Clinical Uses of Microarrays in Cancer Research

Carl Virtanen and James Woodgett

Samuel Lunenfeld Research Institute, Toronto and Microarray Centre, Clinical Genomics Centre,
University Health Network Toronto, Canada

Abstract

Perturbations in genes play a key role in the pathogenesis of cancer. Microarray-based technology is an ideal way in which to study the effects and interactions of multiple genes in cancer. There are many technologic challenges in running a microarray study including annotation of genes likely to be involved, designing the appropriate experiment and ensuring adequate quality assurance steps are implemented. Once data are normalized, they need to be analyzed and for this there are numerous software packages.

Keywords

cancer; microarray; annotation; experimental design; quality metrics; normalization; filtering; analysis

1. Introduction

Cancer is a genetic disease. Though this seems hardly worth mentioning nowadays it is, in fact, a relatively new idea. Until the 19th century, cancer was believed (due to Hippocrates) to be a disease caused by an excess of black bile. At the beginning of the 20th century the developmental biologist Theodor Boveri first suggested a linkage between tumour formation and improper chromosome segregation based on observations he made in sea urchins (1). Since then, the discovery of the structure of DNA in the 1950's (2), combined with genetic evidence of heritance of cancer and the discoveries of oncogenes (via the Rous sarcoma virus and the src gene in the 1970's (3)) and tumour suppressor genes e.g. retinoblastoma in the early 1980's (4), led the way to a more general understanding of the causal processes involved in neoplasia.

That cancers are caused by changes at the DNA level is now a given. However, few cancers, especially those of epithelial origin (so-called solid tumors), make it to an advanced stage by the mutation of a single gene. Instead, current thinking is that tumorigenesis is a haphazard stepwise process involving multiple short circuits of regulatory pathways, adaptation, avoidance of immune surveillance and finally gain of properties that allow disseminated growth – metastasis – typically the ultimate cause of death (5). Indeed, most cancers initiate with a single mutation conferring a growth advantage compared to neighboring cells (6).

The act of clonal expansion becomes associated with further mutations which endow subsequent cell populations with invasive and metastatic potential. The path is often convoluted and is also associated with cul-de-sacs and cell death. Unfortunately, the selective pressures and numbers of events that accrue over the many years of tumour development are statistically high enough that cancer is a leading cause of death. While many of the key contributors and cellular pathways that push a cell into neoplasia have been determined, there are clearly many more that are unknown.

Identification of all the factors involved in cancer should provide not only greater understanding of the biology of this collection of diseases, but also new therapeutic targets and diagnostic markers. A critical milestone on the road to this understanding (and this is true for many other diseases) was the completion of the rough and final drafts of the human genome (7). Indeed, one of the most important findings by the genome sequencing project was that the actual number of genes that are present in humans, roughly 25,000 by last estimate (8), is far less than the previously anticipated ~100,000 (9).

With ~25,000 defined genes in the human genome, there is now a hard limit on the space through which biological causation for neoplasia has to be searched. This is not to say the task will not be formidable since the relative paucity of human genes (compared with plants, for example) is countered by the complex processing of these genes through differential splicing, post translational modifications and exquisite control of function. There are probably at least 10 fold the number of human proteins compared to genes due to splice variants. For some proteins, there is already a plethora of knowledge as to their biological effect, either directly determined or gathered using a whole organism or systems biology approach. Unfortunately, for most genes and proteins, very little information exists regarding their biological activity. Also, each of the levels of information about a gene enriches our knowledge base but adds a level of complexity to any analysis, promoting informatics to the forefront if there is to be any hope of interpreting this deluge of data. Integrating all of this information in a cohesive and meaningful way is one of the goals of bioinformatics (10). The Cancer Genome Anatomy Project, the Human Cancer Genome Project, and the Cancer Genome Project represent three large scale programs attempting to coalesce a wide range of cancer related genomic data into a single resource (11) (*see Note 1*).

2. Microarray Technology

Microarrays leverage biological information with bioinformatics knowledge to bring new insights to many biological systems. They have been applied most intensively to the field of cancer research. The basic premise behind a microarray is that thousands of fragments of DNA (the probes) representing various genes are attached to the surface of an inert material. Reverse transcribed messenger RNA labeled with a fluorescent dye from an experimental sample (the target) is then co-hybridized on the microarray. After removal of non-selectively

¹The three cancer-related programs described involve centers in the UK, USA and Brazil (11). The extensive molecular analysis of various cancers from the programs has DNA sequencing as the common link. The output from this work involves the sequencing of cDNA libraries, SAGE libraries (SAGE – Serial Analysis of Gene Expression is another form of microarrays), and mutation testing of genomic DNA from various tumors. Without a planned and informatics-based approach the accumulated data would not be available to the wider research community, and without this integrated approach the effectiveness of data mining would be diminished.

bound fluorescent material, the microarray is scanned at a high resolution to quantify the amount of fluorescent signal over the surface. Since the location of the gene probes is predetermined, the relative or absolute amount of RNA for each of the genes on the array can be calculated. A microarray therefore measures the levels of mRNA transcripts in a sample. Since thousands of gene fragments can be located on an array, it can provide a genome-wide view of gene expression in cancer.

The two most popular microarray technologies employ the use of either double stranded cDNA probes spotted onto slide surfaces, pioneered by the work of Pat Brown's laboratory (12), or single stranded oligonucleotide probes. Oligonucleotide microarrays are made either by depositing pre-synthesized oligonucleotides directly onto a slide surface mechanically, e.g. using ink jet technology (www.agilent.com), or assembled base by base using photolithography in a process somewhat analogous to semiconductor etching using photoactivatable nucleotides and multiple masks. The photolithographic technique was developed by Stephen Fodor in the early 1990's (13,14) and is the basis behind the microarray platform used in the Affymetrix GeneChips (www.affymetrix.com).

