Chapter 15

Microarray Data Analysis

Jane Fridlyand

I met Terry when I was a beginning graduate student in the Department of Statistics at UC Berkeley. The first year is the time when bright-eyed and idealistic graduate students start thinking about what they want to do for the next 30 years of their lives, or at least until they are handed their PhD diploma and a job offer from Wall Street. I was in awe of Terry, but gathered my courage to approach him with that crucial question: “Would you work with me?” Now that I write this, I find that it sounds rather like a marriage proposal. And indeed, it becomes one: a covenant of commitment between a student and the mentor, with all the ups and downs, for better or for worse, lasting a lifetime. I wanted to work with Terry because he inspired me, as a scientist and as a person, and his interests in biological and medical applications were close to my heart. I also had reason to hope for a positive response – I was told in confidence by several people that in his 20 years of working with students Terry had never turned anyone down. So, here I was asking “Would you work with me?”. His reply was immediate and crushing: “Why?”. I did not know what to say – with all the mental rehearsals I had done, I was not prepared for this comeback. I must have blushed, mumbled something and run away. I guess there is always a first one to be turned down, unfortunately it just happened to be me!

My despair did not last long. The next day I found a thick stack of papers on statistical genetics and schizophrenia research in my mailbox with a note asking me to read them and meet Terry the next day at a specific time to discuss. And this is how our work together began. Although I transitioned from schizophrenia research, our working relationship had been established, and Terry has remained a very important part of my life since (15 years and counting!).

J. Fridlyand
Department of Biostatistics, Genentech
e-mail: fridlyand.jane@gene.com

Microarrays and high-dimensional data

Starting in the late nineties, the field of applied statistics in biomedical research has transformed from the traditional paradigm of many samples and few variables to a situation that had not been greatly considered before by statisticians outside of the machine learning community – one of few samples and an enormous number of variables, also known as the “small n , large p ” problem. Unlike the past when existing theory and methods foreshadowed (or even dictated) data types that would occur in practice, this time the technology came first along with excitement and great promise. cDNA microarrays and high-density oligonucleotide chips allowed measurement of many thousands, and eventually, millions, of gene products simultaneously. High-density SNP (Single Nucleotide Polymorphism) arrays enabled high-throughput genetic profiling of living organisms. Taken together and occurring in parallel with the ongoing human and other genome projects, these breakthroughs in technology generated exciting possibilities in biomedical research: human disease prognosis and classification, new drug targets, mammalian models, basic research, and, finally, the ability to conduct discovery experiments on a scale previously unimaginable. As new technologies were quickly adopted by researchers and clinicians, questions encompassing a broad spectrum of statistical issues arose, including:

- “How reliable and reproducible are the measurements?” (quality assessment and control)
- “Can I really find a needle in a haystack?” (experimental design, estimation, testing)
- “What can one do with so many variables at a time?” (modeling, prediction techniques)
- “How can I minimize the false leads?” (testing)

Re-formulating and addressing such questions falls in the purview of statisticians, who are able to draw on their knowledge of experimental design, prediction techniques, modeling, estimation and testing, and adapt and expand existing concepts to work with these new and unprecedentedly large datasets.

When I think of Terry’s approach to statistics and mentorship, a few quotes from Albert Einstein come to mind – *In theory, practice and theory are the same. In reality, they are not* and *Everything should be made as simple as possible, but not simpler*. These points could not have been more appropriate or timely than when excited statisticians, physicists, and computer scientists started working on high-dimensional biomedical problems.

It is difficult to overemphasize Terry’s contributions to the field of high-dimensional data analysis in biomedical research. He stepped in at the very start and, with vigor, rigor and great enthusiasm, began to transform the analytical methods used in the field. Generically, we can consider two levels of microarray data analysis: low-level analysis, which is concerned with preprocessing the raw data into meaningful and analyzable measures; and high-level analysis, which is the statistical analysis of the resulting data matrix. Most methodological researchers tend to specialize in

one or the other of these. Terry has made major, fundamental, and very widely used contributions across the analysis spectrum.

The May 8th 2011 PubMed search for “TP Speed” reveals that Terry has co-authored in excess of 150 peer-reviewed publications, a large number of which focus on the analysis of high-dimensional biological data. Here, I provide a historical commentary to only a few of the most ground breaking of those.

Your results will only be as good as the information you put in (more commonly known as “garbage in, garbage out”)

Perhaps the most widely cited microarray contributions from Terry have been from his work on low-level preprocessing of the measurements. Early on, it was recognized that there are many sources of systematic variation in both cDNA and oligonucleotide microarrays. Although understanding the underlying physical reasons for the observed variation is useful, it is not always feasible. Terry recognized that simple empirical normalization approaches may be competitive with more complex biophysical models. Terry also proposed a number of what are now among the most commonly used quality control visualization tools assuring that appropriate preprocessing has been done (e.g. MA-plots, chip pseudo-image plots). Finally, for a formal evaluation of preprocessing methods, relevant biological calibration experiments had to be designed and conducted.

Yang et al. [10] and Irizarry et al. [4] represent some of the papers describing revolutionary microarray normalization (preprocessing) techniques for cDNA arrays (lowess) and short oligonucleotide chips (RMA), respectively. RMA, or some subsequent variant of it, is the most frequently used and cited preprocessing technique for short oligonucleotide chips. Rabbee and Speed [7] describe a multi-chip, multi-SNP approach to genotype calling for Affymetrix SNP chips, providing the first such alternative to the standard (at the time) genotype calling procedures, which processed all the features associated with one chip and one SNP at a time.

Microarray data analysis $\neq$ clustering

In the very early days of microarray data analysis, probably due to the high dimensionality of the data, virtually all analyses included a cluster analysis – regardless of the scientific question under study (for which clustering may or may not be appropriate).

