Handout for ISMB 2008 tutorial: 'Introduction to microarray analysis'

Mark Reimers,
Department of Biostatistics,
Virginia Commonwealth University

Abstract

This tutorial is intended to introduce non-specialist researchers in systems biology and in medicine to the best current practices for the analysis of microarray technologies.

Introduction

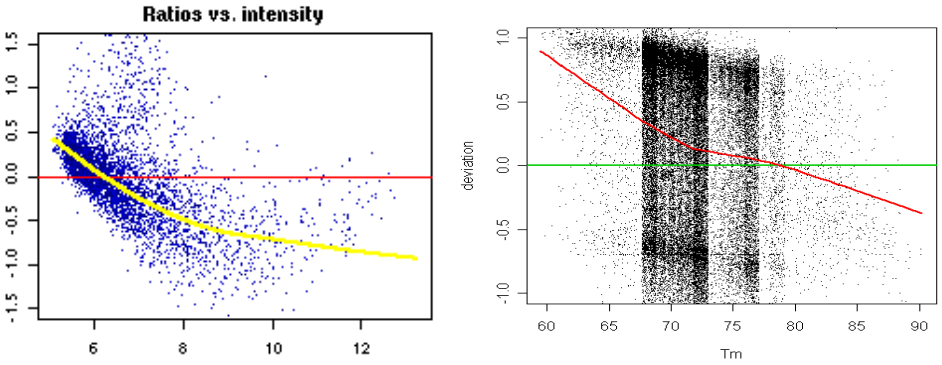
Modern microarray technologies enable researchers to gather data in quantities unimaginable only a decade ago. These data acquisition technologies are changing the nature of research in biology and are poised to revolutionize medical diagnosis and treatment.

Biostatisticians have accepted the challenge of analyzing these new data sets. There are two main stages in the analysis. First we must extract a clear signal from the technologies; this is the work of quality control and normalization. Second we try to identify clues to the biological processes at work: these may be individual genes or gene groups whose transcription change consistently. In analyzing such data we researchers are like the prisoners in Plato's Cave; with our measures we perceive only a shadow of the reality. We must think imaginatively and critically to infer that reality.

Quality Assessment

The first issue is how to decide which data are worth the investment of analysis effort. Although microarrays give quick streamlined high-throughput measures, much of the physical chemistry is hidden (by design) from the user. There are few places where the user can monitor the process and in particular compare the technical characteristics of one array against those of another. Often technical factors particular to one array can give very odd results without any obvious indication in the metrics that are routinely monitored. For example there may be particles of dust or scratches on a chip, air bubbles in the hybridization, or fingerprints or wipe-marks on a glass cover. These are usually not visible to the naked eye but can make a big difference in data.

A general approach to many kinds of normalization and QA issues is to compare any given chip to an ideal reference and look for unusually large departures that seem related with any known technical variables. In practice we don't know the ideal reference chip but usually a (robust) mean across all microarrays of actual probe values approximates the ideal reasonably well. Terry Speed noticed the intensity-dependent bias of many two-color arrays (see Figures 1A and 4) by plotting departure of log-ratios from ideal average (i.e. a log-ratio of 0) against mean intensity as the technical variable[1]. Similar plots for melting temperature or average probe intensity show often dramatic systematic departures from the ideal profile. Some of the most striking images come from analysis of spatial variability across a chip (see Figures 2 and 3)[2].



A B

Figure 1.
A. Plot of log ratio of intensity of chip GSM25377 (from GSE2552) relative to the average across all chips plotted against that average as a technical variable.
B. Plot of log ratio of chip AG1_1A4 against the average of series AG1_1 (from the MAQC project data) plotted against Tm for the probes as a technical variable.

