STATISTICS AND GENE EXPRESSION ANALYSIS

TERRY SPEED

Department of Statistics,
University of California at Berkeley
Division of Genetics & Bioinformatics,
Walter & Eliza Hall Institute of Medical Research
Melbourne, Australia

1. Introduction

A (protein coding) gene is determined to be expressed in a cell or group of cells when its transcribed messenger RNA (mRNA) or the resulting protein product is detected. There are a wide variety of techniques for determining and quantifying gene expression, and many of these have substantial statistical components to them. In this article, we review some of statistical models and methods used in analysing gene expression data, focussing entirely on approaches quantifying mRNA. The large-scale measurement of protein is under active development, and while that too has its statistical problems, these are too broad to be dealt with here.

Before discussing statistical matters, it will be helpful to present a small sample of the extensive biological and technological background to gene expression analysis.

Why do we measure gene expression? The most common experiment is comparative: we want to compare the mRNA levels of one or more genes in cells from different sources. Comparisons of interest include tumour vs normal cells, cells from a specific organ in a mutant or genetically modified organism vs cells from the same organ in a normal organism of the same strain, and cells before and after an intervention such as a drug treatment. Another important class is the time-course experiments, where cells are sampled at different times, e.g. after the administration of a drug, or as the cell cycle or development proceeds, and interest is in temporal patterns of gene expression. Yet other experiments focus on spatial patterns of gene expression. There are many other kinds of gene expression experiments, essentially as many as there are organisms, cell types and conditions of biological interest.

How do we measure gene expression? As stated above, there are many techniques for doing so, but most rely on DNA-RNA or DNA-DNA hybridization. This is the process through which single-stranded DNA or RNA molecules find and base-pair with their complementary sequences amidst a complex mixture of many molecules of the same

kind. The terminology we adopt names the sequence representing a gene of interest the probe, while the pool within which a complementary copy of the probe is sought is named the target DNA or RNA. Other terminologies are the reverse of ours.

On what scale do we measure gene expression? Much of the recent interest by statisticians in this area stems from the availability of data sets giving expression measurements on tens of thousands of genes, so-called microarray gene expression data. However, nylon membrane filters with thousands of genes spotted on them have been around for over a decade, and smaller-scale quantitative expression data for much longer. We begin with a discussion of the first and simplest method of quantifying RNA, as many of the features of the high-throughput methods are already present here.

2. Low-throughput methods

2.1. **Quantitative northern blots.** Isolated RNA is separated according to size by electrophoresis, and transferred by blotting to an immobilizing matrix such as a nylon membrane. A labelled DNA probe is incubated with the blot under conditions which promote annealing, and the probe will then bind to the RNA molecules on the blot complementary to it. This is the hybridization reaction. The result is then imaged, either directly (e.g. by laser scanning or with the use of a CCD camera), or indirectly, by exposing an X-ray film to the blot.

The amount of RNA can be quantified by measuring the intensity of the signal in the image in regions corresponding to the probe of interest. Usually control RNA is measured at the same time, typically a gene that is thought to be expressed at a more or less constant level, (a so-called housekeeping gene), and the expression level of the gene of interest is then given relative to the control gene.

Although this technique has been in use for over 20 years, it has attracted little attention from statisticians. In part this is because low-throughput assays with simple read-outs are usually seen as outside the domain of statistical analysis, apart from such simple matters such as analysing replicate data. This attitude changes when the assay becomes high-throughput, or when much more data are collected on a given unit. These considerations lead naturally into our next topic, which is an important development of the northern blot.

2.2. **Quantitative PCR, including kinetic or real-time PCR..** The Polymerase Chain Reaction (PCR) can be used to estimate the concentration of a particular target RNA relative to a reference. The mRNA is converted to complementary DNA (cDNA) using the enzyme reverse transcriptase (sometimes abbreviated by RT), and the result is amplified exponentially using the PCR. References are control sequences (such as housekeeping genes) that are present in the same

preparation of RNA as the target sequence. Quantification is achieved amplifying the target RNA (and the reference RNA) to a more readily detectable quantity, and by comparing the amount of amplified product generated by the reference and the target sequence. There are many variants. The older assays measured end-product, but the method works better if the amplified products are measured during the exponential phase of the chain reaction, particularly if the reference and target sequence are present in approximately equal concentrations, and if they amplify with equal efficiency. More accurate variants involves adding reference molecules in known amounts to a series of amplification reactions.