For the problem of identifying genes that are differentially expressed in two conditions, a more natural, statistically based approach would be to use the mean difference or standardized mean difference of the expression levels, separately for each gene. However, these statistics are problematic. A large mean difference may be due to large variability or an outlier. But taking account of the variability by using the

standardized difference is also problematic because when the number of replicates is small the estimate of variance is less reliable, and in particular may be artificially small. In this case, a small average difference can be highly statistically significant, yet biologically meaningless.

Lönnstedt and Speed [6] address this issue using an empirical Bayes approach that avoids these problems. They use a Bayes log posterior odds for differential versus equal expression to select differentially expressed genes. Tai and Speed [9] extend the model to allow for analysis of time-course microarray data.

Do complex datasets require complex methods?

A new laboratory technology without an established methodology for analysis of the resulting data may be attractive to aspiring quantitative analysts eager to apply new sophisticated analytical methods “brewing” in their labs yet lacking an exciting application. This situation violates a firm rule that Terry had for his students: it is the real life problems that motivate methodological research, not the reverse. Thus, when human cancer microarray datasets were first publicly released for re-analysis by other groups, Terry questioned whether the complex, state-of-the-art prediction methods that were being published with the aim of addressing biomedical research problems (e.g. prediction of a patient’s tumor subtype or treatment outcome) do indeed outperform more simplistic approaches that place tight restrictions on the parameter space, such as a linear discriminant analysis with diagonal covariance matrix. Another question that came up was how to measure the relative performance of multiple candidate predictors in the absence of true, independent datasets, particularly when many parameters are estimated.

These two issues are discussed in-depth in Dudoit et al. [2]. Somewhat surprisingly, the main conclusion of this work (later supported theoretically by Levina and Bickel [5]) was that for small n , large p problems the simplest methods, with the most restrictive assumptions, perform as well as or better than the latest machine learning approaches. Moreover, unbiased assessment of performance can be challenging and must be done through a careful and valid cross-validation – an important caveat ignored by several groups in early publications. In view of these results, it is not surprising that in high-dimensional genomics, the rate of independently validated predictions remains low. Nevertheless, much of the progress that has been made is due to Terry’s work on formulating and disseminating the appropriate message.

If you torture data enough, it will confess

Testing many thousands of genes for association with the phenotype of interest invariably presents an issue. From the statistical point of view, testing must be performed at an exceedingly stringent alpha level to control the overall number of false

positive findings. At the time of this writing, this idea seems obvious; however, even 5 years ago, it was not – a change of mindset was required as a great majority of papers reported the significance of individual tests without regard to the number of the comparisons performed. The initial discussion of permutation-based adjusted (rather than nominal) p -values took place in Callow et al. [1]; an extensive review of approaches to multiple testing was presented in Ge et al. [3].

And finally...

On a very personal note, I would like to conclude with the story of a paper that Terry and I have never written but work on which manifests in my mind many of the wonderful qualities that Terry possesses: inspiration, mentorship and willingness to always give one more chance. In 1998, I hit a creativity wall, a not uncommon occurrence in the life of a PhD student. Terry had many PhD students, but each of us was important to him as an individual. Terry brought me to Australia with him and there I was able to stumble on a topic that excited and reinvigorated me – the search for interactions in high-dimensional SNP studies. There we were able to utilize binary tree partitioning techniques to discover epistatic genes without prominent independent (marginal) effect, while at the same time illuminating an underlying interaction structure. The results summarizing the application of our approach are described in Symons et al. [8]; however, the methodological paper was never written. Nevertheless, this is our joint work for which I am most grateful to Terry, and that ultimately motivated me to start and finish my PhD dissertation. A lesson in this to all the mentors out there: do not give up on your students, and eventually you will be thanked in print!

References

- [1] M. J. Callow, S. Dudoit, E. J. Gong, T. P. Speed, and E. M. Rubin. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res.*, 10(12):2022–2029, 2000.
- [2] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, 97(457):77–87, 2002.
- [3] Y. Ge, S. Dudoit, and T. P. Speed. Resampling-based multiple testing for microarray data analysis. *TEST*, 12(1):1–44, 2003.
- [4] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

- [5] E. Levina and P. J. Bickel. Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.
- [6] I. Lönnstedt and T. P. Speed. Replicated microarray data. *Stat. Sinica*, 12: 31–46, 2001.
- [7] N. Rabbee and T. P. Speed. A genotype calling algorithm for Affymetrix SNP arrays. *Bioinformatics*, 22(1):7–12, 2006.
- [8] R. C. A. Symons, M. J. Daly, J. Fridlyand, T. P. Speed, W. D. Cook, S. Geron-dakis, A. W. Harris, and S. J. Foote. Multiple genetic loci modify susceptibility to plasmacytoma-related morbidity in $E\mu-v-abl$ transgenic mice. *Proc. Natl. Acad. Sci. USA*, 99(17):11299–11304, 2002.
- [9] Y. C. Tai and T. P. Speed. A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Stat.*, 34(5):2387–2412, 2006.
- [10] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, 30(4):e15, 2002.

Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation

Yee Hwa Yang¹, Sandrine Dudoit², Percy Luu³, David M. Lin³, Vivian Peng^{3,4}, John Ngai^{3,4,5} and Terence P. Speed^{1,6,*}

¹Department of Statistics, ²Division of Biostatistics, ³Department of Molecular and Cell Biology, ⁴Functional Genomics Laboratory and ⁵Neurogenomics Center, Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720-3860, USA and ⁶Division of Genetics and Bioinformatics, The Walter and Eliza Hall Institute, Melbourne, Australia

Received October 3, 2001; Revised November 20, 2001; Accepted December 1, 2001

ABSTRACT

There are many sources of systematic variation in cDNA microarray experiments which affect the measured gene expression levels (e.g. differences in labeling efficiency between the two fluorescent dyes). The term normalization refers to the process of removing such variation. A constant adjustment is often used to force the distribution of the intensity log ratios to have a median of zero for each slide. However, such global normalization approaches are not adequate in situations where dye biases can depend on spot overall intensity and/or spatial location within the array. This article proposes normalization methods that are based on robust local regression and account for intensity and spatial dependence in dye biases for different types of cDNA microarray experiments. The selection of appropriate controls for normalization is discussed and a novel set of controls (microarray sample pool, MSP) is introduced to aid in intensity-dependent normalization. Lastly, to allow for comparisons of expression levels across slides, a robust method based on maximum likelihood estimation is proposed to adjust for scale differences among slides.