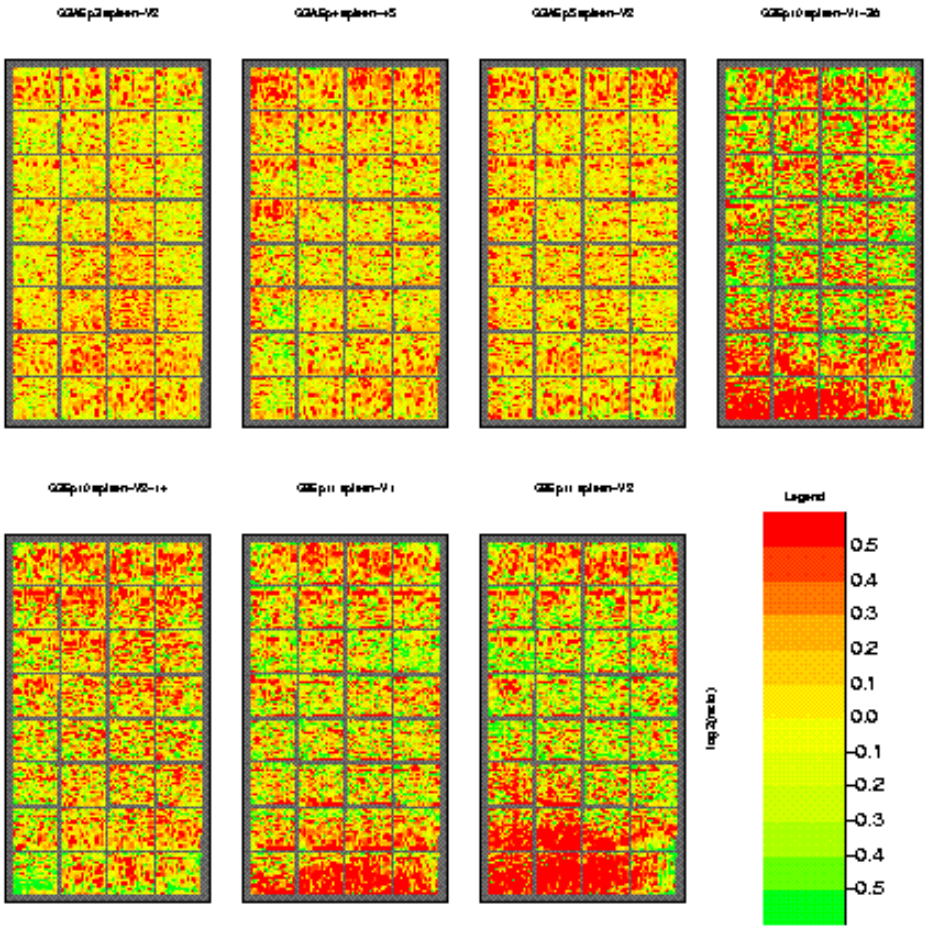


Figure 2.
Spatial plots of differences between log ratios on each of 9 NCI spotted microarrays and the average of log-ratios for the corresponding probes across nineteen samples from the same tissue.

