

BIOCHIP PLATFORMS AS FUNCTIONAL GENOMICS TOOLS FOR DRUG DISCOVERY

Ivan Wick¹ and Gary Hardiman^{1,2}

¹Biomedical Genomics Microarray Facility (BIOGEM), and ² Department of Medicine, University of California San Diego, La Jolla CA 92093-0724

Tel.: 858 822 3792

Fax: 858 822 6430

E-mail: ghardiman@ucsd.edu

Keywords Microarrays, Biochips, gene expression, platform, data integration

Improvements in DNA microarray technology have generated data on a scale that for the first time permits detailed scrutiny of the human genome. This provides the infrastructure to understand not only the wiring diagram and annotation of the human genome, but also the molecular basis of genetic defects. These advances have the potential to significantly improve healthcare management by improving disease diagnosis and specifically targeting molecular therapy. Herein, we review the current state of the technology, we compare and contrast the commercial platforms used by the biopharmaceutical industry and we explore recent efforts at cross-platform data integration.

INTRODUCTION

At the present time the pharmaceutical industry faces an upswing in research and development costs, whilst the number of novel molecular entities reaching the market grows at a significantly slower rate. This has forced the industry to devise and adapt methodologies that have the potential to increase the number of new drug candidates in the pipeline, within a much shorter time frame (1). In the past decade high-density DNA microarrays and biochips have revolutionized the field of biomedicine and helped accelerate target validation and drug discovery efforts (2). Microarrays are still predominantly used for gene expression analyses, but they are also finding application in genotyping and re-sequencing applications, in addition to comparative genomic hybridization studies. They have been utilized to address *in vitro* pharmacology, and toxicology issues and are being widely applied to improve the processes of disease diagnosis, pharmacogenomics, and toxicogenomics (3-6).

The power of microarray technology lies in its ability to perform massive parallel profiling of gene expression from a single sample. A global view of gene expression provides a snapshot of the transcriptome in healthy and disease states. This information is highly useful as it uncovers gene families or more specifically pathways that are affected, but also reveals those that are unaffected (7). Hypotheses about genes with unknown function can also be formed by comparison of their expression levels with genes of known function (8). Similar expression profiles imply that genes may be co-regulated.

Microarray experimentation is a complex process and significant time and effort are required to design biologically sound and statistically robust experiments. Once target genes are identified additional time and expense are required to validate their selection and relevance. Drug discovery programs utilizing microarray technologies must therefore consider all available technologies before allocating precious resources. Several complementary microarray technologies for measuring gene expression have evolved. However platform evaluations are impractical for the majority of researchers, as this involves considerable expenditure, and often a commitment to dedicated hardware and software (9,10).

FROM CANCER CLASSIFICATION TO DRUG DISCOVERY

The seminal publication that unleashed the power of gene expression and microarray technologies, demonstrated accurate classification of hematologic malignancy-acute myeloid leukemia (AML) and acute lymphoid leukemia (ALL) (11). This study demonstrated that molecular signatures could clearly classify patients at the clinic. Furthermore it revealed the utility of microarrays in several areas: tumor classification, prediction of tumor classes, molecular diagnostics, in addition to elucidating the genetic defects that lead to cancers. Subsequently, gene expression in many solid tumors was studied using DNA microarrays. In the case of breast and lung cancers, the focus switched from tumor classification (12-14) to dissecting solid tumors in the context of patient survival (15), or defining tumors by metastasis signatures (16,17). These cancer studies uncovered gene expression-class based molecular classification of cancer. Thus the notion arose that transcriptional profiling could help discover new tumor classes, pathway defects, patient stratification for treatment and discovery of new drug candidates.

MICROARRAY EXPRESSION PLATFORMS

Many competing technologies have gained acceptance by the pharmaceutical industry including full-length cDNA arrays, or pre-synthesized or *in situ* synthesized oligonucleotides as probes (17,18). The standard experimental paradigm compares mRNA abundance in two different biological samples, on the same or replicate microarrays. Each of the respective platforms has been optimized to work with either a single or dual color detection system. The key trends recently have been a shift from cDNA- to oligonucleotide-based microarrays and from 'in-house or home-brew' to higher quality commercial platforms. The salient features of those platforms predominantly utilized by the pharmaceutical industry are presented below. Advances in laboratory automation have improved the sensitivity, specificity and reproducibility of microarray experimentation, including advances in automated hybridization, sample preparation, and preparation.