A comparative approach is typically taken to identify genes important in neoplasia. For example, tumorous tissue may be compared to normal tissue, or tumors of different pathological types or grades may be looked at. The relative fold difference between the categories is determined for each particular gene being measured on the microarrays, and significant groups of genes can be aggregated based on similar behavior and looked at either as a whole or individually. Morphologically indistinct tumors may actually belong to clinically and biologically distinctive categories. For example, Her2/neu/ERBB2 positive status in breast cancer is widely used to assess trastuzumab (Herceptin) treatment, which is, in fact, the target of the drug (15). Microarrays are now routinely used in the hope of finding such biomarkers.

Microarray expression profiling of cancers has now been used in a wide variety of settings. Correlation to pathological subtype was first shown by Golub et al. (16) in comparing acute myelogenous leukemia and acute lymphocytic leukemia. In malignant lymphomas, Alizadeh et al. (17) were the first group to show a correlation to therapeutic outcome with gene expression profiles in B-cell lymphoma. A distinct signature for metastatic tumors was found by Ramaswamy et al. (18) in primary lung adenocarcinomas. Meta-analysis of gene expression data sets has been used to find a generalized signature for proliferation (19). Prediction of patient outcome was demonstrated successfully in breast cancer by grouping genes together into "meta" genes (20) (*see* ^{Note 2}). A commercially available diagnostic based on microarray data for the recurrence of breast cancer in tamoxifen-treated patients is now available (21). Extending the concept of finding gene signatures, grouping genes according to the pathways they are involved in has also proved successful (22). In a recent

²Meta analysis is a well accepted approach in evidence based medical practices and works through statistically combining the results of previous experiments or studies. In the analysis of microarray studies, it is proposed in (20) to optimize the data generated by combining through metagenes i.e. the aggregate patterns of variation of subsets of potentially related genes. The example given is breast cancer. Predicting prognosis in this cancer is a major clinical challenge. Although staging through assessment of lymph node involvement, estrogen receptor status, age and tumor size are major prognostic factors, it still is not possible to predict consistently the outcomes. Using various statistical approaches a combination of clinical and genomic (metagene) data is likely to be used increasingly to guide decision making.

study, Glinsky et al (23) combined data from mouse and human prostate tumor samples to derive an 11 gene signature that consistently holds a stem-cell type expression profile predicting poor outcome across a variety of cancer types. Microarrays have also been used extensively to profile cancers of the prostate (24,25), ovaries (26), kidney (27), breast (28,29), and lung (30,31), along with a number of other neoplasms (32–34).

For the remainder of this chapter, we discuss the various factors involved in the design and interpretation of a typical microarray project focused on a clinical application. Fig. 1 illustrates a general flow chart of the process. The approach will necessarily be a general one, and the merits of various options at each stage will be discussed. We will focus on techniques specific to cDNA arrays, though comparable methodologies exist for the oligonucleotide arrays. The “Tumor Analysis Best Practices Working Group” has published an extensive review of microarray techniques focused on Affymetrix arrays (35).

3. Design and Interpretation

3.1 Annotation

One of the most important informatic steps that need to be done before a single microarray has even been hybridized is annotation of the array platform. Out-of-date annotation can severely impact interpretation of the dataset, as Dai et al. (36) have shown. The basic information as to what a spot on an array is, either EST accession ID or sequence information, is the first starting point. However, this information is somewhat meaningless unless placed in an appropriate context with higher levels of meta-annotation. Unfortunately, all of the meta-annotations one might need to use are in disparate public and private databases. These databases are themselves dynamic in nature and therefore constantly changing at different rates, making contemporary annotation a non-trivial task (10). The SOURCE database attempts to pull together various annotations for users of microarrays, and is also updated on a regular basis (37), however the information it contains is only basic. As a consequence many producers of microarrays now invest substantial effort in maintaining array annotation. At the University Health Network Microarray Centre in Toronto, for example, all databases and annotations on arrays are fully re-annotated quarterly and this seems to be a reasonable schedule.

A database, such as MySQL (www.mysql.org), installed on a server running Linux (<http://distrocenter.linux.com/platforms>) is a good starting point for pulling data annotations together. There is no “one size fits all” approach to annotation and much will depend on a researchers’ particular interest. The starting point usually involves retrieving sequence for each of the elements on an array and finding their genomic position using BLAT (38). For cDNA probes, searching for repeats in sequences using REPEATMASKER (39) is important, since these can confound results if not taken into account. This should not be an issue for oligonucleotide arrays as the design of the representative oligonucleotide includes avoidance of non-unique sequences.

Once a gene’s position in the genome is known, further annotation can usually be derived by searching through the UCSC “knowngene” data table, available as a free download (<http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/knownGene.txt.gz>). From this table a

reference to kgXref (known gene cross references table, also available at UCSC) can be used to find Uniprot (40) and NCBI REFSEQ protein ids (41). With the REFSEQ id, a link can be made to the UNIGENE (42) database, available for free by downloading and searching directly (<ftp://ftp.ncbi.nih.gov/repository/UniGene>). Unigene is a database of non-redundant gene ids assembled by clustering GenBank sequences. Although commonly used, it can be unreliable (43) and it is therefore recommended that links to the Entrez Gene ids (44) be made. This can be done by downloading and searching the table gene2accession found at NCBI (<ftp://ftp.ncbi.nih.gov/gene/DATA/gene2accession.gz>). With Entrez Gene ids, links to OMIM can be made through the table mim2gene (available at: <http://omim.org/downloads>). OMIM (45) is a curated catalog of human genes related to genetic disorders. Linking genes to actual published data found in PubMed (42) can be made using the gene2pubmed table (<ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz>). Finally, Gene Ontology (46) annotations can be derived using the gene2go table (<ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz>). Gene Ontology (GO) classifications are quite useful in pulling together higher level concepts surrounding a gene product. For example, a gene product may be part of the parent GO term “signal transduction”. Further annotation is limited only by how deep a researcher wants to go, as many hundreds of specialized databases exist. The journal Nucleic Acids Research (<http://nar.oxfordjournals.org>) publishes yearly an issue devoted to the most popular online molecular biology related databases.