INTRODUCTION

DNA microarrays are part of a new class of biotechnologies that allow the monitoring of expression levels in cells for thousands of genes simultaneously. In a typical microarray experiment utilizing 'spotted arrays', the two mRNA samples to be compared are reverse transcribed into cDNA, labeled using two different fluorophores (usually a red fluorescent dye, Cy5, and a green fluorescent dye, Cy3) and then hybridized simultaneously to the glass slide. Intensity values generated

from hybridization to individual DNA spots are indicative of gene expression levels, and comparisons in gene expression levels between the two samples are derived from the resulting intensity ratios (1). Applications of microarrays range from the study of gene expression in yeast under different environmental stress conditions (2,3) to the comparison of gene expression profiles for tumors from cancer patients (4–9).

In order to accurately and precisely measure gene expression changes, it is important to take into account the random (experimental) and systematic variations that occur in every microarray experiment. For example, a well-known source of systematic variation arises from biases associated with the different fluorescent dyes. This can most easily be seen in an experiment where two identical mRNA samples are labeled with different dyes and subsequently hybridized to the same slide (10). In this instance, it is rare to have the dye intensities equal across all spots between the two samples. Even though such systematic biases may be comparatively small, they may be confounding when searching for subtle biological differences. Dye biases can stem from a variety of factors, including physical properties of the dyes (heat and light sensitivity, relative half-life), efficiency of dye incorporation, experimental variability in hybridization and processing procedures, or scanner settings at the data collection step. Furthermore, the relative gene expression levels from replicate experiments may have different sample variances due to differences in experimental conditions. Many of these factors, whether internal or external to the target samples, make distinctions between differentially and constantly expressed genes difficult [in this article we adopt the definitions of 'probe' and 'target' from the January 1999 supplement to *Nature Genetics* (11), whereby the term target refers to the samples hybridized to the array and the term probe refers to the DNA sequences spotted on the array].

The purpose of normalization is to minimize systematic variations in the measured gene expression levels of two co-hybridized mRNA samples, so that biological differences can be more easily distinguished, as well as to allow the comparison of expression levels across slides. Current methods of normalization

*To whom correspondence should be addressed at: Department of Statistics, University of California, 367 Evans Hall, #3860 Berkeley, CA 94720-3860, USA. Tel: +1 510 642 2781; Fax: +1 510 642 7892; Email: terry@stat.berkeley.edu

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

e15 Nucleic Acids Research, 2002, Vol. 30, No. 4

fail to account for important sources of systematic variation (e.g. intensity- or spatially-dependent dye biases). In this article we propose a composite normalization procedure, based on robust local regression, to accommodate different types of dye biases and the use of control sequences spotted on the array. The selection of a suitable set of control spots for use in the normalization procedure is critical for proper normalization. To this end, we introduce a novel control sample (microarray sample pool, MSP), with minimal sample-specific bias over a large intensity range, and show that it is effective in many types of microarray experiments.

MATERIALS AND METHODS

Biological samples

Preparation of RNA samples and microarray analysis. Tissues were dissected, solubilized in Trizol (Gibco BRL) and total RNA was prepared according to the manufacturer's suggested protocol. Prior to reverse transcription and labeling, total RNA samples were treated with DNase using RQ RNase-free DNase (Promega) for 20 min at 37°C. RNA samples were reverse transcribed and labeled for microarray analysis using standard techniques (6,12). Briefly, RNA samples were reverse transcribed with Superscript II reverse transcriptase in the presence of 2-aminoadenyl-dUTP. Samples were purified and coupled to Cy3 or Cy5 as described (6,12,13). Labeled targets were resuspended in hybridization buffer and applied to glass microarrays. Hybridizations were performed overnight at 50–55°C. Washed and dried slides were imaged in an Axon GenePix 4000A scanner.

Experiment A: apolipoprotein AI (apo AI) experiment. The treatment group consisted of eight mice with the apo AI gene knocked out and the control group consisted of eight control C57BL/6 mice. For each of these 16 mice, target cDNA was obtained from mRNA by reverse transcription and labeled using a red fluorescent dye, Cy5. The reference sample used in all hybridizations was prepared by pooling cDNA from the eight control mice and was labeled with a green fluorescent dye, Cy3. Target cDNA was hybridized to microarrays containing 6384 cDNA probes, which included 257 genes thought to be related to lipid metabolism. Probes were spotted onto the glass slides using a 4 × 4 print head and each of the corresponding 16 print tip groups was laid out in a 19 × 21 array or sub-grid. For further details the reader is referred to Callow *et al.* (14).

Experiment B: olfactory bulb experiment. In this experiment, comparisons were made between different spatial regions of the mouse olfactory bulb to screen for possible region-specific differences in gene expression (D.M.Lin, Y.H.Yang, J.Scolnick, L.Brunet, V.Peng, T.Speed and J.Ngai, submitted for publication). The target cDNA was hybridized to glass microarrays containing ~18 000 isolated expressed sequence tags (ESTs) from the RIKEN Release 1 mouse cDNA library (15). The olfactory bulb is an ellipsoidal structure, so in order to make a 3-dimensional representation using binary comparisons, bulbs were separately sub-dissected into three sections along each of the three orthogonal axes. RNA was collected from a number of different mice and samples from the same anatomical

domains were harvested and pooled (D.M.Lin, Y.H.Yang, J.Scolnick, L.Brunet, V.Peng, T.Speed and J.Ngai, submitted for publication). Comparisons were made between maximally separated regions: anterior versus posterior, medial versus lateral and dorsal versus ventral regions.