Affymetrix

Affymetrix (Santa Clara, Ca) pioneered this field and has dominated for many years, applying photolithographic technologies derived from the semiconductor industry to the fabrication of high-density microarrays. The GeneChip has become the pharmaceutical industry standard owing to its extensive genetic content, high levels of reproducibility and minimal start up time. The GeneChip consists of short single stranded DNA segments, oligonucleotides or oligos, which are built to order by

chemical synthesis (17). A major advantage of GeneChips is that they are designed *in silico*, thereby eliminating management of DNA clone libraries, and the possibility of misidentified tubes, clones, or features (19). The disadvantage of this platform is that it demands a dedicated scanner and utilizes short 25-mer oligonucleotides, which are less sensitive than the longer 60-mers utilized in other technologies. Additionally multiple oligonucleotides are required for transcript detection.

Agilent

Alternative platforms are emerging and are challenging the dominance of Affymetrix. One of these, a two-color Agilent Technologies (Palo Alto, CA) relies on the *in situ* synthesis of probes by ink jet printing using phosphoramidite chemistry. Ink jet technology has been utilized in the past by Agilent to fabricate spotted cDNA arrays from PCR amplicons. This cDNA-based platform has largely been retired in favor of the superior oligonucleotide format, which consists of 60-mers contrasting with the short 25-mers probes employed by Affymetrix. Although short oligonucleotides should in theory provide the greatest discrimination between related sequences, they often have poor hybridization properties. The 60-mers provide enhancements in sensitivity over 25-mers in part to the larger area available for hybridization. Another advantage is that one 60-mer per gene or transcript is required (20).

Amersham CodeLink Bioarray

Another widely used platform by the pharmaceutical industry is the CodeLink Bioarray platform from Amersham Biosciences (now part of GE Healthcare) (Piscataway, NJ). 30-mer oligonucleotides are synthesized *ex situ* using standard phosphoramidite chemistry. This has the advantage that the probes can be validated by mass spectrometry prior to non-contact piezoelectric deposition on a proprietary three-dimensional polyacrylamide gel matrix. Covalent attachment is achieved via covalent interactions between 5' amine groups on the oligonucleotide probes and functional groups on the slide surface. Advantages with this platform is that the three dimensional nature of the slide surface promotes an aqueous biological environment and solution phase kinetics which enhance assay sensitivity. It is an open system that can be utilized with any microarray scanner. The disadvantages are that one oligonucleotide probe, a 30-mer, is used to interrogate a particular gene, which is potentially less sensitive than a 60-mer (21).

Applied Biosystems Expression Array System

A recent addition to the field is the Expression Array System from Applied Biosystems (Foster City, CA). Standard phosphoramidite chemistry is employed to synthesize 60-mer oligonucleotides, which similar to the CodeLink platform are validated offline by mass spectrometry, and are then deposited onto a modified nylon microarray substrate. The 3' end of the oligonucleotide is covalently coupled to the nylon via a carbon spacer, which raises the oligonucleotide and helps avoid steric-hindrance. Chemiluminescence detection distinguishes the ABI system from the other commercial platforms and provides increased detection sensitivity, as unlike fluorescence detection, an excitation step is not required, thereby minimizing background noise. The disadvantages with this platform are that it is not amenable to customization and it requires a dedicated chemiluminescence reader.

Illumina BeadChip and Sentrix Array Matrix

Illumina (San Diego, Ca) has developed a bead-based technology for SNP genotyping and gene expression profiling applications on two distinct substrates, the Sentrix LD BeadChip and the Sentrix Array Matrix (which facilitate up to 8 and 96 sample respectively). Both are fabricated to utilize an 'array of arrays' format, which enables the processing of multiple samples at the same time. Each array on each substrate contains thousands of tiny etched wells, into which thousands to hundreds of thousands of 3-micron beads self-assemble in a random fashion. 50-mer gene-specific probes concatenated with 'address or zip-code' sequences are immobilized on the bead surface. After bead assembly, each array is 'decoded' via a proprietary process, to determine which bead type containing which sequence, is present in each well of the substrate. The advantages of this platform are its sensitivity and reproducibility. The oligonucleotide probes can be validated off-line, and the low-density bead chip format can be utilized with any Microarray Scanner, capable of scanning at 5 μ resolution. The Sentrix Array Matrix offers major increases in throughput to the pharmaceutical industry, but this high-density format requires a dedicated scanner.