A recent technique of the second kind for quantitating RNA is called kinetic or real-time PCR. There the target and reference sequences are amplified and the products detected in the same instrument, and the endpoint is when the reported fluorescence passes a fixed threshold above baseline. Note that "real-time" here is sometimes abbreviated by RT, and so can be confused with the same abbreviation for reverse transcriptase. In fact kinetic PCR is really RT RT PCR! There are a large number of different protocols, including TaqMan, and a number of different instruments for carrying out this assay. Details can be found in the technical notes below. Rather more statistical research has been devoted to improving quantitifation methods for RT-PCR, see e.g. Pfaffl (2001), but there are still many issues remaining. This is a very fertile area for biostatisticians. I for one am getting into it, but there is plenty of room for others.

It is important to point out here that the gene whose expression levels we wish to measure needs to be specified in advance, indeed parts of its DNA sequence need to be known in order to allow the preparation of primers necessary for the PCR

3. High-throughput methods: SAGE

Serial analysis of gene expression (SAGE) is a method for the comprehensive analysis of gene expression patterns. It is the main quantitative approach to gene expression not based upon hybridization. More importantly, one does not need to know the sequences of the mRNA transcripts in advance.

Three principles underlie the SAGE methodology:

i) a short sequence tag (10-14bp) contains sufficient information to uniquely identify an mRNA transcript, provided that that the tag is obtained from a unique position within each transcript;
ii) sequence tags can be linked together to form long serial molecules that can be cloned and sequenced efficiently and relatively cheaply; and
iii) a count of the number of times a particular tag is observed provides the expression level of the corresponding transcript. Thus

we actually count here, rather than measure indirectly as with all other technologies discussed so far.

A typical SAGE experiment would involve two sources of mRNA, say tumour and the corresponding normal tissue. For each source a set (called a library) of (say) 50,000 tags would be derived using the SAGE protocol. In these two libraries there might be 20,000 distinct (termed unique) tags observed, and for each unique tag, the frequency with which that tag appeared in each library could be calculated. The data for this comparative experiment are then two lists of counts, one for each unique tag observed.

The first question a biologist asks here is: which tags are significantly differentially represented in the two libraries? For any given tag, say tag i, the natural null hypothesis here is Hi : the proportions of tag i in the two libraries coincide. Rejection of this null hypothesis leads to the conclusion that the gene corresponding to tag i is differentially expressed between the two sources of RNA. Making an independence assumption that might be difficult to verify, one current approach to this question starts with the observation that under Hi, the number of times tag i appears in library 1, say, given the total number across the two libraries, is binomial with $p = 1/2$. This is the basis of a test of Hi, and when this is done for all $i = 1, ..., 20,000$, a Bonferroni adjustment can be used. The test just described is one of a number in use, (Audic & Claverie, 1997, Man et al, 2001). There are a range of outstanding questions with these data including dealing with sequencing errors, which might be of the order of 1-3% per base in the tags; considering the independence assumption leading to the binomial model; and seeking a valid multiple testing correction less conservative than Bonferroni. The difficulty is that because of the co-expression of genes, different tag counts in a library cannot be regarded as independent. However, the extent to which this matters is not yet clear. When more SAGE libraries accumulate in a given context, questions will undoubtedly arise which lead naturally to classification and cluster analyses, see below in the context of microarray data. As with the technologies outlined above, there seem to be many opportunities for biostatistical research involving SAGE data. A general source on this topic http://www.sagenet.org.

4. High-throughput methods: array based approaches

The principal class of high-throughput methods for quantifying gene expression are those based on microarrays, although the term macroarray is also used for the older nylon technology. Broadly speaking there are three basic microarray technologies: nylon membrane arrays, spotted arrays, and high-density oligonucleotide arrays. The special supplements Nature (1999, 2002) provide a good overviews of the production and utilisation of the last two technologies. We explain each briefly

before turning to statistics. There we will attempt to discuss the issues in a general way when applicable to two or more of these technologies, and leave to the reader to consult the references for material on topics rather more specific to the different technologies.

