C. and G. Tononi (1999). "Differences in brain gene expression between sleep and waking as revealed by mRNA differential display and cDNA microarray technology." J Sleep Res 8 Suppl 1: 44-52.

Dahlquist, K. D., N. Salomonis, et al. (2002). "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways." Nat Genet 31(1): 19-20.

DeRisi, J. L. and V. R. Iyer (1999). "Genomics and array technology." Curr Opin Oncol 11(1): 76-9.

DeRisi, J. L., V. R. Iyer, et al. (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science 278(5338): 680-6.

Eberwine, J., H. Yeh, et al. (1992). "Analysis of gene expression in single live neurons." Proc Natl Acad Sci U S A 89(7): 3010-4.

Geschwind, D. H. (2000). "Mice, microarrays, and the genetic diversity of the brain." Proc Natl Acad Sci U S A 97(20): 10676-8.

Geschwind, D. H. (2001). "Sharing gene expression data: an array of options." Nat Rev Neurosci 2(6): 435-8.

Geschwind, D. H., Gregg J (2002). Microarrays for the Neurosciences: An Essential Guide. Cambridge, MA, The MIT Press.

Geschwind DH, Ou J, Easterday MC, Dougherty JD, Jackson RJ, Chen Z, Antoine H, Terskikh A, Weissman IL, Nelson SF, Kornblum HI (2001). A genetic analysis of neural progenitor differentiation. Neuron 29:325-339.

Ginsberg (2001). Gene Expression Profiling Using Single Cell Microdissection Combined with cDNA Microarrays. Society for Neuroscience Short Course. ed. Geschwind DH. San Diego, Society for Neuroscience.

Grundy, W. N., T. L. Bailey, et al. (1997). "Meta-MEME: motif-based hidden Markov models of protein families." Comput Appl Biosci 13(4): 397-406.

Hughes, J. D., P. W. Estep, et al. (2000). "Computational identification of cis-regulatory elements associated with groups of functionally related ge nes in Saccharomyces cerevisiae." J Mol Biol 296(5): 1205-14.

Jenssen, T. K., A. Laegreid, et al. (2001). "A literature network of human genes for high throughput analysis of gene expression." Nat Genet 28(1): 21-8.

Karsten, S. L., Geschwind D. H. (2002). Gene expression analysis using cDNA microarrays. Current Protocols in Neuroscience. 20: unit 4.28 (in press).

Karsten, S. L., V. M. Van Deerlin, et al. (2002). "An evaluation of tyramide signal amplification and archived fixed and frozen tissue in microarray gene expression analysis." Nucleic Acids Res 30(2): E4.

Li, C. and W. H. Wong (2001). "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection." Proc Natl Acad Sci U S A 98(1): 31-6.

Lockhart, D. J., H. Dong, et al. (1996). "Expression monitoring by hybridization to highdensity oligonucleotide arrays." Nat Biotechnol 14(13): 1675-80.

Lockhart, D. J. and E. A. Winzeler (2000). "Genomics, gene expression and DNA arrays." Nature 405(6788): 827-36.

Luo, Z. and D. H. Geschwind (2001). "Microarray applications in neuroscience." Neurobiol Dis 8(2): 183-93.

Miles, M. F. (2001). "Microarrays: lost in a storm of data?" Nat Rev Neurosci 2(6): 441-3.

Mirnics, K. (2001). "Microarrays in brain research: the good, the bad and the ugly." Nat Rev Neurosci 2(6): 444-7.

Mischel P, N. S., Cloughesey T et al. (2002). personal communication. D. Geschwind.

Nadon, R. and J. Shoemaker (2002). "Statistical issues with microarrays: processing and analysis." Trends Genet 18(5): 265-71.

O'Dowd, B. F., T. Nguyen, et al. (1998). "Discovery of three novel G-protein-coupled receptor genes." Genomics 47(2): 310-3.

Okubo, K., N. Hori, et al. (1992). "Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression." Nat Genet 2(3): 173-9.

Pilpel, Y., P. Sudarsanam, et al. (2001). "Identifying regulatory networks by combinatorial analysis of promoter elements." Nat Genet 29(2): 153-9.

Pomeroy, S. L., P. Tamayo, et al. (2002). "Prediction of central nervous system embryonal tumour outcome based on gene expression." Nature 415(6870): 436-42.

Praz, V., R. Perier, et al. (2002). "The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data." Nucleic Acids Res 30(1): 322-4.

Quandt, K., K. Frech, et al. (1995). "MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data." Nucleic Acids Res 23(23): 4878-84.

Roth, F. P., J. D. Hughes, et al. (1998). "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation." Nat Biotechnol 16(10): 939-45.

Saal, L. H., Troein, C. Vallon-Christersson, J. Gruvberger,S. Borg ?. and Peterson C. (2002, in press). "BioArray Software Environment: A Platform for Comprehensive Management and Analysis of Microarray Data." Genome Biology.

Sabatti, C., S. L. Karsten, et al. (2002). "Thresholding rules for recovering a sparse signal from microarray experiments." Math Biosci 176(1): 17-34.

Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science 270(5235): 467-70.

Speed, T. P. (2002). Statistical Analysis of Gene Expression Microarray Data. Boca Raton, Fl, CRC Press LLC.

Stein, L. (2002). "Creating a bioinformatics nation." Nature 417(6885): 119-20.

Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A 98(9): 5116-21.

Various Authors (1999). "The Chipping Forcast." Nature Genetics 21(Supplement).

Various Authors (2001). DNA Microarrays: The New Frontier in Gene Discovery and Gene Expression Analysis. ed. DH Geschwind. Society for Neuroscience Short Course, San Diego, Society for Neuroscience. Available at : http://web.sfn.org/Template.cfm?Section=BrowsebyType&template=/Ecommerce/ProductDisplay.cfm&ProductID=87

Velculescu, V. E., L. Zhang, et al. (1995). "Serial analysis of gene expression." Science 270(5235): 484-7.

Wang, K., L. Gan, et al. (1999). "Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray." Gene 229(1-2): 101-8.

Werner, T. (2000). "Identification and functional modelling of DNA sequence elements of transcription." Brief Bioinform 1(4): 372-80.

Werner, T. (2001). "Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data." Pharmacogenomics 2(1): 25-36.

Werner, T. (2001). "The promoter connection." Nat Genet 29(2): 105-6.

Werner, T. (2001). "Target gene identification from expression array data by promoter analysis." Biomol Eng 17(3): 87-94.

Yang, Y. H., Buckley, M. J. , Dudoit, S. and Speed, T.P. (2001). Normalization for cDNA Microarray Data. SPIE BiOS. San Jose, Ca.