WHICH PLATFORM IS SUPERIOR?.

Each technology possesses inherent advantages and disadvantages, and currently no one platform offers superiority. In short, longer probes provide greater sensitivity at the expense of reduced specificity. Consequently the use of 25-mers requires the use of multiple oligonucleotides probes per transcript. For example the optimal feature of the Amersham CodeLink system is its excellent sensitivity owing to the three-dimensional nature of the surface. This platform is also advantageous in

that it is open, and can be used with a range of microarray scanners. However as this platform is fabricated via deposition rather than *in situ* synthesis, certain features may exhibit poor spot morphology and contamination artifacts. Affymetrix has in its favor the fact that it is a mature platform with an extensive array catalog that has been widely used by the pharmaceutical industry. There is however a requisite commitment to dedicated hardware. The Agilent platform is highly reproducible and the most sensitive of the various array platforms. Furthermore considerable cost savings are realized, as this is a two-color platform. However, the two-color approach also has potential disadvantages as different fluorescently labeled nucleotides may be incorporated with different frequencies, altering ratios due to enzymatic parameters rather than actual transcript abundance. Additionally, multiple experiment comparisons are not possible without replicating the reference sample (which, in the case of biopsy material, may be difficult to obtain).

Figure 1 highlights the evolution and overlap of catalog human arrays from Affymetrix, Amersham CodeLink and Agilent respectively. With improvements in the technologies and the availability of annotated human genome sequence, each of the major providers has released higher density arrays with smaller feature sizes. Unigene identifiers were the most complete and common identifiers among probes on every array, so this information was used to determine the genes commonly represented amongst all the arrays.

EXPRESSION PLATFORMS AND DATA VARIABILITY

Several DNA chip technology formats have evolved and carefully designed studies have been performed to evaluate the interchangeability of data from various platforms. The combination of expression measurements from different technologies within a single analysis would realize considerable cost savings and reduce the need for duplicate experiments in separate laboratories. The outlook was initially bleak for cross platform data integration from the early comparisons that were carried out. The discordance observed was attributed to the disparity inherent in each of the respective platforms. Differences arise from the intrinsic properties of the biochips themselves, and the various processing and analytical steps involved. This led to the following series of questions. Is one platform generating incorrect data? Do different biochips accurately reflect true biological expression? If differences are reported, then what actually is the true biological expression profile?

Affymetrix versus cDNA microarrays

Variability in measured gene expression levels, associated with different platforms, particularly cDNA-based microarrays has hindered integrative efforts. Kothapalli *et al.*, first reported inconsistencies in cross-platform data (22). Gene expression levels in peripheral blood mononuclear cells (PBMC) of a large granular lymphocyte leukemia patient and a healthy control were determined using Affymetrix oligonucleotide GeneChips and Unigem V spotted cDNA microarrays (Incyte, Wilmington, DE). This study highlighted problems often encountered with cDNA microarray platforms, namely inconsistent sequence fidelity of the spotted microarrays, variability in differential expression levels, low specificity of cDNA microarray probes, discrepancies in the fold-change calculation compared to Affymetrix, and lack of probe specificity for different isoforms of a gene. Considerable variation exists in the cDNA libraries used to generate spotted cDNA microarrays and errors in handling bacterial clones and cross contamination have been well documented (19).