3.2 Experimental Design

3.2.1. General considerations—Suffice to say the experimental design is the most important stage for any scientist wishing to use microarrays in their clinical research. A good experimental design relies on a good hypothesis and appropriate scaling of the power of the experiment to the type and availability of the samples to be profiled. For example, one hypothesis might be that there is a difference in gene expression between normal and tumor samples from the same group of patients. Or, that there is a difference between two types of cancer present in the same tissue types in different patients (lung squamous cell carcinoma versus lung adenocarcinoma). The experimental conditions should then be set up according to the general statistical principles which all experiments are guided by. Care must be taken to balance the sample population for extraneous factors: smokers and non-smokers, age, sex, tumor stage, and so on. These types of factors may all effect gene expression and an obvious bias in one of the experiments hypothesis driven categories will be confounded if not balanced correctly. Furthermore, it is now well known that variation between studies of the same type of sample are influenced by the samples themselves and by human related factors (47–49). Therefore, one must balance between experimental groups the technicians responsible for processing and handling samples, the slide lots used during the experiment, and possibly even the days during which RNA extraction and hybridization are being performed.

In two color experiments in which two RNA samples are labeled with two fluorescent dyes with distinct emission spectra, one channel is usually designated for the experimental channel and one for a suitable control channel. There are two generally accepted methodologies for choice of control channel. In the first, the control channel is not a control *per se*, but is instead one of the experimental samples from one of the classes being

examined e.g. tumor type A versus tumor type B. In the “balanced block design” (50) each sample from the two classes is matched up to a member of the other class and the two are hybridized on the same array. In the “loop design” (51), each sample is split into two subsamples and each subsample is used to connect arrays together in a loop pattern. In the other methodology, a reference sample is used as a control on each chip and comparisons between classes is accomplished by comparing ratios of experimental to control. This is by far the most common design for two color cDNA experiments. It has the disadvantage that more microarray chips are needed compared to a block design since each sample has to be run on a separate chip. But one advantage is that multiple experiments (such as future studies that are done as follow up) using the same reference sample can be compared directly. Also, in cases where the object is to find novel classes based on expression patterns, only the reference design can be used. Finally, a reference design reduces the need to perform dye-flips (52) to some extent, because the reference samples are all labeled with the same dye between which comparisons are being made. Dye-specific bias is also less of an issue when using indirect labeling (53).

It is worth mentioning the value of profiling normal tissue types when deciding on an experimental setup. It is quite easy to find groups of genes which vary across normal tissue types (31,54). This is likely due to the heterogeneous nature of pathological samples, where a bias in tissue type, cells at a particular phase of the cell cycle or hormonal status, or even inter-individual expression signatures may cause certain genes to change between samples. Obviously it is of value to remove such genes from further analysis during the filtering stage.

3.2.2. Sample numbers—Once a general experimental design has been decided upon, the next important choice will be the number of samples to be profiled. Unfortunately, there is no easy answer for determining sample size. The multiplex nature of microarray studies, many thousands of genes versus small numbers of samples, makes power calculations that are traditionally used in setting up clinical trials difficult to apply, and often dependent on the final analysis regimen. A point worth mentioning is that choosing a sample size based on traditional estimates of statistical power is fine when all that is necessary is to make an estimate on the confidence of the error term, but this can yield different numbers than what is necessary for assessing the power of a classification algorithm. Unfortunately, all of this is compounded by the fact that final analysis of microarray data is usually more exploratory and unknown at the start of a study. More often than not, the choice of sample size will be governed by the amount of pathological samples that are actually available, or the cost limitations. Still, there are some methods available for determining appropriate sample size (50,55–57).

Tibshirani’s method (56) of sample size estimation is a good choice and does not assume equal variances or independence of genes on an array. The fact that it is based on estimating false discovery rates, a common methodology for finding genes that vary between groups, makes it even more suitable for use. The basic premise behind Tibshirani’s method is to start with the output from a typical permutation based analysis routine and then estimate false discovery and false negative rates for different sample sizes. By way of example, to find a two-fold difference between two groups of 10 samples e.g. normal versus tumor, each on an

array with 1,000 genes the sample size must be increased to roughly 30 samples in each group to get optimal results.

3.3. Quality Metrics

Measuring the quality of slides after hybridization and scanning is an important part of analysis. Visually inspecting images for gross defects should always be done. However, more robust methods are recommended. The best software for assessing data quality is found in the LIMMA package which is part of the Bioconductor (58) framework for the R (www.R-project.org) statistical language. We currently use Bioconductor version 1.8 with R version 2.3.1. Although there is not one particular number or function which will definitively point to a chip being of “pass” or “fail” quality, the joint use of these functions allows for an overall assessment to be made and for outliers i.e. bad slides, to be identified.