MSP titration series. Total EST collections were generated from amplification of PCR products for microarray fabrication. Samples corresponding to all 18 816 ESTs from the RIKEN Release 1 cDNA library were pooled and precipitated. An MSP was also made from a randomly picked non-normalized plasmid library generated from mouse cerebellum (A.Finn and T.Serafini, unpublished results). Precipitated samples were resuspended and serially diluted in preparation for printing. Six steps were used in the dilution series and the samples were then spotted in the middle of the first or last row of each of the print tip groups. Microarrays were prepared as discussed previously (6).

Image processing

Each hybridization produced a pair of 16-bit images, which were processed using the software package Spot (16). The main quantities of interest produced by the image analysis methods (segmentation and background correction) are the (R, G) fluorescence intensity pairs for each gene on each array (where R = red for Cy5 and G = green for Cy3). Note that we call the spotted DNA sequences 'genes', whether they correspond to actual genes, ESTs or DNA sequences from other sources.

Statistical methods

An 'MA-plot', as described in Dudoit *et al.* (10), is used to represent the (R, G) data, where $M = \log_2 R/G$ and $A = \log_2 \sqrt{(R \times G)}$. We have found MA-plots to be helpful in terms of identifying spot artifacts and detecting intensity-dependent patterns in the log ratios M . They are also very useful for the purpose of normalization, as illustrated next with several location normalization procedures. Within-slide normalization for location consists of subtracting a function $c(\cdot)$ from individual intensity log ratios, where the function $c(\cdot)$ is computed separately for each slide, using only data from that hybridization.

Global normalization. Global methods assume that the red and green intensities are related by a constant factor, i.e. $R = kG$, and the center of the distribution of log ratios is shifted to zero

$$\log_2 R/G \rightarrow \log_2 R/G - c = \log_2 R/(kG)$$

A common choice for the location parameter $c = \log_2 k$ is the median or mean of the intensity log ratios M for a particular gene set.

Intensity-dependent normalization. We use the robust scatter plot smoother 'lowess', implemented in the statistical software package R (17), to perform a local A -dependent normalization

$$\log_2 R/G \rightarrow \log_2 R/G - c(A) = \log_2 R/[k(A)G]$$

where $c(A)$ is the lowess fit to the MA-plot. The lowess scatter plot smoother performs robust locally linear fits. In particular, it will not be affected by a small percentage of differentially expressed genes, which will appear as outliers in the MA-plot. The user-defined parameter f is the fraction of the data used for smoothing at each point; the larger the f value, the smoother the fit. We typically use $f = 40\%$.

Within-print tip group normalization. Within-print tip group normalization is simply a (print tip + A)-dependent normalization, i.e.

$$\log_2 R/G \rightarrow \log_2 R/G - c_i(A) = \log_2 R/[k_i(A)G]$$

where $c_i(A)$ is the lowest fit to the MA-plot for the i th grid only (i.e. for the i th print tip group), $i = 1, \dots, I$, and I denotes the number of print tips.

Scale normalization. Starting from data which have been location normalized as just described, we suppose that the log ratios from the i th print tip group follow a normal distribution with mean zero and variance $a_i^2 \sigma^2$, where σ^2 is the variance of the true log ratios and a_i^2 is the scale factor for the i th print tip group. In order to perform scale normalization, the scale factors a_i for the different print tip groups are estimated and then eliminated. Enforcing the natural constraint $\sum_{i=1}^I \log a_i^2 = 0$, with I denoting the total number of print tip groups on the array (or the number of slides, for multiple slide normalization discussed below), the maximum likelihood estimate for a_i is

$$\hat{a}_i^2 = (\sum_{j=1}^{n_i} M_{ij}^2) / [I \sqrt{(\prod_{k=1}^I \sum_{j=1}^{n_k} M_{kj}^2)}]$$

where M_{ij} denotes the j th log ratio in the i th print tip group, $j = 1, \dots, n_i$. A robust alternative to this estimate, which we find preferable, is

$$\hat{a}_i = (MAD_i) / [I \sqrt{(\prod_{k=1}^I MAD_k)}]$$

where the median absolute deviation MAD is defined by

$$MAD_i = \text{median}_j \{ |M_{ij} - \text{median}(M_{ij})| \}$$

Composite normalization. For a given print tip group the composite normalization curve is a weighted average of the MSP lowess curve and the lowess curve based on all genes in the print tip group. The weights are dependent on the cumulative number of genes at different intensity levels A . An outline of this procedure for a spot in the i th print tip group is as follows. (i) Estimate $\hat{f}_i(A)$, the lowess fit to the MA-plot for the i th print tip group. (ii) Estimate $\hat{g}(A)$, the lowess fit to the MA-plot using only spots from the MSP titration series. (iii) Calculate the weighted average, $c_i(A) = \alpha_A \hat{g}(A) + (1 - \alpha_A) \hat{f}_i(A)$, where α_A is defined as the proportion of genes less than a given intensity A .

Comparison between different normalization methods. After image processing and normalization, the gene expression data can be summarized by a matrix X of intensity log ratios $M = \log_2 R/G$, with p rows corresponding to the genes being studied and n columns corresponding to the different hybridizations. In the apo AI experiment $p = 6384$ and there were $n_1 = 8$ control (C57Bl/6 mice) and $n_2 = 8$ treatment (apo AI knockout mice) hybridizations. Differentially expressed genes were identified by computing two-sample Welch t -statistics. For gene j the t -statistic comparing gene expression in the control and treatment groups is

$$t_j = (\bar{x}_{2j} - \bar{x}_{1j}) / \sqrt{[(s_{1j}^2/n_1) + (s_{2j}^2/n_2)]}$$

where $\bar{x}_{1j}$ and $\bar{x}_{2j}$ denote the average background-corrected and normalized expression level of gene j in the eight control and eight treatment hybridizations, respectively. Similarly, s_{1j}^2 and s_{2j}^2 denote the variances of gene j expression level in the control and treatment hybridizations, respectively. Large absolute t -statistics suggest that the corresponding genes have different expression levels in the control and treatment groups. The statistical significance of the results was assessed based on P -values adjusted for multiple comparisons. These adjusted

P -values were estimated by permutation, using Westfall and Young's step-down adjusted P -value procedure in algorithm 4.1 (18). The analysis of the apo AI experiment is described in detail in Dudoit *et al.* (10).