Subsequent studies revealed similar discrepancies in cDNA based platforms (23-25). Li *et al.*, employed Incyte cDNA arrays and Affymetrix GeneChips to analyze gene expression changes induced by *tert*-butylhydroxyquinone treatment of human neuroblastoma cells (24). Cross-hybridization of the cDNA probes partially contributed to the discrepancies of the data generated by the two platforms. Data generated from the oligonucleotide microarrays were more reliable for interrogating changes in gene expression when compared to data from the cDNA microarrays. Kuo *et al.*, carried out the first large-scale analysis of reproducibility between spotted cDNA microarrays and Affymetrix GeneChips. The cDNA microarrays studies contained 9,703 cDNA probes whilst the Affymetrix HU6800 arrays contained 7,245 probe sets. A comparison of mRNA measurements of 2,895 sequence-matched genes in 56 cell lines from the standard National Cancer Institute (NCI 60) panel of 60 cancer cell lines revealed extremely poor correlation and discordance between the two platforms. An important limitation with this study, which negated the value of this data, was the lack of replicates. Additionally the data derived from separate microarray experiments that were carried out in two different laboratories using different materials and experimental protocols. Although identical cells lines were studied, the cells had been cultured independently and both the mRNA samples and hybridization targets were prepared separately. This revealed that even though differences existed between the technologies, the low correlations observed were likely also due in part to inherent systemic variations caused by the nature of the slide chemistries, target labeling, printing, and the scanning instrumentation (19).

Woo et al., compared the variability in measured gene expression levels associated with three types of microarray platforms, Affymetrix Mouse Genome Expression Set 430 GeneChips (MOE430A and MOE430B), spotted cDNA microarrays, and spotted oligonucleotide microarrays. Total liver RNA from four male mice, two each from inbred strains A/J and C57BL/6J were assayed on all three platforms. Variances associated with measurement error were observed to be comparable across all microarray platforms. The MOE430A GeneChips and cDNA arrays were found to have higher precision across technical replicates than the MOE430B GeneChips and oligonucleotide arrays. The Affymetrix platform revealed the greatest range in the magnitude of expression levels followed by the oligonucleotide arrays. Good concordance was observed in both the estimated expression level and statistical significance of common genes between the Affymetrix MOE430A GeneChip and the oligonucleotide arrays. Despite high precision, cDNA arrays showed poor concordance with other platforms (25).

Arrays containing one single long oligonucleotide probe for each gene have become a popular arrays format and are rapidly replacing the problematic cDNA-based arrays (24,26). Spotted long oligonucleotide arrays are created by deposition of long oligonucleotides 49-90 bases in length. Gene expression was analyzed in two dissimilar RNA samples, K652 an erythroleukemia cell line and Stratagene Universal Reference RNA (Stratagene, La Jolla, CA). Two platforms were compared, home-spotted oligonucleotide arrays fabricated from 70-mer probes, and Affymetrix. Each measurement was expressed as a pair of log-transformed differential expression (M) and total signal (A) values. Comparison of the expression measurements for 7344 genes that were represented in both platforms revealed strong correlations (0.8 - 0.9) between relative gene expression measurements. Preliminary analysis of the probe sequences used in that study suggested a high degree of overlap between the probes on both platforms (26).

Comparison of commercial platforms

The first comprehensive study of data generated from the three most widely used commercial platforms, was carried out by Tan *et al.*, (27) on PANC-1 cells, which have a pancreatic ductal phenotype. The authors examined gene expression measurements generated from identical RNA preparations on all platforms. This experimental design facilitated a direct comparison of all the array formats and eliminated experimental or biological variation that may have arisen from independent cell culture and RNA extraction. Microarrays from Affymetrix (U95Av.2 GeneChips, multiple 25-mer

oligonucleotide probe sets), Agilent (Human I, cDNA probes) and Amersham Biosciences (CodeLink Uniset Human I Bioarrays, 30-mer oligonucleotide probes) were hybridized with PANC-1 RNA collected from cells grown in serum-rich medium were compared with serum-depleted cells (24 hours following transfer to serum-free media). Three biological replicates were generated for each condition, and three experimental or technical replicates were produced for the first biological replicate. In order to compare the data, GenBank IDs were chosen over Unigene IDs for the comparison to eliminate variability owing to platform-dependent probes for different splice variants. In total 2,009 common genes were identified as being present on all three platforms and correlations in expression levels and comparisons for significant expression changes in this subset were calculated. This revealed considerable divergence across the three platforms. Unsupervised clustering and principle component analysis (PCA) suggested that the largest variation in measurements from the commercial platforms was attributable to the platforms themselves. Despite the fact that the Affymetrix and Amersham platforms were one color, short oligonucleotide platforms, no significant agreement was produced with these two platforms. The best level of agreement between the target gene sets was only 21% (between Amersham CodeLink and Agilent cDNA using a 2-fold induction and P value $P < 0.001$ criteria). Although gene sets overlapped to some extent across these platforms, the majority of genes identified as differentially expressed by each technology were uniquely identified by that technology. An encouraging observation was that when the 200 highest ranking down-regulated genes were classified according to biological themes rather than individual genes, stronger concordance was observed between the Agilent and Amersham Biosciences platforms. This suggests that enough genes within distinct gene ontology (GO) categories were detected by each platform to arrive at a common biological theme. The low concordance across the technologies can partly be attributed to the detection of distinct types or sets of alternatively spliced transcript variants.