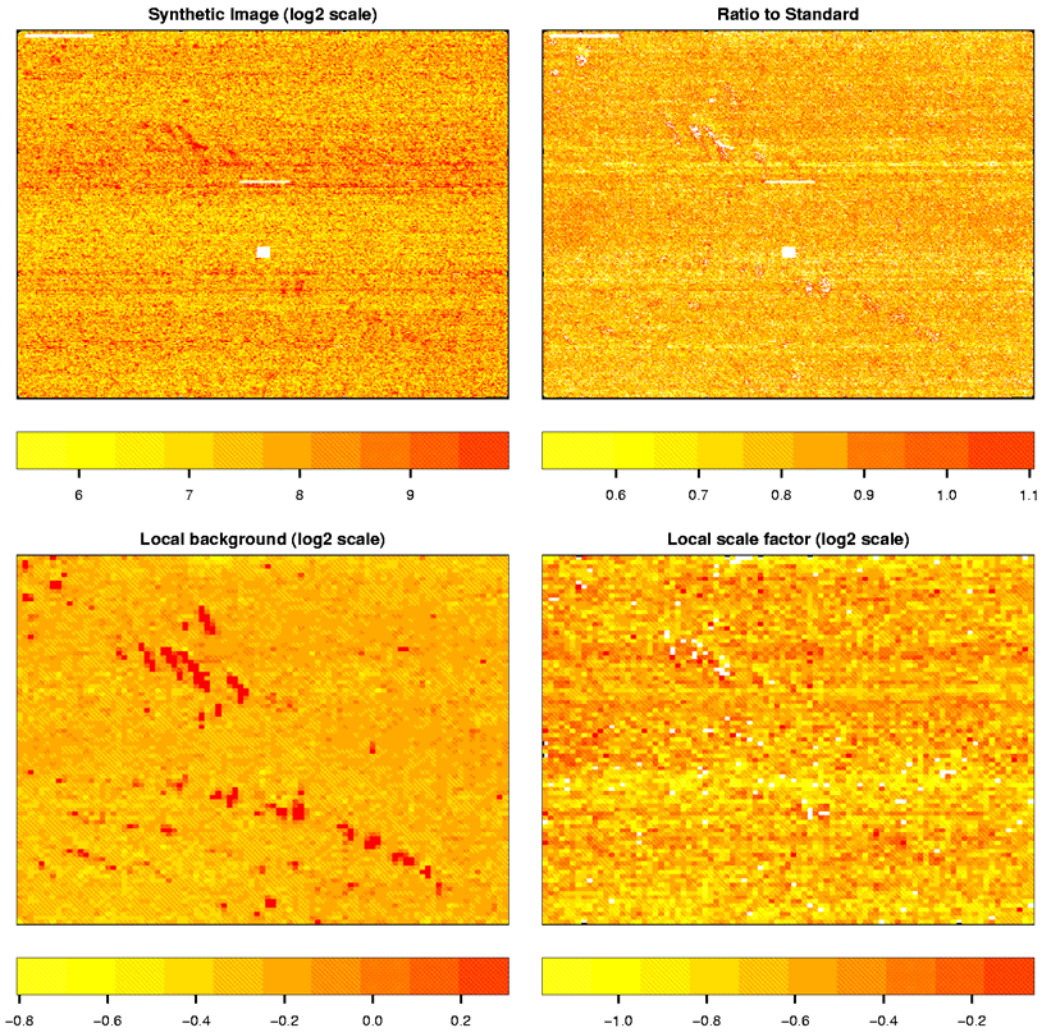


Figure 3.
Top left: spatial plot of intensities of one chip among the 23 replicates in the Gene Logic spike-in data set.
Top right: log ratios of the same chip relative to the average across all chips
Bottom left: log ratios of only the probes in the lowest 20% of all intensities (these probes are distributed across the chip)
Bottom right: log ratios of only the probes in the highest 20% of all intensities.

Issues in normalization of expression arrays

When we compare measures on biological samples obtained by a high-throughput assay such as a microarray, the differences in measures are due to biological differences in the samples and also to differences in the complex procedures by which those measures are produced. The most that normalization can hope to do is to compensate for the effects of the differences in procedures among the samples being compared. Data analysts generally don't have records of the procedures in enough detail to identify crucial differences in technique; and if such differences were clear the technicians would minimize them.

The technical differences most obvious to microarray pioneers in the late 1990's were that some arrays had much brighter scans than others. These types of differences had clear technical causes: variations among arrays in the amount of cDNA that was hybridized, the efficiency of the labelling reaction, and the scanner setting. Hence these large differences were very unlikely to be due to real biological differences among samples and researchers sought ways of compensating them. The simplest compensation was by dividing all the values on each chip by a chip-specific fudge factor that estimated the overall brightness[3]. This normalization made the mean value of all gene measures the same for each chip. For two colour arrays the normalization by was done in each channel separately, so the mean of each channel was the same. The computation is as follows, where f_i represents an intensity measure for gene i: and represents the normalized values.

$$(1)\quad C_{\text{red}} = \sum_{i=1}^{N} f_i^{\text{red}} \;;\quad C_{\text{green}} = \sum_{i=1}^{N} f_i^{\text{green}}$$
$$f_i^* = f_i^{\text{red}} / C_{\text{red}} \;;\; f_i^* = f_i^{\text{green}} / C_{\text{green}}$$

This normalization was often done directly on log-transformed data by subtracting the chip mean log-ratios so that the mean log-ratio becomes 0 for each chip.