4.1. **Different array technologies.**

4.1.1. *Nylon Membrane Filters.* This is the oldest array technology, but one which is still widely used around the world. A typical filter microarray has $5,000$ complementary DNA (cDNA) clones $600 - 2,400$ bases in length, spotted in a grid on the membrane. Radio-labeled target cDNA derived from the mRNA of interest is hybridized to the array, and the filter is then exposed to X-ray film and the film imaged. The resulting digital image constitutes the raw data from the experiment.

A very high-density variant of the traditional filter-based microarray is the oligonucleotide filter array, which can have $50,000$ spots consisting of pools of 10-mers, Meier-Ewart et al, 1998.

4.1.2. *Spotted cDNA Microarrays.* Introduced in Schena et al (1995), a typical spotted array consists of up to $40,000$ cDNA probes of length $600 - 2,400$ bp placed in a regular pattern on a glass microscope slide. The main advantage of the non-porous glass support is that it facilitates miniaturization and the use of fluorescence (rather than radiolabel) based detection. Essentially all spotted arrays use two sources of mRNA, each labelled with its own fluorophore. These are mixed in equal quantities and competitively hybridized to the spots on the slide. In an obvious sense, each spotted array experiment may be regarded as several thousand paired comparisons. Following the hybridization, laser excitation stimulates the spots to fluoresce, and the photons emitted are collected, amplified, converted to digital form and presented as two digital images of the slide, each quantifying the amount of cDNA on the spots labelled by one of the two fluorophores. These two digital images are the raw data of a spotted microarray experiment.

A variant of the spotted arrays uses as probes long ($60 - 75$ bp) oligonucleotides representing part of a gene or EST, Hughes et al (2001). These are put onto the glass using an ink-jet printer device, and generally lead to higher quality data, As with the original spotted arrays, a two-colour system is used, although the technology may well be good enough to provide reliable single colour quantification.

4.1.3. *High-Density Oligonucleotide arrays.* A quite different technology can be used to place up to $500,000$ short (25 bp) oligonucleotide probe pairs on a small glass chip, with 11-20 of these probe pairs representing a part or all of a single gene, see Fodor et al (1991) and Lockhart et al (1996). Each probe pair consists of a perfect match (PM) probe, and a mismatch (MM) probe, the latter being the same as the former

apart form a single nucleotide change ($A \leftrightarrow G$ or $C \leftrightarrow T$) in the middle (13th) position. A tagged target cRNA sample hybridizes with the complementary oligonucleotides on the chip, and detection is via laser excitation followed by the collection of fluorescence emission as with spotted arrays. As with the approaches already discussed, the image is the starting point of analysis.

5. Statistical issues

5.1. **Design of experiments.** The careful design of microarray experiments is in its infancy. Most work to date concerns spotted array experiments, which require more care by virtue of the paired nature of each experiment. Also, many users of spotted arrays construct the arrays themselves, whereas filter arrays and high-density oligonucleotide arrays tend to be bought "off the shelf". In spotted array experiments, there are thus two main aspects to the design question:

(i) the design of the array itself, i.e., deciding which cDNA probe sequences to print on the slide, whether to use replicated spots and control sequences, and how many and where these should be printed on the slide;
(ii) the allocation of mRNA target samples to the slides, i.e., deciding how mRNA samples should be paired for hybridization, the dye assignments, and the type and number of replicates.

Proper experimental design is needed to ensure that questions of interest can be answered and that this can be done accurately and efficiently given experimental constraints, such as cost of reagents and availability of mRNA. Designs specifically suited for the question of interest and judicious pairing of mRNA samples for hybridization can greatly improve the efficiency of microarray experiments by ensuring the precise measurement of relevant effects. A number of statisticians have been involved in these questions, but there is little literature so far. For some initial work in this area, see Kerr & Churchill (2001), and Yang and Speed (2002). We can expect much more published research on this topic in the near future.

5.2. **Image analysis.** As explained above, the "raw data" arising from all microarray technologies are images: of labelled probes on a nylon filter, a glass slide or a glass chip. There seems little doubt that the results of downstream analyses can be appreciably influenced by the initial image analysis, though few studies of this topic exist at present, see Yang et al (2001b) for one such.