A subsequent publication expanded on this study to examine six platforms, two cDNA based (Agilent and homemade), three short-oligonucleotide (Affymetrix, Amersham CodeLink and Mergen, San Leandro, CA) and one long oligonucleotide platform (Agilent) (28). The authors explored the hypothesis that gene expression profiles were biological rather than technological. They utilized the mouse *lacZ* model transgenic mutation test system (Muta™ Mouse) and RNA from whole lung tissue and an immortalized lung cell line (FE1) were compared on the various microarrays. Pearson product-moment correlations were performed on common genes across the six platforms, based on comparisons of UniGene identity for the reporter. The resulting correlation coefficients provided a simple measure of

agreement between platform pairs. The oligonucleotide arrays were highly correlated with each other, and moderately correlated with Agilent cDNA chips. Using condition tree analysis, the nature of the tissue was found to account for the measured differences in gene expression among microarray slides regardless of the particular platform. The FE1 and lung samples were observed to split on two main branches, with the Mergen spotted 30-mer oligonucleotide array (MO3) and academic cDNA platforms falling outside these clusters. Within these main branches, the biological replicates clustered within a platform type. Within tissue types the primary determinant of clustering was the platform, with biological replicates grouping together within a platform. The data support the hypothesis that for broad comparisons between two sample types, run on different commercial and homemade platforms at different times, the primary determinants of microarray gene expression changes result from true biological differences, rather than artifacts of platform choice. Using (significance analysis of microarrays) SAM, differences were found amongst the platforms in their ability to detect differential gene expression. The most differential gene expression was observed using CodeLink, Affymetrix and Agilent oligonucleotide based arrays.

Sequence-based probe matching

Figure 2 demonstrates the poor overlap between the Affymetrix, Agilent and Amersham CodeLink platforms using Unigene identifiers. Only 7171 probes are in common amongst all three platforms, using this reporter. Mecham et al., matched probes across different platforms on the basis of sequence rather than gene identifiers (29). Breast cancer cell line derived RNA was examined using Agilent cDNA and Affymetrix oligonucleotide microarray platforms to assess the advantage of this method. They noted that with regard to gene expression ratios and difference calls, cross-platform consistency was significantly improved by sequence-based matching. Sequence-based probe matching was found to produce more consistent results when comparing similar biological data sets obtained by different microarray platforms. This strategy allowed a more efficient transfer of classification of breast cancer samples between data sets produced by cDNA microarray and Affymetrix gene-chip platforms

The lower correlation shown by non-sequence-overlapping but Unigene-matched probes can be explained by several factors. It may reflect splice variants or 3'-5' degradation of microarray signals along genes (30,31). Unigene clusters assemble putative genes from cDNA clones using a variety of algorithms. However a subset of these clusters are incorrect (32). A significant fraction of these errors have been removed in updates of Unigene and by comparison to annotated human genome sequence.

However, the actual Unigene build utilized in some of the cross comparative studies may still contain several cases when two cDNA clones, are incorrectly listed as belonging to the same Unigene cluster. In one possible scenario, the cDNA feature on the spotted microarray and the Affymetrix probe, may thus measure the expression levels of two entirely different transcripts. The low correlation observed between non-overlapping probe sets are for the most part due to this reason, and this likely explains the discordance amongst cross-platform data. Clearly sequence-based probe matching is required to adequately compare data. Currently, this is not a trivial pursuit as the probe sequences for certain of the platforms represents proprietary information.