The next development in array normalization came in 2001 when Terry Speed noticed that in fact the normalization done by (1) for two-colour arrays hadn't worked entirely. Some residual bias could be seen by plotting the log-ratio (log(R) / log(G)) against the average brightness in the two channels[1]. The same thing can be seen in one–colour arrays by plotting the intensities against the average as a reference. The bias could be estimated by a (non-parametric) curve, which could be estimated by a local regression (loess), as shown below at left. The values of log-ratios were adjusted by subtracting the height of the loess curve at the same average brightness, as shown below at right.

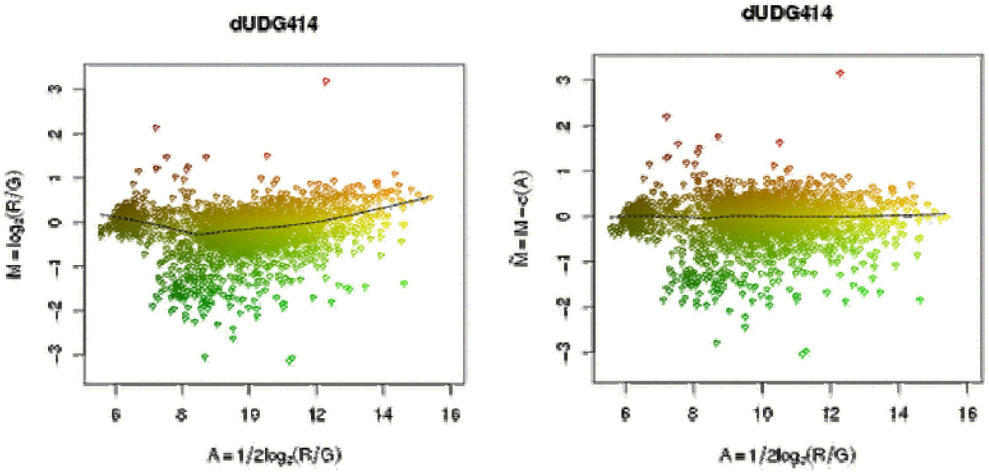


Figure 4.
Ratio-intensity bias normalization by local regression (loess). The left picture shows log-ratio on chip dUDG414 plotted against mean log intensity levels in the two channels; the loess curve is drawn in black. The right picture shows the same data after subtracting the local regression estimate of the mean log-ratio at each intensity. (courtesy Henrik Bengtsson)

Excursion: Data transformations

Most microarray data is log2-transformed, which makes it comparable to rt-PCR measures. This transformation is justified by the theory of proportional errors: if the errors are proportional to means then after a logarithm transform the errors in variables of very different sizes are roughly equal. However by 2002 some other statisticians had noticed that the variance of log transformed data was much wider at low intensities. This suggested to them that the model of proportional errors was incomplete. Their further refinement was that the error should be considered composed of two components, a roughly constant 'speckle noise' component, supposedly reflecting chip surface irregularities and electrical noise in the scanner, and a proportional error, supposedly reflecting variations in the hybridization process.

(2) $f_i \sim y_i + \sigma(y_i)\varepsilon_i^1 + \varepsilon_i^2$

Two researchers simultaneously published variants of a method for estimating the relative proportions of these different error types in actual array data, and devising a transform to 'stabilize' the error variances of different genes[4]. However the transforms are complex and hard to interpret, and these methods have not really caught on. They are sometimes used today for Illumina arrays, where many other sources of error are reduced. In most types of microarray the errors at the low end are usually dominated by systematic effects and are not well modelled by equation (2).