In order to compare the different within-slide normalization procedures, we considered their effect on the location and scale of the log ratios M using box plots. A Gaussian kernel density estimator (the 'density' function from the statistical software package R, bandwidth size 0.17) is also used to produce density plots of the log ratios for each of the normalization methods. For experiment A we considered the effect of the normalization procedures on the t -statistics for the knockout gene.

RESULTS

Within-slide normalization: intensity- and spatially-dependent systematic error

We first address within-slide normalization, i.e. normalization issues associated with data obtained from a single slide. A well-known source of error can be attributed to biases linked to the different dyes used at the labeling step. Current methods of global normalization assume a uniform grading of systematic error across all variables in an experiment. Two major assumptions are usually made: (i) all cDNA species within a sample will incorporate an equivalent amount of dye per mole cDNA; (ii) there are no other variables (e.g. spatial location, overall intensity, plate) that contribute to dye biases across the slide. These assumptions are too simplistic to account for the multiple sources of systematic error typically encountered in microarray experiments. The problem is best illustrated in an experiment where identical mRNA samples are labeled with Cy3 and Cy5 and subsequently hybridized to the same slide [self-self comparison; described in Dudoit *et al.* (10)]. In a 'perfect' self-self hybridization the intensity log ratios M in an MA-plot should be evenly distributed around zero across all intensity values A . However, this is rarely the case, and systematic error often manifests itself in terms of non-zero log ratios M . Furthermore, the imbalance in the red and green intensities is usually not constant across the spots and can vary according to overall spot intensity A (indicated by a curvature in the MA-plot), location on the array, plate origin and possibly other variables.

Intensity-dependent dye bias can be seen in the apo AI experiment (14). Apo AI is a gene known to play a pivotal role in high-density lipoprotein metabolism. The goal of the experiment was to identify genes with altered expression in the livers of mice with the apo AI gene knocked out compared with inbred C57Bl/6 control mice. In this instance, it was found that the vast majority of genes examined on the microarray showed no difference in expression level. The clear curvature in the MA-plot in Figure 1A strongly suggests the existence of an intensity-dependent dye bias.

Some systematic differences may exist between the print tips, such as slight differences in the length or in the opening of the tips, and deformation after many hours of printing. We therefore also performed individual lowess fits within each print tip group. The arrays in the apo AI experiment were printed with a 4×4 print head, so each lowess fit in Figure 1 corresponds to spots printed with a single print tip. Four

Figure 1. Within-slide normalization. (A) MA-plot demonstrating the need for within-print tip group location normalization. (B) MA-plot after within-print tip group location normalization. Both panels display the lowess fits ($f = 40\%$) for each of the 16 print tip groups (data from apo AI knockout mouse number 8 in experiment A).

Figure 2. Within-slide normalization: box plots displaying the intensity log ratio distribution, for each of the 16 print tip groups before and after different normalization procedures. The array was printed using a 4×4 print head and the print tip groups are numbered first from left to right, then from top to bottom, starting from the top left corner (data from apo AI knockout mouse number 8 in experiment A). (A) Before normalization, but before scale adjustment. (C) After within-print tip group location and scale normalization.

within-print tip group lowess curves stand out from the remaining 12 curves, indicating strong print tip or spatial effects. These four curves correspond to the last row of print tips in the 4×4 print head (print tips 13–16). This pattern was visible in the raw images, where the bottom four grids tended to have a higher red signal. We further examined the spatial effects by considering box plots of the log ratios M for each print tip group. Figure 2 shows that print tip groups 13–16 have a larger spread in their log ratios than any of the other 12 print tip groups. Such a difference in spread may result in misidentification of genes that are differentially expressed in the knockout mice compared to the control mice. Thus, normalization for scale across print tip groups seems desirable here.

Within-slide normalization using the majority of genes on the microarray

For the apo AI experiment considered in Figures 1 and 2, global normalization, in which a constant adjustment is used to force the distribution of the log ratios to have a median zero within each slide, would result in a vertical translation of the MA-plot. It would not correct for intensity- or spatially-dependent effects, including local differences in the spread of the log ratios M . As a first pass towards eliminating intensity and spatial biases, we considered a normalization procedure in which the majority of genes on the array are used for normalization.

This is a reasonable assumption when there are good reasons to expect that (i) only a relatively small proportion of the genes will vary significantly in expression between the two co-hybridized mRNA samples or (ii) there is symmetry in the expression levels of the up/down-regulated genes. The data shown in Figures 1 and 2 are good examples of this situation.

To address both intensity and spatial normalization issues, we first incorporated an intensity modifier into our normalization procedures. We used the scatter plot smoother lowess to produce robust location estimates of the intensity log ratios M for various intensity levels A and to adjust each gene with a different normalization value depending on its overall intensity. Other variables that may contribute to systematic bias include differences in print tips and spatial location. Because every grid in an array is printed using the same print tip, print tip groups can also be used as proxies for spatial effects on the slide. Thus, we also incorporated a print tip modifier into the intensity-dependent normalization. It might be thought that the layout of genes on the slide could lead to one or more print tip groups being enriched for differentially expressed genes and, hence, invalidate the assumption underlying print tip group normalization. While we cannot rule out chance imbalances in the spatial distribution of differentially expressed genes, the mechanics of spotting cDNA onto the slide makes a large effect of this kind unlikely. Even if one had a collection of

Figure 3. Within-slide normalization: MA-plot for comparison of the anterior versus posterior portion of the olfactory bulb. These samples are very similar and we do not expect many genes to change. The cyan dots represent the MSP titration series and the cyan curve represents the corresponding lowest fit. The red curve corresponds to the lowest fit for the entire dataset. Control genes are highlighted in yellow (tubulin and GAPDH), green (mouse genomic DNA) and orange (an approximate rank-invariant set of genes with $P = 0.01$ and $I = 25$). (Left) MA-plot before normalization. (Right) MA-plot after within-print tip group location normalization.

genes known or expected to be differentially expressed in one or more microtiter plates, they would be spotted evenly across the slide by the printer.