Listing 1 shows a typical R script that can be used to assess raw data files after quantification. The script will output a series of quality related graphs into the working directory. The data files in this case were all quantified in ArrayVision (www.imagingresearch.com/products/ARV.asp).

Line 1	Loads the appropriate LIMMA module from Bioconductor into R.
Lines 2 – 7	Load the data (files ending in “.txt” and starting with a number from 1 to 9) in the current working directory and associate the correct layout of the array with the data itself.
Line 8	Copies the data into a new matrix called “MA” and can be used to change normalization method (method=“none” in this case) or background subtraction choice (bc.method=“none”). The choice of performing background subtraction is not always obvious. The underlying hypothesis of background subtraction is that the intensity of a spot on a microarray is the combined value of the background intensity around the DNA spot plus the foreground measurement of intensity within the limits of the actual spot. This hypothesis is somewhat flawed though, as negative control spots i.e. water or DMSO spots with no DNA in them, often are of lower intensity than the surrounding background (an effect termed “ghosting”). As such, background subtraction can often introduce rather serious artefacts in the data. However, it can be useful if the slide surface consistencies are non-uniform, e.g. if fluid was not distributed evenly during hybridization. If considering background subtraction, it is recommended that the procedures in this section be repeated with background subtraction applied to see what effect it may have on the data.
Lines 9, 10	Are used to set a filename and size of graph that will serve as the output from the next line.
Line 11	Shows a density histogram is generated. Fig. 2a shows a typical density histogram for a series of 20 slides. The red and green lines both overlap substantially and the curves are roughly symmetrical around the peaks. Compare this to a series of slides (Fig. 2b) where there was a problem during RNA extraction and labeling. The histograms are very inconsistent and a problem is immediately assessed. In this case the entire experiment should be repeated.
Lines 12–15	Plot pseudo-colored spatial images of the data for both channels. Separate files for both channels and the foreground (spot) and background intensities are generated. Figs. 3a and 3b show a typical output for the green channel foreground and background signal of two slides with acceptable quality. Notice that there is a lot of consistency across the slide surface. Contrast these two images with Figs. 3c and 3d, which obviously display a problem that occurred during hybridization. These slides will probably need to be repeated.
Lines 16 – 19	Setup and output a boxplot of all the current slides into a file. A boxplot shows the overall mean (horizontal line in the centre of the box) along with the interquartile range (the box) and outliers (circles). Once again, relative consistency is appropriate, and a box plot can allow for an outlier slide to be easily determined. In Fig. 4 it is easy to see that slide number 5 is different from the other slides, and closer examination of this slide would be necessary to determine if it should be accepted or rejected for the study.
Lines 20 – 33	Iterate through all the arrays loaded into the dataset and output MA plots (6 plots per page as determined on line 28). An MA plot shows any intensity dependant effects in the raw data where $M = \log_2(R/G)$ and $A = (0.5) * \log_2(R * G)$. The x-axis represents a log ratio of equal intensities in both channels. The majority of data in a typical microarray dataset should spread out roughly along this line. Fig. 5a shows an MA plot from a slide with a good distribution of data. Fig. 5b, on the other hand, clearly has an intensity-dependant issue that goes beyond that which can be corrected later by normalization.

3.4 Normalization and Filtering

3.4.1 Normalizing data—Once assured of some measure of slide quality, the data will first need to be normalized before valid comparisons between sample subjects are made. This is because differences in labeling efficiencies, amounts of starting RNA materials, fluctuating hybridization conditions and other biases may be present from slide to slide. The box plots in Fig. 4 (aside from the problem slide) show that the distribution for individual

slides is not quite the same, even though they are expected to be. This platform related variation should be accounted for before looking for biological variation. The most commonly applied normalization is a “per-chip” normalization that balances out systematic biases within a slide itself. It is well known that there are intensity-dependant issues in microarray ratio measurements (with low intensity measured log ratios typically deviating from zero) and that these furthermore vary across pin groups on spotted slides (59,60). Therefore, typically a locally weighted linear regression curve (loess) is fit on a pin to pin basis to an MA plot of the data and values in the control channel are adjusted up or down to correct for deviation from this line. Fig. 6a shows an MA plot of a slide before loess normalization and Fig. 6b shows the data after normalization. Further normalization usually consists of normalizing each genes’ log ratio to zero across chips to account for genes that are either highly or lowly expressed across all samples. It is usually easier to visualize genes that are changing between samples after such a normalization. Normalization is a standard procedure and can be performed in any microarray data analysis package. For example, to loess normalize data in R, simply change line 8 in Listing 1 to: `MA<-normalizeWithinArrays(RG, method="loess", bc.method="none")`.

3.4.2 Filtering unwanted genes—Filtering is usually the next step in microarray data analysis. Filtering is an important step because it removes genes from further examination that are not of interest. Genes with very low intensity are typically removed because there tends to be much higher variability with their measurements. A good rule of thumb is to flag or remove those genes with intensities that are less than two times the average background intensity before normalization. Ranking genes in percentiles is another way to determine an appropriate cutoff. For example one might wish to remove those genes in the lower 10th percentile of overall intensity measure. Filtering of genes can also be done when there is *a priori* information of spots which may be problematic. On spotted cDNA slides, there are sometimes problems related to production e.g. during amplification of spotted material, which may warrant some spots being removed. It is recommended to remove spots which have a high percentage of repetitive elements in them. Repeat elements such as the human ALU repeat are present in upwards of 10% of the human genome (61), and these are frequently present in the non-coding ends of spotted cDNA’s on microarrays. Since we would expect an inordinate amount of cross hybridization in sequences containing a large amount of repeat (say, greater than 33%) they should be filtered out. Finally, as mentioned earlier, there are many genes which vary naturally to a great extent between individuals or samples. If a study is done with an appropriate number of normal samples, one can first look for genes that vary between individuals and filter them out from subsequent analysis when looking at differences between tumor types.