By 2003 statisticians were considering more complex error models Some noticed that there were pronounced differences of log-ratios in different regions of the same chip; they tried to fit 2-dimensional lowess surfaces to chips. Others tried to meld the lowess approach with the variance-stabilizing approach. These methods all seemed rather complicated. In 2003 Benjamin Bolstad, one of Terry Speed's students, proposed cutting through all the complexity by a simple non-parametric procedure[5]. In essence his proposal was to shoe-horn the intensities of all probes on each chip into one standard distribution shape. The standard or reference distribution was often determined by pooling all the individual chip distributions. The algorithm proceeded by mapping every value on any one chip to the corresponding quantile of the standard distribution (see Figure 5); hence it is called *quantile normalization*. This procedure worked as well as most of the more complex procedures, and certainly better than the regression method that was then the manufacturer's default for Affymetrix chips. Ben Bolstad also made the method available as the default in the *affy* package of Bioconductor (www.bioconductor.org), which has become the most widely used suite of freeware tools for microarrays. For all those reasons quantile normalization has become the *de facto* standard today.

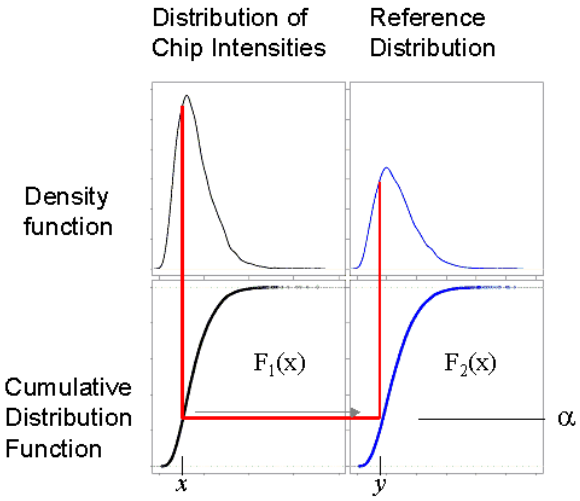


Figure 5
Schematic of quantile normalization
On any chip the value at any specified quantile (for example the 20th percentile) gets changed to corresponding quantile (e.g. the 20th percentile) of the reference distribution.

While quantile normalization was a simple fast one-size-fits-all solution, it engendered considerable problems of its own. Perhaps the first such problem to command attention was that the genes at the high end were shoe-horned into the same distribution shape and thus the changes in those abundant genes were usually distorted and often entirely missed. A recent adjustment to the quantile procedure in the latest versions of the *affy* package fixes that problem. The second issue is more subtle. For reasons that are still not entirely clear, the errors in various probes are highly correlated. Among the 50% of genes that are not expressed in a particular set of samples these correlated errors form a large part of the variation between chips. When quantile normalization acts on these probes, the procedure preserves this apparent but entirely spurious correlation among low-intensity probes.

A radically different approach builds directly on the observation of correlated errors. The idea is that a small number of technical variables cause most of the non-biological variation among chips. The aim of normalization is to try to estimate the technical bias associated with each probe by comparing deviations of other probes with similar technical characteristics but different biology. This procedure requires knowledge of the technical characteristics of each probe and a suitable regression algorithm to map deviations from the ideal profile onto those technical predictors. Early results look promising; however the programming of the algorithms is much more complex and execution times are longer compared to quantile normalization.

Methods for summarizing raw values from oligonucleotide arrays
There are many competing methods for aggregating the measures from multiple-probe oligonucleotide arrays into probe-set summaries. There are still many papers published and more submitted comparing various leading methods on various standard data sets. Broadly speaking the better methods use some non-linear normalization – quantile normalization often does pretty well – and some form of linear model to combine intensities from different probes within each probe set[6]. The key advantage of a linear model over a simple average is that probes of very different intensities can each make independent contributions rather than letting the brightest probes dominate the summary.

Generally speaking these linear model methods do better than the averaging methods that Affymetrix developed, and now Affymetrix uses linear model methods for their arrays. There are several linear model variants in common use. The gcRMA program in Bioconductor attempts to estimate and subtract non-specific hybridization based on an empirically fitted model. This procedure works very well in some circumstances but seems to result in unrealistic values when applied to other data sets. There is also still some disagreement on whether it is best to use a log scale or not. The dChip software (www.dchip.org) fits a linear model in the original scale; the affy and oligo packages in Bioconductor work on a log2 scale.