NORMALIZATION, FILTERING AND META ANALYSIS

Only after appropriate filtering, ratio and intensity data from different platforms can be compared and integrated. A key point is that good agreement between platforms was obtained in the studies of Barczak *et al.*, (26) after filtering out non-reproducible profiles using replicates from different experiments. Following image processing, the data generated for the arrayed genes must be normalized before differentially expressed genes can be identified. This process is necessary to adjust for experimental variables such as target labeling differences and variation in the detection sensitivity of fluorescent labels. Depending on the nature of the experimental design, there are different approaches for calculating normalization factors from the relative fluorescence intensities in the two scanned channels. A popular approach that could help future cross-platform studies is the use of spike in controls, where synthetic xenogenic sequences are added in increasing, but equimolar concentrations to the samples under study. The measured intensities for the added equimolar spike controls should be similar thereby facilitating cross-platform data standardization.

Meta analysis is a combination of techniques where the results of two or more independent studies are statistically combined to answer to a particular question of interest. The underlying rationale is that the combination provides a test with more power than the individual studies themselves. Rhodes *et al.*, utilized meta-analysis to combine multiple datasets from different studies (33). Meta-analysis tests score genes by reporting a P value that expresses the probability that the observed level of differential expression could have occurred by chance. Implementation of this model revealed that four prostate cancer gene expression datasets shared significantly similar results, independent of the method and technology used. This inter-study cross-validation approach identified a panel of genes that were

consistently and significantly dysregulated in prostate cancer. Analysis of these genes revealed a synchronous network of transcriptional regulation in the polyamine and purine biosynthesis pathways. This study established the first model for the evaluation, cross-validation, and comparison of multiple cancer profiling studies. Choi *et al.*, have established an alternative Meta-analysis procedure based on a Bayesian approach where the change of gene expression in cancer was expressed as 'effect size', a standardized index measuring the magnitude of a treatment or covariate effect. The effect sizes were combined to obtain the estimate of the overall mean. The statistical significance was determined by a permutation test extended to multiple datasets (34). Recently Zhou et al., described a 2nd-order expression analysis approach that addresses the challenge of cross-platform data by first extracting expression patterns as meta-information from each data set (1st-order expression analysis) and then analyzing them across multiple data sets. Using yeast as a model system they were able to identify genes of similar function yet without coexpression patterns. Furthermore they elucidated the co-operativities between transcription factors for regulatory network reconstruction by overcoming a key obstacle, namely the quantification of activities of transcription factors.

CONCLUSIONS

Major Findings

DNA microarray and oligonucleotide genechips have emerged as powerful tools for gene expression profiling on a genomic scale, and for establishing functional relationships between large number of genes involved in distinct cellular processes. In addition to detection of DNA copy-number and localization of transcription factor binding nucleic acid arrays have been extensively utilized for the detection of gene transcription. Several DNA chip technology formats have evolved and carefully designed studies have been performed to evaluate the interchangeability of data from the various platforms. The outlook for cross platform integration of data to date is more encouraging than the initial studies suggested. The discordance observed is attributable to the differences inherent in each of the respective platforms. The probes utilized in cDNA arrays may cause inaccurate expression measurements owing to overlap with related gene family members and the inability to discriminate between splice variants. In view of these studies, data from microarray analysis needs to be interpreted cautiously, and preferably using sequence matched probes.

Recent meta-analysis studies have been encouraging. They have provided a much-needed model for the evaluation, cross-validation and comparison of multiple cancer profiling studies.

Future Directions

As commercial manufacturers adopt standard DNA chip manufacturing practices, and arrays become clinical diagnostic tools, many of the quality control methods currently employed in the semiconductor industry will appear. This will result in higher quality, higher density arrays with greater sensitivity and reproducibility, facilitating a more robust analysis of subtle changes in cellular gene expression. However an underlying disadvantage with microarrays from a drug discovery perspective is that mRNA abundance in a cell often correlates poorly with the amount of protein synthesized, and proteins rather than mRNA transcripts are the major effector molecules in the cell. DNA microarrays have little utility in identifying physiologically relevant post-translational modifications of proteins, which influence protein function. Therefore there will remain a need to perform diverse assays in addition to transcriptome analysis.