3.5 Analysis

3.5.1 Differentially expressed genes—Analyzing microarray data usually first takes the form of finding genes which are differentially expressed, either between experimental and control channels on chips or between samples. Finding differentially expressed genes is usually done first, both to find genes of interest and to further filter data before application of more sophisticated data mining techniques such as clustering. When attempting to find genes that are over or under expressed one typically chooses a threshold, such as a 2-fold

difference. This number was originally determined by concordance analysis for one data set (62), but has become a guideline criteria now used in many different analysis. More sophisticated measures, such as using a Z-score (63) to estimate fold changes in an intensity dependant manner can also be used. A one-sided t-test can be performed if replicates were done for each sample. This must be multiple-test corrected (see next paragraph). Once a threshold is decided on, the usual course is to apply that threshold to being found across a percentage of the samples. For example, if 100 samples were obtained and profiled, one might choose to only look at genes that have at least a 2-fold difference in 33% of them. Choosing a 2-fold change level will undoubtedly lead to removal of true differences that are lower, but it will still allow for finding the less conservative changers. This is a testament to microarrays being a screening technology where one usually is looking for the “low hanging fruit”.

3.5.2 Categorizing samples—Attempting to categorize samples can be done in one of two (or both) ways. Unsupervised methods are exploratory in nature. Agglomerative hierarchical clustering is one such technique. In this method, two genes that have the most similar expression profiles across experiments, based on a similarity measure such as their Pearson Correlation, are found. The average is taken between these two genes and then a new gene most similar to this “average gene” is found in the rest of the set. The process is iterated and a tree type diagram can be built up (Fig. 7). The length of tree branches is related to the degree of similarity between adjoining groups. Individuals sample are similarly clustered according to their nearest neighbors. A cluster diagram allows one to explore those categories of genes or samples which are nearest to one another and hypothesis regarding biological meaning can be generated. For example, it is often hypothesized that genes that are closely clustered together are related to one another in some manner, such as belonging to the same molecular pathway. Care should be exercised in interpreting clusters since clustered genes can rotate around a branch and hence distance between genes on the edges of neighbouring clusters is much larger than genes within a cluster. Cluster diagrams were first applied by Eisen (64) to microarray data and are now very common. Some clustering methods, such as those with bootstrapping, aim to assess the statistical significance of tree branch locations (65). Principal components analysis is another unsupervised method to attempt to find biologically meaningful groups in microarray results (66). By projecting the multiple dimensioned space of a microarray dataset onto axis in lower dimensional space, those samples or genes most closely related will be grouped together.

3.5.3 Data mining—Supervised methods of data mining are used when *a priori* information regarding categories in the data exist. The most common method is simply to perform a t-test or an ANOVA (67). However, when performing such tests it is necessary to correct for multiple testing of hypothesis. With very many genes on an array, performing multiple t-tests or ANOVA's will result in many false positive results. For example, at the 5% significance level for an array with 10,000 genes, one would expect 500 genes to be falsely identified as varying. Therefore, a correction to the test statistics must be made. The simplest is a Bonferonni correction, where the significance p-value is simply weighted by dividing its value by the number of tests. This is usually too strict though, and typically

something such as a Benjamini and Hochberg (68) false discovery rate multiple test correction is used.

A popular supervised mining technique is “significance analysis of microarrays” (SAM), first developed by Tibshirani (69). SAM is similar to a t-test but uses a term called a “d-value” instead. If there are two categories of samples, I and U, then let $\bar{x}_I$ and $\bar{x}_U$ be the mean expression measurements of gene i in the two categories, respectively. The test statistic is then:

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_o}$$

where:

$$s(i) = \left[\left(\frac{1/n_1 + 1/n_2}{n_1 + n_2 - 2} \right) \left\{ \sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_n [x_n(i) - \bar{x}_U(i)]^2 \right\} \right]^{1/2}$$

n_1 and n_2 are the numbers of expression values in each category and Σ_m and Σ_n are summations of the differences of expression measurements from the mean expression measurement in categories I and U, respectively. The idea is then to estimate for each gene an expected value of d by randomly shuffling across all samples the expression values, recalculating d and then taking the average value after a fixed number of iterations. When plotting the observed d -values versus the expected d -values, any gene that deviates far from the “observed=expected” line is likely to be significant. A tuning parameter can be chosen which sets the upper and lower bounds on how far a deviation has to be before being called significant.

Other popular supervised data mining techniques include “gene-shaving”(70), which can be also used in an unsupervised fashion, and “weighted-voting”, which was first introduced by Golub (16). Of course, there are many more methods for mining data and it is often overwhelming trying to decide on a course of analysis. A good approach is to run multiple analyses and look for common pools of genes that are significant across all methods.

3.5.4 Validation—Validation is crucial at the end of the analysis pipeline. Although every attempt is made to control for extraneous factors affecting gene expression measurements, including statistical techniques to minimize false positive results, nevertheless there will be mistakes. Currently the best methodology to validate results is RT-PCR (71). Microarrays have a low dynamic range of measured gene expression and tend to underestimate changes, whereas RT-PCR has a high dynamic range. Hence, validation is used to assess overall observed trends rather than duplicate results.