In principle, after within-print tip group location normalization, the log ratios from the different print tip groups should be centered around zero (Fig. 2B). However, it is possible that the log ratios from the various print tip groups have different spreads; if this is the case, a scale adjustment may be required. Figure 2 displays box plots of the intensity log ratios M for a slide in experiment A, before normalization (Fig. 2A), after within-print tip group location normalization (Fig. 2B) and after within-print tip group location and scale normalization (Fig. 2C). In Figure 2B there is a disproportionately large number of extreme log ratios in the lower four grids. After scale normalization, the extreme log ratios are evenly distributed on the array (Fig. 2C). Again, this procedure assumes that a relatively small proportion of the genes vary significantly in expression between the two co-hybridized mRNA samples, as would be expected when comparing samples from wild-type mice versus mice harboring a mutation in a single gene. In addition, it is assumed that the spread of the distribution of the log ratios should be roughly the same for all print tip groups. The robust statistic MAD , like the robust lowess smoother, will not be affected by a small percentage of differentially expressed genes, which will appear as outliers in the MA-plots.

In another example of within-slide location normalization, Figure 3 shows an MA-plot from experiment B, for a comparison of mRNA levels in the anterior and posterior portions of the mouse olfactory bulb. These mRNA samples are biologically very similar and very few genes are expected to be differentially expressed. Indeed, the MA-plot shows very little divergence from the lowest fit to all genes and the scatter plot is roughly symmetrical about the lowess curve. We thus performed a print tip group normalization using all genes. Figure 3 displays the intensity data before (Fig. 3, left) and after normalization (Fig. 3, right). The normalization procedure resulted in a scatter plot centered around an M value of zero across the A intensity range, thus indicating that the types of systematic errors we have identified have been minimized.

Within-slide normalization using MSP

Frequently, the expression profiles in biological samples are more divergent in nature than in the examples investigated above. Thus, normalization based upon all genes may be inaccurate. A control sample that spans the intensity range and exhibits a relatively constant expression level across biological samples is desirable. Yeast genomic DNA has been used for normalization in that system. Since all species within an mRNA sample can hybridize to this control, sample-specific bias is reduced. The genomic DNA approach does not, however, directly extend to more complex metazoan systems, where the high ratio of non-coding to coding DNA effectively reduces the signal from such a control below the detection threshold in a microarray experiment.

We therefore constructed a novel control sample ensemble, MSP, inclusive of all genes present on the microarray. This sample should be analogous to genomic DNA without the intervening sequences and, thus, provides a potential probe for every species within a labeled cDNA target. We titrated this sample over the intensity range of a typical microarray experiment in order to account for all levels of intensity-dependent bias. The utility of this control is demonstrated in Figure 3, which highlights the MSP titration series (cyan dots) and the corresponding lowess fit to the MSP spots (cyan curve). Notice that the MSP curve is the same as the lowess fit to the MA-plot based on all genes (red curve). An intensity-dependent normalization using the MSP control as a reference would thus be similar, in this case, to that using all genes.

In experiment B we made a more divergent comparison between mRNA samples from the medial and lateral portions of the olfactory bulb. Due to the presence of vascular tissue near the medial bulb, medial samples have a higher representation of blood tissue. Figure 4A displays the MA-plot for the medial versus lateral comparison. The genetic divergence between the samples is evident in the increased spread of the log ratios, particularly in the high intensity range. The lowess curve based on all genes (red) and the lowess curve based on the MSP titration series (cyan) are different at high intensity values. In such a case, where samples are widely divergent at high intensities, normalization based on the MSP titration series appears to be more accurate. However, whereas the MSP

Figure 4. Within-slide normalization: MA-plot for comparison of the medial versus lateral portion of the olfactory bulb. The cyan dots represent the MSP titration series and the cyan curve represents the corresponding lowest fit. The red curve corresponds to the lowest fit for the entire dataset. The green curve represents the composite normalization curve. Control genes are highlighted in yellow (tubulin and GAPDH), green (mouse genomic DNA) and orange (an approximate rank-invariant set of genes). (A) MA-plot before normalization. (B) MA-plot after composite normalization.

spots produce more accurate estimates of expression levels, these estimates may be less stable in the context of spatial normalization, due to the small number of MSP spots per print tip group.

Comparisons of MSP to other control samples

In some instances a small number of known genes, for example housekeeping genes whose expression is expected to be constant across samples, are utilized for microarray normalization. Such genes are often highly expressed, as illustrated in Figure 3 for tubulin and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (yellow). Typically, housekeeping genes are not representative of all intensity values A and are therefore limited in their utility for intensity-dependent normalization. In addition, there is a sample-specific bias for many genes which may not be predictable; this is again non-ideal for use as a control.

Another approach is to select a rank-invariant set of genes. A set of genes is said to be rank-invariant if their ranks are the same for the red and green intensities. In practice, a maximal invariant set tends to be too small and an iterative procedure for finding an approximately invariant set of genes has been proposed (19,20). These genes are highlighted in orange in Figure 3 and were obtained using the method described in Tseng *et al.* (20) with $P = 0.01$ and $l = 25$. The value P is chosen such that a conserved set of genes is selected. Notice that this set of spots overlaps the lowest fit to the MA-plot based on all genes.