Three broad analysis issues can be identified with microarray images, although not all approaches proceed in this way: finding the probe centres (registration), partitioning the pixels in the image into probe and non-probe regions (segmentation), and assigning summary values to probe intensity and background (quantification). Rather than

assign pixels to probe and non-probe categories, some approaches (especially with with nylon filters) use parametric, semi-parametric or non-parametric modelling to determine probe intensity. Once summary values of probe intensities are calculated, there remains the question of combining these to measure absolute or relative gene expression. With nylon filter and spotted arrays, intensity is usually the difference of foreground and background values, and ratios of these quantities are the main vehicle for later analysis. In general there are many ways of carrying out the image analysis, and several commercial and freely available packages for doing so, see Carlisle et al (2000) for nylon filter arrays, Yang et al (2001b) and references therein for spotted arrays, and Schadt et al (2000) for high-density oligonucleotide arrays. Brandle et al (2001) is a good overall reference, and other articles in that volume can be consulted on this topic, and Buhler et al (n.d.) is also useful.

In the case of high-density oligonucleotide arrays, the image analysis does not result in expression values, but in PM and MM probe intensity values. One further analytical step is necessary with this technology before se have a gene (or probe set) expression value: the 16 or 20 PM, MM pairs must be summarized. This is not entirely straightforward and research on it is continuing, but see Li & Wong (2001a, b) for the most thorough published discussion to date, and Irizarry et al (2003).

5.3. **Preprocessing tasks: normalization.** As indicated earlier, the most common gene expression experiment is the comparative one. With nylon filter arrays this leads us to compare the images from two hybridizations on to copies of the same basic filter. Sometimes this is done by stripping the results of a first hybridization and re-using the filter, but more commonly a new filter is used. Because the nylon substrate is not solid, there may be warping, and this can make registration across different filters a challenging problem. When this is adequately addressed, interest focusses on comparing the two expression levels for each of the genes spotted onto the array. An entirely analogous situation arises when we have reduced the two images of a single spotted array or two high-density oligonucleotide array experiments to lists of gene expression values. We are back to the same (biologist's) question that we met with SAGE data: which genes seem to be significantly differentially expressed between the two mRNA sources?

Before we can address this question in the microarray context, however, there is usually a need for normalization. This is a generic term describing the identification and removal of systematic sources of variation, other than differential expression, from the measured gene expression values. Systematic effects can come from different labeling efficiencies, different scanning parameters, and a variety of other causes, see Schuchhardt et al (2000) for a good list. These effects can be related to intensity, location on the filter, slide or chip, and other features of the

process such as reagent batch and lab conditions. The need for normalization can be seen most clearly in experiments involving two identical mRNA samples hybridized to different membranes or chips, or on the same glass slide, as long as the results are appropriately visualized.

Pairs of gene expression values, say from a treated (T) and a control (C) source, are usually displayed by plotting the $\log_2$ or $\log_{10}$ intensities against one another, e.g. $\log_2 T$ vs. the $\log_2 C$. Such plots give an unrealistic sense of concordance between the two sets of intensities and can mask important features of the data. It is better to plot $M = \log_2(T/C)$ against $A = \log_2(TC)$, which amounts to a rotation of the previous plot and a rescaling of the axes. Assuming, as is almost always the case, that we expect the majority of genes to be expressed at about the same level in both cell samples, regardless of overall intensity, the MA-plot should be scattered around the horizontal (A-) axis, in a more or less symmetric manner, and the histogram of M values should be centered around zero. This is rarely found to be the case.

A standard normalization for nylon filter and spotted array data is to shift the log ratios so that their mean or median is zero. Frequently there is a strong enough intensity dependence that a smoothing of M values along the A axis defines a better, A-dependent centering. Spatial effects require a modified solution, and there are yet others effects that need to be dealt with from time to time. For a discussion of these issues in the context of spotted arrays, see Yang et al (2001a, 2002), while Schuchhardt et al (2000) is also of interest. Normalization is also relevant to the high-density oligonucleotide technology, but is less well discussed and somewhat more complex, see Li & Wong (2001a, b) and Bolstad et al (2002).