REFERENCES

1. Kennedy T: **Managing the drug discovery/development interface.** *Drug Discovery Today* (1997); **2**(10): 436-44.
2. Marton M.J., DeRisi J.L., Bennett H.A., Iyer V.R., Meyer M.R., Roberts C., Stoughton R., Burchard J., Slade D., Dai H., Bassett Jr. D.E., Hartwell L.H., Brown P.O., Friend S.H. **Drug target validation and identification of secondary drug target effects using DNA microarrays.** *Nat. Med.* (1998) **4**:1293-1301.
3. Waring J.F., Ciurlionis R., Jolly R.A., Heindel M., Ulrich R.G. **Microarray analysis of hepatotoxins in vitro reveals a correlation between gene expression profiles and mechanisms of toxicity.** *Toxicol Letters* (2001) **120**:359-68.
4. Hamadeh HK, Amin RP, Paules RS, and Afshari CA: **An overview of toxicogenomics.** *Curr. Issues Mol. Biol* (2002) **4**(2): 45-56.
5. Johnson JA: **Drug target pharmacogenomics: an overview.** *Am. J. Pharmacogenomics.* (2001); **1**(4): 271-81.
6. Kruglyak L, and Nickerson DA: **Variation is the spice of life.** *Nature Genetics* (2001); **27**: 234 – 236.
7. van Someren EP, Wessels LF, Backer E, Reinders MJ. **Genetic network modeling.** *Pharmacogenomics* (2002); **3**:507-25.

8. Vilo J, Kivinen K. **Regulatory sequence analysis: application to the interpretation of gene expression.** *Eur Neuropsychopharmacol* (2001); 11:399-411
9. Hardiman G. **Microarray Platforms - Comparisons and Contrasts.** *Pharmacogenomics* (2004) 5(5), 487-502.
10. Stafford P and Liu P. **Microarray Technology Comparison, Statistical Analysis, and Experimental Design.** *Microarray Methods and Applications – Nuts and Bolts (DNA Press Eagleville PA)* 273 – 324 (2003). **••Overview of four separate microarray platforms with sample data, an examination of differences in spot morphology, quality control methodology, and a procedure for determining technical and biological variability and the corresponding resolution of the expression array.**
11. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* (1999). 286(5439):531-7. **••Seminal publication that unleashed the power of gene expression and microarray technologies, demonstrated accurate classification of hematologic malignancy-acute myeloid leukemia (AML) and acute lymphoid leukemia.**
12. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. **Molecular portraits of human breast tumours.** *Nature* (2000); 406(6797):747-52.
13. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, Petersen I. **Diversity of gene expression in adenocarcinoma of the lung.** *Proc Natl Acad Sci U S A.* 2001 98(24):13784-9.
14. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *Proc Natl Acad Sci U S A.* (2001) 98(24):13790-5.
15. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley

- PS, Bernards R, Friend SH. **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature.* (2002) **415**(6871):530-6.
16. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. **Gene-expression profiles predict survival of patients with lung adenocarcinoma.** *Nat Med.* 2002 **8**(8):816-24.
17. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP. **Accessing genetic information with high-density DNA arrays.** *Science* (1996) 274:610-4.
18. Schena M, Shalon D, Davis RW, Brown PO. **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* (1995) **270**:467-70.
19. Knight J: **When the chips are down.** *Nature* (2001) **410**:6831 **. *The enormous power of DNA microarray technology is also the source of many of its problems. Errors in handling cDNA clones and cross contamination problems are discussed in this article.*
20. Hughes T.R., Mao M., Jones A.R., Burchard J., Marton M.J., Shannon K.W., Lefkowitz S.M., Ziman M., Schelter J.M., Meyer M.R., Kobayashi S., Davis C., Dai H., He Y.D., Stephanians S.B., Cavet G., Walker W.L., West A., Coffey E., Shoemaker D.D., Stoughton R., Blanchard A.P., Friend S.H., Linsley, P.S. **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.** *Nat. Biotech* (2001) **19**:342-347.
21. Ramakrishnan,R., Dorris,D., Lublinsky,A., Nguyen,A., Domanus,M., Prokhorova,A., Gieser,L., Touma,E., Lockner,R., Tata,M., Zhu,X., Patterson,M., Shippy,R., Sendera,T.J. and Mazumder,A. **An assessment of Motorola CodeLink™ microarray performance for gene expression profiling applications.** *Nucleic Acids Res* (2002) 30, e30.
22. Kothapalli,R., Yoder,S.J., Mane,S. and Loughran,T.P. **Microarray results: how accurate are they?** *Bioinformatics* (2002) **3**, 22. ••*Critical evaluation of microarray data obtained from two different commercially available systems revealed several inconsistencies in the data obtained from the two different microarrays.*
23. Kuo,W.P., Jenssen,T.K., Butte,A.J., Ohno-Machado,L. and Kohane,I.S. **Analysis of matched mRNA measurements from two different microarray technologies.**