3.6 Higher Level Analysis

With lists of significantly co-regulated genes in hand, the next step is to place them into a biologically meaningful context. Comparing GO annotations in lists of genes is a good starting place. The basic premise here is that there may be over-representations of particular

traits in one list of genes compared to another and this may represent unique biological features. To do this, an appropriate statistical measure to compare gene lists must be used, such as a Chi-square or Fisher's Exact Test. Of course, as with previous statistical tests, a correction for multiple hypothesis testing must be made, and methodologies for this exist (72,73). One can also combine GO annotation with expression data during the data mining phase of analysis (22).

It is well known that genomic instability, a distinctive trait of cancer, leads to increased or decreased copy numbers of certain genes. That this will have some effect on gene expression is probable; whether this effect is linear is unknown. However, by integrating expression data with array comparative genomic hybridization (CGH) data, this question can be at least partially answered. Bussey et al. (74) looked at copy number and expression in the NCI-60 panel of human cancer cell lines and found a positive correlation between the two. Somewhat related to this, integrating genomic information for a particular region with mRNA expression and protein data allowed Mootha et al. (75) to find a candidate gene for cytochrome c oxidase deficiency.

Reconstructing genetic networks is another area that has been well explored. The underlying hypothesis is that correlation in expression can be related to co-regulation and hence involvement in similar pathways. The representational analysis presented earlier using GO functional categories is one way to dissect out these interactions. Combining protein-protein interaction data with gene expression measurements is another method that has been used (76). Expression data combined with phylogenetic conservation has also proven successful (77). Indeed, there is some evidence that genes that are co-expressed are often clustered together on the genome, hence bringing issues of selective pressure into play (78).

4. Conclusions

There is no single approach in using microarrays to study neoplasia. Instead, there are a number of methodologies that can be chosen depending on the questions being asked. Here, we have presented a general methodology for the design and analysis of experiments using microarrays. With careful selection and appropriate controls, new insight into the functional underpinnings of cancer can be studied leading to better diagnostics and possible future treatment. The most critical element in employing microarray technology is selection of the most appropriate experimental design that complements the sample set available. It is far simpler to alter analysis strategies of a good dataset than to try to remediate deficiencies in a poor one.

Abbreviations

GO	gene ontology
RT-PCR	reverse transcriptase PCR

References

1. Balmain A. Cancer Genetics: From Boveri and Mendel to Microarrays. *Nat Rev Cancer*. 2001; 1:77–82. [PubMed: 11900254]

2. Watson JD, Crick FH. Genetical implications of the structure of deoxyribonucleic acid. *Nature*. 1953; 171:964–967. [PubMed: 13063483]
3. Stehelin D, Varmus HE, Bishop JM, Vogt PK. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*. 1976; 260:170–173. [PubMed: 176594]
4. Friend SH, Bernards R, Rogelj S, Weinberg RA, Rapaport JM, Albert DM, et al. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature*. 1986; 323:643–646. [PubMed: 2877398]
5. Hanahan D, Weinberg RA. The Hallmarks of Cancer. *Cell*. 2000; 100:57–70. [PubMed: 10647931]
6. Nowell PC. Tumor progression: a brief historical perspective. *Semin Cancer Biol*. 2002; 12:261–266. [PubMed: 12147207]
7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409:860–921. [PubMed: 11237011]
8. International Human Genome Sequencing Consortium. Human Genome Sequencing C. Finishing the euchromatic sequence of the human genome. *Nature*. 2004; 431:931–945. [PubMed: 15496913]
9. Pennisi E. Bioinformatics: Gene Counters Struggle to Get the Right Answer. *Science*. 2003; 301:1040–1041. [PubMed: 12933991]
10. Stein LD. Integrating biological databases. *Nat Rev Genet*. 2003; 4:337–345. [PubMed: 12728276]
11. Strausberg RL, Simpson AJG, Wooster R. Sequence-Based Cancer Genomics: Progress, Lessons and Opportunities. *Nat Rev Genet*. 2003; 4:409–418. [PubMed: 12776211]
12. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995; 270:467–470. [PubMed: 7569999]
13. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science*. 1991; 251:767–773. [PubMed: 1990438]
14. Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL. Multiplexed biochemical assays with biological chips. *Nature*. 1993; 364:555–556. [PubMed: 7687751]
15. Ludwig JA, Weinstein JN. Biomarkers in Cancer Staging, Prognosis and Treatment Selection. *Nat Rev Cancer*. 2005; 5:845–856. [PubMed: 16239904]
16. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*. 1999; 286:531–537. [PubMed: 10521349]
17. Alizadeh AA. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000; 403:503–511. [PubMed: 10676951]
18. Ramaswamy S, Ross K, Lander E, Golub T. A molecular signature of metastasis in primary solid tumors. *Nat Genet*. 2003; 33:49–54. [PubMed: 12469122]
19. Rhodes DR, Chinnaiyan AM. Integrative analysis of the cancer transcriptome. *Nat Genet*. 2005; 37(Suppl):S31–37. [PubMed: 15920528]
20. Pittman J, Huang E, Dressman H, Horng CF, Cheng SH, Tsou MH, et al. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc Natl Acad Sci U S A*. 2004; 101:8431–8436. [PubMed: 15152076]
21. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *N Engl J Med*. 2004; 351:2817–2826. [PubMed: 15591335]
22. Segal E, Friedman N, Kaminski N, Regev A, Koller D. From signatures to models: understanding cancer using microarrays. *Nat Genet*. 2005; 37(Suppl):S38–45. [PubMed: 15920529]
23. Glinsky GV, Berezovska O, Glinskii AB. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *J Clin Invest*. 2005; 115:1503–1521. [PubMed: 15931389]
24. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, et al. Delineation of prognostic biomarkers in prostate cancer. *Nature*. 2001; 412:822–826. [PubMed: 11518967]
25. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A*. 2004; 101:811–816. [PubMed: 14711987]