Pre-processing of CGH, SNP, and Epigenetic Arrays

The new array types pose most of the same types of pre-processing issues and several important new issues. The same issues of intensity-dependent bias and spatial artifacts recur with these arrays, but new issues arise specific to the type of data under consideration. For most of these arrays the effects of variation in technical conditions and the probe-specific biases are quite strong and so accounting for technical variables in normalization helps even more than in expression arrays.

With CGH arrays the unique new issue is segmentation. For various reasons (e.g. that most cancer samples are actually mixtures of tumour and normal) on most CGH arrays the reported ratio of signals is less than the ratio of actual copy numbers; hence the signal to noise ratio is low for most probes. However even within a complex cancer genome most genes have the same copy number as their neighbours. Thus it becomes critical to take advantage of the dependence between neighbours to get the most efficient estimation. Most practical methods for segmentation try to maximize the t-score between adjacent regions declared as ‘segments’. They differ mostly in how they set about it and the degree of mathematical rigor behind the method [7].

With chromatin immuno-precipitation arrays (ChIP-on-chip) the issue is separation of true enrichment signal from noise. Similar issues arise when analysing expression data from tiling arrays. The first studies in this area used a relatively crude approach: they fit a Normal curve to the center and left half of the distribution of signals, and declared the signals more than 3 standard deviations above the mean on the right end of this Normal curve to be significant. This rough and ready procedure ignores the characteristics of the probes that contribute to signal differences. More sophisticated procedures attempt to estimate the signal due to the particular characteristics of each probe sequence, so as to make a more accurate guess as to when particular signals exceed what should be considered as ‘background’. Probably the best current procedure is MAT from Shirley Liu’s lab[8]. MAT fits a regression to variables recording the identity of each individual base in any probe sequence.

Genotyping arrays aim to distinguish one among three possible genotypes at a locus. Here the effects of technical variation acting differently on different probes has been addressed. Rafa Irizzary’s group has put together a fairly complex algorithm CRLMM, which accounts for technical biases in two stages, and seems to give extremely accurate genotype calls on the CEPH standard samples[9]. Their algorithm fits a regression to individual bases and to the length of the PCR fragments which contain target SNPs.

Identifying individual genes of interest

Corrected p-values and FDR

Multiple comparisons issues play a large role in microarray experiments because of the large number of variables (genes). Suppose that you have a typical microarray with 20,000 genes represented, and are comparing two conditions that (unknown to you the experimenter) actually differ not at all. If you perform a typical t-test at a significance level of 0.1% then you might expect to select 200 genes that appear different.

Definitions

The (default) Null Hypothesis about a gene is that no change occurs in that gene between the groups of samples that are being compared. We number genes 1, …, M, and we let M denote the corresponding Null Hypotheses to be tested by H_1, ..., H_M. We suppose that M0 < M of these Null Hypotheses are actually true (no true changes in those genes). To keep track of the numbers we'll use the table below:

Hypotheses	Accepted	Rejected	Number
True	U	V	M0
False	T	S	M-M0
	W	R	M

The number of null hypotheses rejected wrongly ('false discoveries') is denoted by V and the number of rejected null hypotheses is denoted by R. Note that R is an observed random variable, S, T, U, and V are all unobservable (random) values, while M and M0 are fixed numbers, although M0 is unknown to the investigator.

We distinguish three types of errors.
1. False positives occur when a true null hypothesis is rejected; this is known as a Type I error.
2. False negatives occur when a false null hypothesis is retained, i.e. a true difference is not discovered; this is a Type II error.
3. Type III errors when a null hypothesis is correctly rejected, but with a wrong directional decision. These should not happen much in theory but opposing results from different labs suggest they occur often in microarray data.