5.4. **Comparative analyses.** Once the log ratios of intensities have been normalized, interest focusses on those which seem to be genuinely different from zero, i.e. which correspond to genes which are differentially expressed. There is no reliable method of assigning statistical significance to log ratios from unreplicated experiments, although a number of model-based approaches claiming to do this can be found in the literature, see Dudoit et al (2002b) for a discussion of this issue in the context of spotted microarrays. For a single comparison, the best approach is probably to apply a careful normalization to the log ratios, rank them and construct a normal qq-plot of them. Typically the plot will not be linear, but an examination of the extremes in conjunction with the MA-plot can give a good sense of the outlier log-ratios.

It is also advisable to carry out a quality examination of the spots corresponding to extreme log ratios. Exactly where to draw the line with ranked log-ratios, when determining putatively differentially expressed genes, will depend on a variety of factors such as the shape of the qq-plot, the level of false positive and false negative rates deemed

acceptable, and the nature and number of follow-up experiments envisaged. No simple guidelines seem possible, and no formal statistical approach seems available which deals with the question. The situation is different when there are replicate pairs of filters, slides or chips. We broaden the context somewhat to discuss the issue of multiple testing more generally.

5.5. **Multiple testing.** The identification of differentially expressed genes, i.e., genes whose expression levels are associated with a response or covariate of interest, is but one of the testing problems which arise with microarray data. The covariates could be either polytomous e.g. treatment/control status, cell type, drug type, or continuous, e.g. dose of a drug, time, and the responses could be, for example, censored survival times or other clinical outcomes. The biological question of differential expression can be restated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of no association between the expression levels and the responses or covariates. As a typical microarray experiment measures expression levels for several thousands of genes simultaneously, we are faced with an extreme multiple testing problem. Special problems arising from the multiplicity aspect include defining an appropriate Type I error rate (i.e. false positive rate) and devising powerful multiple testing procedures which control this error rate and account for the joint distribution of the gene expression levels.

A number of recent papers have addressed the question of multiple testing in the context of microarray experiments (Efron et al, 2000, Golub et al, 1999, and Tusher et al, 2001) However, the proposed solutions were not cast in the standard statistical framework and do not provide adequate Type I error rate control. When going from single to multiple hypothesis testing, several definitions of the Type I error rate are possible and include: the per-comparison error rate (PCER), defined as the expected value of (number of Type I errors/number of hypotheses); the family-wise error rate (FWER), defined as the probability of at least one Type I error; and the false discovery rate (FDR), or expected proportion of Type I errors among the rejected hypotheses. In general, for a given multiple testing procedure, PCER (FWER and FDR (FWER, one should thus decide on an appropriate error rate to control for the problem under consideration. It is important to note that the expectations and probabilities above are conditional on assumptions concerning which hypotheses are true, i.e., on which genes are differentially expressed. A fundamental, yet often ignored distinction in multiple testing, is that between strong and weak control of the Type I error rate. Strong control refers to control of the Type I error rate under any combination of true and false hypotheses, i.e., for any combination of differentially and constantly expressed genes. In contrast, weak control refers to control of the Type I error rate only

when none of the genes are differentially expressed, i.e., under the complete null hypothesis that all the null hypotheses are true. In general, weak control without any other safeguards is unsatisfactory. In the microarray setting, where it is very unlikely that none of the genes are differentially expressed, it seems particularly important to have strong control of the Type I error rate.

Adjusted p-values provide useful and flexible summaries of the strength of the evidence in favour of differential expression. The adjusted p-value for a particular gene reflects the overall false positive error rate for the family of hypotheses when genes with smaller p-values are declared differentially expressed. Adjusted p-values may also be used to summarize and compare the results from different multiple testing procedures.

In their 1993 book, Westfall & Young (1993) proposed resampling-based p- value adjustment procedures which are highly relevant in the context of microarray experiments. In particular, these authors defined adjusted p-values for multiple testing procedures which control the family-wise error rate and take into account the dependence structure between test statistics (their $\min P$ and $\max T$ adjusted p-values). In Dudoit et al (2001b) these ideas are applied in the context of microarray data. It is clear that this area is undergoing rapid development, see the recent review Ge et al (2003).