26. Schaner ME, Ross DT, Ciaravino G, Sorlie T, Troyanskaya O, Diehn M, et al. Gene expression patterns in ovarian carcinomas. *Mol Biol Cell*. 2003; 14:4376–4386. [PubMed: 12960427]
27. Vasselli JR, Shih JH, Iyengar SR, Maranchie J, Riss J, Worrell R, et al. Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor. *Proc Natl Acad Sci U S A*. 2003; 100:6958–6963. [PubMed: 12777628]
28. Sorlie T, Perou C, Brown P, Botstein D, Borresen-Dale A. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*. 2001; 98:10869–10874. [PubMed: 11553815]
29. Sotiriou C. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci USA*. 2003; 100:10393–10398. [PubMed: 12917485]
30. Garber ME. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci USA*. 2001; 98:13784–13789. [PubMed: 11707590]
31. Jones MH, Virtanen C, Honjoh D, Miyoshi T, Satoh Y, Okumura S, et al. Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet*. 2004; 363:775–781. [PubMed: 15016488]
32. West RB, van de Rijn M. The role of microarray technologies in the study of soft tissue tumours. *Histopathology*. 2006; 48:22–31. [PubMed: 16359534]
33. Tinker AV, Boussioutas A, Bowtell DDL. The challenges of gene expression microarrays for the study of human cancer. *Cancer Cell*. 2006; 9:333–339. [PubMed: 16697954]
34. Wadlow R, Ramaswamy S. DNA microarrays in clinical cancer research. *Curr Mol Med*. 2005; 5:111–120. [PubMed: 15720274]
35. The Tumor Analysis Best Practices Working Group. Expression Profiling - Best Practices for Data Generation and Interpretation in Clinical Trials. *Nat Rev Genet*. 2004; 5:229–237. [PubMed: 14970825]
36. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucl Acids Res*. 2005; 33:e175. [PubMed: 16284200]
37. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, et al. Source: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucl Acids Res*. 2003; 31:219–223. [PubMed: 12519986]
38. Kent WJ. BLAT---The BLAST-Like Alignment Tool. *Genome Res*. 2002; 12:656–664. [PubMed: 11932250]
39. Smit, A., Hubley, R., Green, P. RepeatMasker Open 3.0. 1996–2004. www.repeatmasker.org
40. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The Universal Protein Resource (UniProt). *Nucl Acids Res*. 2005; 33:D154–159. [PubMed: 15608167]
41. Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet*. 2000; 16:44–47. [PubMed: 10637631]
42. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. *Nucl Acids Res*. 2006; 34:D173–180. [PubMed: 16381840]
43. Murray CG, Larsson TP, Hill T, Bjorklund R, Fredriksson R, Schioth HB. Evaluation of EST-data using the genome assembly. *Biochem Biophys Res Commun*. 2005; 331:1566–1576. [PubMed: 15883052]
44. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucl Acids Res*. 2005; 33:D54–58. [PubMed: 15608257]
45. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucl Acids Res*. 2005; 33:D514–517. [PubMed: 15608251]
46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000; 25:25–29. [PubMed: 10802651]
47. Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, et al. Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods*. 2005; 2:351–356. [PubMed: 15846362]

48. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J. Independence and reproducibility across microarray platforms. *Nat Methods*. 2005; 2:337–344. [PubMed: 15846360]
49. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, et al. Multiple-laboratory comparison of microarray platforms. *Nat Methods*. 2005; 2:345–350. [PubMed: 15846361]
50. Simon RM, Dobbin K. Experimental design of DNA microarray experiments. *Biotechniques*. 2003; (Suppl):16–21.
51. Kerr MK, Churchill GA. Experimental design for gene expression microarrays. *Biostatistics*. 2001; 2:183–201. [PubMed: 12933549]
52. Dobbin K, Shih JH, Simon R. Questions and Answers on Design of Dual-Label Microarrays for Identifying Differentially Expressed Genes. *J Natl Cancer Inst*. 2003; 95:1362–1369. [PubMed: 13130111]
53. Cox WG, Singer VL. Fluorescent DNA hybridization probe preparation using amine modification and reactive dye coupling. *Biotechniques*. 2004; 36:114–122. [PubMed: 14740493]
54. Virtanen C, Ishikawa Y, Honjoh D, Kimura M, Shimane M, Miyoshi T, et al. Integrated classification of lung tumors and cell lines by expression profiling. *Proc Natl Acad Sci USA*. 2002; 99:12357–12362. [PubMed: 12218176]
55. Mukherjee S, Tamayo P, Rogers S, Rifkin R, Engle A, Campbell C, et al. Estimating dataset size requirements for classifying DNA microarray data. *J Comput Biol*. 2003; 10:119–142. [PubMed: 12804087]
56. Tibshirani R. A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics*. 2006; 7:106. [PubMed: 16512900]
57. Tsai CA, Wang SJ, Chen DT, Chen JJ. Sample size for gene expression microarray experiments. *Bioinformatics*. 2005; 21:1502–1508. [PubMed: 15564298]
58. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004; 5:R80. [PubMed: 15461798]
59. Quackenbush J. Microarray data normalization and transformation. *Nat Genet*. 2002; 32(Suppl): 496–501. [PubMed: 12454644]
60. Smyth GK, Yang YH, Speed T. Statistical issues in cDNA microarray data analysis. *Methods Mol Biol*. 2003; 224:111–136. [PubMed: 12710670]
61. Weiner AM. SINEs and LINEs: the art of biting the hand that feeds you. *Curr Opin Cell Biol*. 2002; 14:343–350. [PubMed: 12067657]
62. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet*. 1996; 14:457–460. [PubMed: 8944026]
63. Yang I, Chen E, Hasseman J, Liang W, Frank B, Wang S, et al. Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol*. 2002; 3:R0062.
64. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*. 1998; 95:14863–14868. [PubMed: 9843981]
65. Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. *Bioinformatics*. 2001; 17:309–318. [PubMed: 11301299]
66. Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*. 2000:455–466. [PubMed: 10902193]
67. Cui X, Churchill G. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*. 2003; 4:210. [PubMed: 12702200]
68. Benjamini Y, Hochberg Y. Controlling the false discover rate: a practical and powerful approach to multiple testing. *J Royal Stats Soc*. 1995; 57:289–300.
69. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*. 2001; 98:5116–5121. [PubMed: 11309499]