We describe four characterizations of the false positive rates
i) The *per-comparison error rate* PCER = E(V)/M is the expected proportion of type I errors among the M decisions.
ii) The *family-wise error rate* FWER = P(V > 0)is the probability of committing at least one error.
iii) The *false discovery rate* (in the sense of *Benjamini and Hochberg*), $FDR_{B\text{-}H}$ = E(V/R | R > 0) P(R > 0), is the long run average fraction of false positives among all gene lists, counting empty gene lists as having no false positives.
iv) The *false discovery rate* (in the sense of *Storey*) FDR_{Storey} = E(V/R | R > 0) is the long run average fraction of false positives among non-empty gene lists.

Both senses of FDR are related to, but differ from the PCER, which is the expected proportion of false positives among the apparently significant results. The Storey FDR is perhaps closer to our intuitive sense of false discovery rate, while the Benjamini and Hochberg FDR is easier to work with and is widely used[10]. Usually in practice the

intuitive FDR is greater than the reported FDR ($FDR_{Storey} > FDR_{B-H}$): if there are many highly significant genes the two FDR's will be similar; if there are only a few genes with acceptable values of the reported FDR, then the intuitive FDR (FDR_{Storey}) will be notably higher than the FDR_{B-H} and there may be no genes at an acceptable level of intuitive FDR[11].

Permutation procedures for FDR

The most common approach to estimating significance levels is to permute the sample labels many times and count the number of (false) positives obtained by performing the selection procedure each time. Benjamini and Hochberg presented their theory in terms of the mean value of (their) FDR, and it has become standard to use the mean value of the positive count from the various permutations as an estimate of the numbers of false positives in any procedure. In fact the distribution of false positive counts is almost always highly skewed, so that the mean is unstable, and is probably not the figure of most interest. Some researchers specify an upper 90% confidence bound for the number of false positives, based on the permutation distribution.

Identifying recurrent genomic aberrations in CGH and methylation

Cancer research leads the way in studies of changes in gene copy number or methylation status. For example CDKN2A (p16), an important regulator of the cell cycle, is frequently deleted in cancers of many tissues. This fact suggests that p16 loss may be an important step on the way to malignancy for a many cancers; however loss of this gene is not necessary since p16 loss is not observed in a majority of cancers of any tissue.

Tumours are an illustration of an issue that faces researchers increasingly in the age of genomics: biology provides many mechanisms to a common phenotype. Individual variation will increasingly be an issue in research on tumour biology as well as in providing personalized medicine.

Thus it is necessary to devise methods that are designed to detect changes that recur frequently rather than ideal changes that are confounded with random noise. The classical t-test and its many more efficient variations used in genomics are aimed at this Platonic ideal.

Several authors have published methods for this problem as it occurs in CGH data. Their methods are usually based on adding up the aberrations for each locus across samples. The significance of each score is determined by permutations of the segments within the data[12].

Identifying gene groups of interest

As microarrays are becoming more comprehensive and more reliable the number of genes detected as differentially expressed in many studies has exceeded the researchers' capacity to interpret. One approach to automated interpretation is to leverage gene groups available from Gene Ontology or curated databases of pathways such as BioCarta. The idea is that genes in a gene group are carrying out some co-ordinated function and if these genes are simultaneously up- (or down-) regulated then it is plausible that this function or pathway is a major player in the condition under study.

Three kinds of approaches have been developed. The earliest and simplest approach is simply to take the list of differentially expressed genes from the t-test and ask whether any of the functional groups under consideration is over-represented in that list. This kind of simple discrete (categorical) thinking is often an easy first step in analysis. However it was noted in a paper from the Broad Institute that biologically meaningful co-ordinated changes may not include many (or any) genes with sufficiently large changes to achieve statistical significance (after multiple comparisons correction). Furthermore statisticians know that procedures on continuous variables are more powerful than procedures involving discretized variables. Hence a variety of procedures have been invented by combining t-scores for individual genes in pathways. Finally it seems likely that the co-ordinated changes most biologically significant are those that run counter to the normal covariation. This idea forms the basis of several procedures for selection procedures based on multivariate analysis of gene set co-expression.