5.6. **Classification and clustering.** Microarray experiments have revived interest in both cluster and discriminant analysis, by raising new methodological and computational challenges. In discriminant analysis, also called supervised learning or class prediction, we might have observations on tumor mRNA samples known to belong to prespecified classes, and the task is to build predictors for allocating new observations to these classes. By contrast, in cluster analysis, also called unsupervised learning or class discovery, the classes are unknown a priori and the task is to determine these classes from the data themselves, i.e., to determine the number of classes and assign each observation to one of these classes. Either experiments or genes or both can be clustered, and the commonest approach uses hierarchical procedures based on correlation as a measure of dissimilarity. Clustering of this kind is currently the most popular way of analysing gene expression data, undoubtedly because of the power the technique to group co-expressed genes and hence shed light on the function of uncharacterized genes. For some examples see Eisen et al (1998), Bassett Jr et al (1999), and Alizadeh et al (2000).

The ability to successfully distinguish between tumor classes (already known or yet to be discovered) using gene expression data is an important aspect of this novel genomic approach to cancer classification. There are already many papers on this topic, and almost every technique from the field of machine learning has already been applied

to this problem. How do they compare? Are there advantages to the more recent or more elaborate classification techniques? While it is not possible to give a single long-term answer to this question, it is possible to obtain some insights. The study Dudoit et al (2002a) compared a number of familiar methods for classifying tumors based on gene expression data, including nearest neighbor classifiers, linear discriminant analysis, and classification trees. Two recent machine learning devices known as bagging and boosting were also considered. The discrimination methods were all applied to datasets from three recently published cancer gene expression studies, and the main conclusion, for these datasets, was that simple classifiers such as diagonal linear discriminant analysis and nearest neighbors performed remarkably well compared to more elaborate ones such as aggregated classification trees. These conclusions may change as the size of data sets grows.

5.7. **Other topics.** When expression levels are measured for thousands of genes in time and in space, a challenging problem is to discover and recognize reproducible temporal expression patterns, including ones not previously known. Current approaches to this class of questions with microarray data are rather ad hoc, usually involving one or two-dimensional clustering methods. These methods, typified by Eisen's "heat diagrams" (Eisen et al, 1998), rearrange the order of genes and experiments to map the data onto a plane in a more visually compelling way. The hope is that visual examination of the resulting image will identify patterns to which explanations can be attached. Other researchers rely on multi-dimensional scaling, which uses distances between genes or arrays to produce a scatter plot in the plane for subsequent visual examination. There is a clear need for more biostatistical research on problems like this. For yet other topics, and a overview of the area from a statistical point of view, see Speed (2003), where a much fuller bibliography can be found.

Acknowledgement. This article is a slightly modified and updated version of one which originally appeared in Biostatistical Genetics and Genetic Epidemiology (editors R.C. Elston, J.M. Olson and Lyle Palmer, Wiley) under the title "Gene Expression Analysis".

References

[1] Alizadeh, A. A., Eisen, M. B., Davis, R. E. et al. Different types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503-511.

[2] Audic, S. and Claverie, J-M. (1997) The significance of digital gene expression profiles. Genome Research 7, 986-995.

[3] Bassett Jr, D. E., Eisen, M. B., and Boguski, M S. (1999) Gene expression informatics - its all in your mine. Nature Genetics 21, Supplement, 51-55.

[4] Bittner, M. L., Chen, Y., Dorsel, A.N. and Dougherty, E.R. eds (2001) Microarrays: Optical technologies and informatics. Progress in Biomedical Optics and Imaging Vol 2, No. 23. Proceedings of SPIE Vol 4266.

[5] Bolstad BM, Irizarry RA, Astrand M, Speed TP. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19(2):185-93.

[6] Brandle, N., Bischof, H. and Lapp, H. (2001) A generic and robust approach for the analysis of spot array images In Bittner et al (2001): 1-12.

[7] Buhler, J., Ideker, T.,and Haynor, D. (n.d.) Dapple: Improved techniqes for finding spots of dna microarrays. Technical Report, Department of Molecular Biotechnology, University of Washington, Seattle.

[8] Carlisle, A. J., Prabhu, V. V., Elkahloun, A., Hudson, J., Trent, J. M., Linehan, W. M., Williams, E. D., Emmert-Buck, M. R., Liotta, L. A., Munson, P. J., and Krizman, D. B. (2000) Development of a Prostate cDNA Microarray and Statistical Gene Expression Analysis Package Molecular Carcinogenesis 28:12-22 (2000)

[9] Dudoit, S., Fridlyand, J. and Speed, T. (2002a) Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association. 457 (97):77-87.