70. Hastie T, Tibshirani R, Eisen M, Alizadeh A, Levy R, Staudt L, et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 2000; 1:R0003.
71. Rajeevan MS, Vernon SD, Taysavang N, Unger ER. Validation of Array-Based Gene Expression Profiles by Real-Time (Kinetic) RT-PCR. *J Mol Diagn.* 2001; 3:26–31. [PubMed: 11227069]
72. Beissbarth T, Speed TP. GStat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics.* 2004; 20:1464–1465. [PubMed: 14962934]
73. Zhong S, Tian L, Li C, Storch KF, Wong WH. Comparative analysis of gene sets in the Gene Ontology space under the multiple hypothesis testing framework. *Proc IEEE Comput Syst Bioinform Conf.* 2004:425–435. [PubMed: 16448035]
74. Bussey KJ, Chin K, Lababidi S, Reimers M, Reinhold WC, Kuo WL, et al. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Mol Cancer Ther.* 2006; 5:853–867. [PubMed: 16648555]
75. Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, et al. From the Cover: Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci USA.* 2003; 100:605–610. [PubMed: 12529507]
76. Segal E, Wang H, Koller D. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics.* 2003; 19:i264–272. [PubMed: 12855469]
77. van Noort V, Snel B, Huynen MA. Predicting gene function by conserved co-expression. *Trends Genet.* 2003; 19:238–242. [PubMed: 12711213]
78. Semon M, Duret L. Evolutionary Origin and Maintenance of Coexpressed Gene Clusters in Mammals. *Mol Biol Evol.* 2006; 23:1715–1723. [PubMed: 16757654]

Fig. 1.
Generalized process flowchart for analyzing microarray data.

Fig. 2.
(A) A typical density histogram for a series of 20 slides. The red and green lines each indicate either the red or green channels. (B) A series of slides where there was a problem during RNA extraction and labeling.

Fig. 3.
(A) Foreground and (B) background signal intensities for a slide with a suitable degree of quality. (C) Foreground and (D) background signal intensities for a slide with problem that occurred during hybridization.

Fig. 4. Boxplot diagram for six slides with varying degrees of scatter and quality. A boxplot displays summary statistics such as central tendency and variability for each slide.

Fig. 5. M (log intensity ratios) versus A (log total intensity) plots showing scatter from a typical slide of (A) good quality and (B) poor quality. An MA (intensity scatter) plot compares intensity on two colors (or two chips).

Fig. 6. MA plot of a slide (a) before loess normalization and (b) after loess normalization.

Fig. 7.

A typical dendrogram obtained after performing a 2-way hierarchical clustering. Branches on the left indicate genes which cluster together according to similarity of expression. Likewise, branches across the top indicate the degree of similarity of individual samples. In this case, it is evident that there are two distinct groupings of samples according to their expression across a wide number of genes.

Listing 1

An R script for assessing the quality of microarray slides in an experiment.

```

1 library(limma)
2 files<-dir(pattern="^[1-9].*.txt")
3 RG<-read.maimages(files,source="arrayvision")
4 RG$genes<-readGAL()
5 RG$printer<-getLayout(RG$genes)
6 spottypes<-readSpotTypes()
7 RG$genes$Status<-controlStatus(spottypes, RG)
8 MA<-normalizeWithinArrays(RG,method="none",bc.method="none")
9 fnc<-file.path(".",paste("F-RGdensities",".png",sep=""))
10 png(filename = fnc, width = 6.5 * 140, height = 10 * 140,pointsize=20)
11 plotDensities(MA)
12 imageplot3by2(RG,z="R",low="white", high="red")
13 imageplot3by2(RG,z="Rb",low="white", high="red")
14 imageplot3by2(RG,z="G",low="white", high="green")
15 imageplot3by2(RG,z="Gb",low="white", high="green")
16 fnc<-file.path(".",paste("F-boxplot",".png",sep=""))
17 png(filename = fnc, width = 6.5 * 140, height = 10 * 140,pointsize=20)
18 par(ps=8,las=2)
19 boxplot(MA$M~col(MA$M),names=colnames(MA$M),main="boxplot")
20 numarrays <- ncol(RG)
21 numpages <- ceiling(numarrays/6)
22 for (ipage in 1:numpages)
23 {
24   i1 <- ipage * 6 - 5
25   i2 <- min(ipage * 6, numarrays)
26   fnc <- file.path(".", paste("MA plot", "-", i1, "-", i2, ".png", sep = ""))
27   png(filename = fnc, width = 6.5 * 140, height = 10 * 140,pointsize = 20)
28   par(mfrow = c(3, 2))
29   for (i in i1:i2)
30   {
31     plotMA(MA[, i])
32   }
33 }

```