The discrete approach

The simplest approach (and one that is often a good beginning for a new type of question) is to classify the genes into two sets: those which are significantly changed (or significantly changed in a specific direction), and the rest. Then we may cross-classify genes in terms of change and membership in any group of interest, and test the significance using a standard χ^2-test or a more accurate Fisher's Exact Test[13].

The univariate continuous approach

In 2003 researchers at the Broad Institute introduced the Gene Set Enrichment Analysis (GSEA) procedure, which uses Kolmogorov-Smirnov (K-S) test of distribution equality to compare the distribution of t-scores for a selected gene group with the distribution of t-scores for all genes. Gene sets whose t-scores clearly come from a different distribution are likely to represent changed functions[14].

A year later a simpler approach appeared in the form of Parametric Analysis of Gene Expression (PAGE). It is much easier to detect a very specific change of means as opposed to detecting any of the many possible changes of distribution. The PAGE test statistic is essentially a z-score: $Z = (m_G - m) / S_{All}$, where m is the mean of all fold changes, m_G is the mean of fold changes of genes in group G, and S_{All} is the standard deviation of all fold changes[15].

The multivariate approach

Most statistical procedures are more powerful when the covariation in the different measures is specifically estimated and accounted for. The Hotelling's T-squared method [16] explicitly aims at comparing the size and direction (think of co-ordinated changes) of the difference in means (between two groups) to the size and directions of typical variation within groups: $T^2 = (\bar{x}_1 - \bar{x}_2)^T W^{-1} (\bar{x}_1 - \bar{x}_2)$, where W is the sample covariance matrix. Directions of little variation within groups correspond to small values (actually 'eigenvalues') of W which correspond to big values of W^{-1}. Thus a difference in means between the two groups in a direction in which little normal variation occurs counts for more than a difference between groups that could occur easily within groups. So if within groups two genes are tightly correlated, but between groups one changes up and the other changes down, that is a difference unlikely to arise by random sampling.

References

1. Yang, I.V., et al., *Within the fold: assessing differential expression measures and reproducibility in microarray assays.* Genome Biol, 2002. **3**(11): p. research0062.
2. Reimers, M. and J.N. Weinstein, *Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases.* BMC Bioinformatics, 2005. **6**: p. 166.
3. Quackenbush, J., *Microarray data normalization and transformation.* Nat Genet, 2002. **32 Suppl**: p. 496-501.
4. Durbin, B. and D.M. Rocke, *Estimation of transformation parameters for microarray data.* Bioinformatics, 2003. **19**(11): p. 1360-7.
5. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.
6. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res, 2003. **31**(4): p. e15.
7. Olshen, A.B., et al., *Circular binary segmentation for the analysis of array-based DNA copy number data.* Biostatistics, 2004. **5**(4): p. 557-72.
8. Johnson, W.E., et al., *Model-based analysis of tiling-arrays for ChIP-chip.* Proc Natl Acad Sci U S A, 2006. **103**(33): p. 12457-62.
9. Carvalho, B., et al., *Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data.* Biostatistics, 2007. **8**(2): p. 485-99.
10. Reiner, A., D. Yekutieli, and Y. Benjamini, *Identifying differentially expressed genes using false discovery rate controlling procedures.* Bioinformatics, 2003. **19**(3): p. 368-75.
11. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-5.
12. Beroukhim, R., et al., *Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma.* Proc Natl Acad Sci U S A, 2007. **104**(50): p. 20007-12.
13. Zeeberg, B.R., et al., *GoMiner: a resource for biological interpretation of genomic and proteomic data.* Genome Biol, 2003. **4**(4): p. R28.
14. Mootha, V.K., et al., *PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.* Nat Genet, 2003. **34**(3): p. 267-73.
15. Kim, S.Y. and D.J. Volsky, *PAGE: parametric analysis of gene set enrichment.* BMC Bioinformatics, 2005. **6**: p. 144.
16. Kong, S.W., W.T. Pu, and P.J. Park, *A multivariate approach for integrating genome-wide expression data and biological knowledge.* Bioinformatics, 2006. **22**(19): p. 2373-80.