[10] Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2002b) Statistical methods for identifying genes with differential expression in replicated microarray experiments. Statistica Sinica 12:111-140.

[11] Efron, B., Tibshirani, R., Goss, V. amd Chu, G. (2000). Microarrays and their use in a comparative experiment. Technical Report, Stanford University Department of Statistics.

[12] Eisen, M. B.. Spellman, P. T., Brown, P.O., and Botstein, D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the USA 95, 14863-14868.

[13] Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, and Solas D (1991) Light-directed, spatially addressable parallel chemical synthesis. Science 251, 767-73.

[14] Ge, Y., Dudoit, S. and Speed, TP. (2003) Resampling-based multiple testing for microarray data analysis Department of Statistics, University of California at Berkeley, Tech. Rep. 633. To appear (with discussion) in Test.

[15] Golub, T. R., Slonim, D. K., Tamayo, P., et al Molecular classification of cancer: class discovery and class prediction by gene expression profiling. Science 286, 531-7.

[16] Higuchi, R., Fockler, C., Dolinger, G., and Watson, R. (1993) Kinetic PCR: Real time monitoring of DNA amplification reactions. Biotechnology 11, 1026-1030. Hughes TR, Mao M, Jones AR et al (200i) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nature Biotechnology 19, 342-347.

[17] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research 31(4):e15.

[18] Kerr MK, Churchill GA (2001) Experimental design got gene expression microarrays. Biostatistics 2, 183-201.

[19] Li, C and Wong, W. H. (2001a) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection Proceedings of the National Academy of Sciences of the USA 98: 31-36.

[20] Li, C. and Wong, W. H. (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biology 2, 32.1-32.11.

[21] Lockhart, D. J., Dong, H. L., Byrne, M.C. et al Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology 14, 1675-1680.

[22] Man, M. Z., Wang, X. and Wang, Y. (2000) POWERSAGE: comparing statistical tests for SAGE experiments. Bioinformatics 16, 953-959. Meier-Ewert, S., Lange, J., Gerst, H et al (1998) Comparative gene expression profiling by oligonucleotide fingerprinting Nucleic Acids research 26 2216-2223.

[23] Nature (1999) The Chipping Forecast Supplement to Nature Genetics, 21. See also: http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal /v21/n1s/index.html

[24] Nature (2002) The Chipping Forecast II Supplement to Nature Genetics 32. pp 461-562. See also: http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal /v32/n4s/index.html

[25] PE Applied Biosystems (1997) User Bulletin #2 36pp. PE Applied Biosystems (1998) User Bulletin #5 20pp.

[26] Pfaffl, M. W. (2001) A new mathematical mnodel for relative quantification in real-time RT-PCR. Nucleic Acids Research 29, 2002-2007. Roche Molecular Biochemicals (2000) Technical Note No LC/10.

[27] Schadt, E., Li, C., Su, C. and Wong, W.H. (2000) Analyzing high-density oligonucleotide gene expression data Journal of Cellular Biochemistry 80, 192-202. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct 20;270(5235):467-70.

[28] Schuchhardt, J., Beule, D., Mailk, A., Wolski, E., Eickhoff, H., Lehrach, H and Herzel, H. (2000) Normalization strategies for cDNA microarrays. Nucleic Acids Research 28, E47.

[29] Speed, T. (Editor) Statistical Analysis of Gene Expression Microarray Data. 2003. CRC Press.

[30] Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied ot transcriptional response to ionizing radiation. Proceedings of the National Academy of Sciences of the USA 98, 5116-5121.

[31] Westfall, P. H. and Young, S. S. (1993) Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons.

[32] Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001a) Normalization for cDNA microarray data In Bittner et al (2001): 141-152. Yang, Y.H., Buckley, M. J., Dudoit, S., and Speed, T. P. (2001b) Image processing on cDNA microarray data. Journal of Computational and Graphical Statistics. In press.

[33] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. (2002) Normalization for cDNA microarray data: a robust composite method addressin single and multiple slide systematic variation. Nucleic Acids Research 30(4):e15.

[34] Yang YH and Speed T (2002) Design issues for cDNA microarray experiments. Nat Rev Genet. 3(8):579-88.