Composite within-slide normalization

We propose a composite normalization method to address the limitations of using all genes or only the MSP titration series for normalization. The composite normalization curve is a weighted average of the MSP curve and the within-print tip group lowest curve based on all genes. The weights are dependent on the cumulative number of genes at different intensity levels. Figure 4B shows the MA-plot after within-print tip group normalization. In this figure the green composite normalization curve is a weighted average of the red and cyan colored curves. Note that the divergence of the red from the green curve at high intensity values still persists after normalization. In practice, we find composite normalization necessary in the case of divergent samples. Biologically

significant outliers in experiment B were more consistently identified when composite normalization was incorporated in the analysis (data not shown).

Comparison between different normalization methods

In order to compare the different within-slide normalization methods, we considered their effect on the location and scale of the log ratios M . Figure 5A shows density plots of the log ratios for different normalization methods. Without normalization (black curve) the log ratios are centered around -1 , indicating a bias towards the green (Cy3) dye. A global median normalization (red curve) shifts the center of the log ratio distribution to zero, but does not affect the spread. The dependence of the log ratio M on the overall intensity A is also still present (see Fig. 1). Both the intensity-dependent (green curve) and within-print tip group (blue curve) location normalization methods reduce the spread of the log ratios compared to a global normalization. A within-print tip group scale normalization (cyan curve) further reduces the spread slightly.

The different methods were also evaluated based on their ability to identify genes which are known to be differentially expressed. For experiment A the apo AI gene is knocked out in the eight treatment mice, so one expects the t -statistics to take on very large negative values for this gene. Figure 5B shows a truncated plot of the extreme t -statistics for each of the methods. The global median, intensity-dependent and within-print tip group location normalization methods seem to perform best in terms of their ability to detect the three copies of the knocked out apo AI gene. A good method should enable a clear distinction between differentially and constantly expressed genes as reflected by the t -statistic, i.e. one expects a large jump in the t -statistic between the least extreme of the differentially expressed genes and the most extreme of the remaining genes. The largest jump in P -values is observed for within-print tip group location normalization. Thus, in the situation presented by experiment A, where log ratios from the different arrays have fairly similar spreads (see Fig. 2), within-print tip group location normalization enables the best separation between differentially expressed genes and noise.

Multiple slide normalization

Having addressed location and scale normalization issues within a slide, all normalized log ratios should be centered

Figure 5. Within-slide normalization. (A) Density plots of the log ratios M before and after different normalization procedures. The solid black curve represents the density of the log ratios before normalization. The red, green, blue and cyan curves represent the densities after global median normalization, intensity-dependent location normalization, within-print tip group location normalization and within-print tip group scale normalization, respectively (data from apo AI knockout mouse number 8 in experiment A). (B) Plot of t -statistics for different normalization methods. The numbers 1–8 represent the differentially expressed genes identified in Dudoit *et al.* (10) and confirmed using RT-PCR; indices 1–3 represent the three apo AI genes spotted on the array. Empty circles represent the remaining 6376 genes where no effect is expected. Only t values less than -4 are shown.

around zero. However, in many experiments expression levels must be compared across different slides. It is important to note that individual slides in a multiple slide comparison may need to be adjusted for scale when the different slides have substantially different spreads in their intensity log ratios. Failing to perform a scale normalization could lead to one or more slides having undue weight when averaging log ratios across slides. We can apply the principles used for within-slide print tip group scale normalization to multiple slide scale adjustment.

In practice, the need for scale normalization between slides will be determined empirically. Figure 6 displays box plots of the log ratios for each of the 16 slides in experiment A, after within-print tip group location and scale normalization. The box plots are centered at zero and have fairly similar spreads. In this instance we chose not to adjust for scale, as the noise introduced by a scale normalization of the different slides may be more detrimental than a small difference in scale.

DISCUSSION

Intensity data from microarray experiments are subject to a variety of random and systematic errors. This paper has introduced location and scale normalization methods for different types of cDNA microarray experiments and discussed different sets of control spots utilized in normalization. The location normalization procedure is based on robust local regression of the intensity log ratios on overall spot intensity and accounts for intensity and spatial dependence in the dye biases. A MSP titration series was constructed and used as a set of controls for normalization. The advantages of the MSP are the minimal sample-specific bias and the coverage of a wide intensity range. In addition, we have proposed a composite normalization procedure, whereby the utility of different sets of control spots and normalization methods are combined. The different normalization methods were compared using gene expression data from two experiments: the apo AI experiment (experiment

A), with replicated treatment and control slides, and the mouse olfactory bulb experiment (experiment B). Normalization can be performed at three different levels: (i) within a single slide; (ii) between a pair of slides for dye-swap experiments (21); and (iii) among multiple slides.

Within-slide normalization methods

For within-slide normalization, global methods have been used as pre-processing steps in a number of papers on the identification of differentially expressed genes in single slide cDNA microarray experiments (22,23). Such procedures assume that the red and green intensities can be related by a multiplicative constant. In one of the first proposed normalization methods, Chen *et al.* (22) derived an iterative procedure for estimating normalization constants. Similar approaches have been implemented in widely used microarray software packages [e.g. GenePix (24)]. Kerr *et al.* (25) and R.D.Wolfinger, G.Gibson, E.D.Wolfinger, L.Bennett, H.Hamadeh, P.Bushel, C.Afshari and R.S.Paules (SAS Institute, unpublished data) proposed the use of ANOVA models for normalization purposes. Their methods essentially perform only a global normalization and do not correct for intensity or scale differences. We have found that the standard global median normalization can often be inadequate due to spatially- and intensity-dependent dye biases. We propose instead a within-print tip group location normalization method which is based on robust local regression of the log ratios M on overall spot intensity A (the lowess smoother for MA-plots). Compared with other normalization procedures, this approach provided a clearer distinction between the differentially and constantly expressed genes in experiment A.