- Bioinformatics* (2002) **18**, 405–412. ••*A comparison of mRNA measurements 56 cancer cell lines using Stanford type cDNA microarrays and Affymetrix oligonucleotide microarrays.*
24. Li J, Pankratz M, Johnson JA. **Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays.** *Toxicol Sci.* (2002) 69:383-90. •*Data generated from oligonucleotide rather than cDNA microarrays were found to be more reliable for interrogating changes in gene expression.*
25. Woo Y, Affourtit J, Daigle S, Viale A, Johnson K, Naggert J, Churchill G. **A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms.** *J Biomol Tech.* (2004) 4:276-84. •*Good concordance was observed in both the estimated expression level and statistical significance of common genes between the Affymetrix MOE430A GeneChip and oligonucleotide arrays. Despite high precision, cDNA arrays showed poor concordance with other platforms.*
26. Barczak A, Rodriguez MW, Hanspers K, Koth LL, Tai YC, Bolstad BM, Speed TP, Erle DJ. **Spotted long oligonucleotide arrays for human gene expression analysis.** *Genome Res.* (2003) **13**:1775-85. ••*Comparison of Affymetrix and spotted 70-mer oligonucleotide arrays revealed strong correlations ($r = 0.8-0.9$) between relative gene expression measurements.*
27. Tan, Paul K., Downey, Thomas J., Spitznagel, Edward L., Jr, Xu, Pin, Fu, Dadin, Dimitrov, Dimiter S., Lempicki, Richard A., Raaka, Bruce M., Cam, Margaret C. **Evaluation of gene expression measurements from commercial microarray platforms.** *Nucl. Acids. Res.* **31** (2003): 5676-5684. ••*Comparison of gene expression measurements generated from identical RNA preparations that were obtained using three commercially available microarray platforms.*
28. Yauk CL, Berndt ML, Williams A, Douglas GR. **Comprehensive comparison of six microarray technologies.** *Nucleic Acids Res.* (2004) **32**(15):e124. ••*The authors explore and validate the hypothesis that gene expression profiles are biological rather than technological*
29. Mecham BH, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, Wetmore DZ, Mariani TJ, Kohane IS, Szallasi Z. **Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based**

- gene expression measurements.** *Nucleic Acids Res.* (2004) **32**(9):e74. ••Sequence-based probe matching produces more consistent results than identifier based probe matching.
- 30.** Wang,H., Hubbell,E., Hu,J.S., Mei,G., Cline,M., Lu,G., Clark,T., Siani-Rose,M.A., Ares,M., Kulp,D.C. and Haussler,D. **Gene structure-based splice variant deconvolution using a microarray platform.** *Bioinformatics* (2003) **19** (Suppl. 1), I315–I322.
- 31.** Auer,H., Lyianarachchi,S., Newsom,D., Klisovic,M.I., Marcucci,G., Kornacker,K. and Marcucci,U. **Chipping away at the chip bias: RNA degradation in microarray analysis.** *Nature Genet.*, (2003) **35**, 292–293.
- 32.** Burke,J., Davison,D. and Hide,W. **d2_cluster: a validated method for clustering EST and full-length cDNA sequences.** *Genome Res.*, (1999) **9**, 1135–1142.
- 33.** Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM. **Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer.** *Cancer Res.* (2002) 62:4427-33. ••First meta-analysis model which revealed that four prostate cancer gene expression datasets shared significantly similar results, independent of the method and technology used.
- 34.** Choi JK, Yu U, Kim S, Yoo OJ. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* (2003) 19 Suppl 1:i84-90.
- 35.** Zhou XJ, Kao MC, Huang H, Wong A, Nunez-Iglesias J, Primig M, Aparicio OM, Finch CE, Morgan TE, Wong WH. **Functional annotation and network reconstruction through cross-platform integration of microarray data.** *Nat Biotechnol.* (2005) **23**(2):238-43. ••Using yeast as a model system genes of similar function yet without coexpression patterns were identified.