Other intensity-dependent normalization methods have been proposed in recent articles. Finkelstein *et al.* (26) recommended an iterative linear regression procedure, which essentially amounts to robust linear regression. Sapir and Churchill (27) suggested using the orthogonal residuals from the robust regression of $\log R$ versus $\log G$ as the normalized log ratios.

e15 Nucleic Acids Research, 2002, Vol. 30, No. 4

Figure 6. Multiple slide normalization: box plots displaying the intensity log ratio distribution for different slides/mice for experiment A, after within-print tip group location and scale normalization. The first eight box plots represent the data for the eight control mice and the last eight represent the data for the eight apo AI knockout mice.

Since an MA-plot amounts to a 45° counterclockwise rotation of the $(\log G, \log R)$ coordinate system (up to multiplicative constants), their method is similar to fitting a robust regression line through the MA-plot, instead of a lowess curve. One can view these two linear normalizations as a more constrained version of our intensity-dependent normalization. Kepler *et al.* (28) proposed a more general intensity-dependent normalization approach, which uses a different local regression method instead of the lowess smoother. Most methods suggested thus far do not correct for spatial biases in the log ratios. As we have shown, spatial bias is also a significant source of systematic error, due to hybridization artifacts or print tip effects during printing of the microarray. Our proposed normalization procedures correct for these artifacts.

Within-slide location normalization methods adjust the intensity log ratios M such that they are approximately zero for genes that are constantly expressed in the two co-hybridized samples. The box plots of the location normalized log ratios in each print tip group in Figure 2 suggest that some scale adjustment may also be required within slide. However, within-print tip group scale normalization seems to have decreased our ability to identify the differentially expressed genes in experiment A. We believe that this is due to an increase in the variability (the denominator of the t -statistic) of the log ratios for the eight differentially expressed genes compared to the rest of the genes.

Multiple slide normalization methods

A similar approach to that described for within-slide scale normalization may also be extended to perform scale normalization across slides. In practice, multiple slide normalization aims to adjust for different sample variances in log ratios across slides. Such adjustments are required so that the relative expression levels from one particular slide do not dominate the average relative expression levels across replicate slides. In general there is a trade-off between the gains achieved by scale normalization and the possible increase in variability introduced by this additional step. In cases where the scale differences are fairly small it may thus be preferable to perform only a location normalization. Further investigations are underway to develop an improved procedure for scale adjustment and to identify better comparison criteria to assess the effectiveness of various normalization procedures.

Comparisons of MSP to commonly used control samples

In general, the set of control spots most appropriate for normalization depends on the nature of the experiment. Traditional methods based upon intensity values of housekeeping genes often show sample-specific bias and do not address the issue of intensity-dependent dye biases. Other drawbacks include the possibility that housekeeping genes may actually be regulated within an experimental sample. Housekeeping genes also tend to be highly expressed and, hence, may not be

Table 1. The various normalization methods considered in this article

		Within-slide	Multiple slide			
		Global, location	Intensity-dependent, location	Print tip-dependent, location	Print tip, location and scale	Scale
		$c(.)$ constant, $a(.) = 1$	$c(.) = c(A)$, $a(.) = 1$	$c(.) = c(A, \text{print tip})$, $a(.) = 1$	$c(.) = c(A, \text{print tip})$, $a(.) = a(\text{print tip})$	
All genes	Assumes the majority of genes in the two mRNA samples have similar overall expression levels	Yes	Yes	Yes	Yes	Yes
Housekeeping genes	Usually highly expressed and do not capture intensity-dependent structure	Yes	No	No	No	No
MSP titration series	Doesn't require any prior biological assumption, however, estimating $c(A, \text{print tip})$ based on a small number of spots may not be very stable	Yes	Yes	No	No	No
Rank-invariant set	May not span the whole intensity range	Yes	Yes	No	No	No

For within-slide normalization, the log ratios are normalized by $\log_2 R/G \rightarrow [\log_2 R/G - c(.)]/a(.)$, where $c(.)$ and $a(.)$ correspond to location and scale adjustment, respectively. The columns refer to different normalization methods and the rows correspond to different sets of control spots. The Yes or No in each cell refers to the feasibility of performing the normalization in practice. For example, it is possible in practice to perform global normalization based only on housekeeping genes, but it is not advisable to perform intensity-dependent normalization on housekeeping genes only.

representative of other genes of interest. It is clear that a less localized type of control is required to obtain accurate normalization. The other three types of control examined in this article were chosen for their representation of as many genes and intensity values as possible so as to minimize sample bias.

The MSP titration series was constructed with this specific aim in mind. In the yeast system, normalization is typically performed using yeast genomic DNA, which contains proportionately small amounts of non-coding DNA. In contrast, the genomes of higher organisms such as mice contain a much higher representation of non-coding DNA. The MSP is analogous to genomic DNA as a control, with the exception that non-coding regions are removed. Typically, a concentration titration is done to span as wide an intensity range as possible. However, due to limitations in the construction of the MSP, very high expression values cannot be represented. In practice, one could construct an MSP of lower complexity with a larger representation of highly expressed genes. Since most rare and low expression genes do not contribute significantly to an MSP signal, removing this population is analogous to further removal of non-coding DNA. Theoretically, all labeled cDNA sequences could hybridize to this mixed probe sample, so it is therefore minimally subject to any sample-specific bias.

The use of all genes for normalization offers the most stability in terms of estimating spatially- and intensity-dependent trends in the log ratios. However, in biological samples which show significant divergence, a lowess fit to the MA-plot based on all genes may not produce accurate normalized log ratios. In such instances, it would be more appropriate to normalize using the MSP spots alone. While the MSP and rank-invariant controls are effective for intensity-dependent normalization, we have found that normalization based on all genes is more reliable for spatial normalization. This is due in part to the low representation of MSP and rank-invariant spots per print tip group (6–12 spots per 400 spots) and is an example of bias variance trade-off.

Composite normalization and the MSP titration series

This article has proposed a composite normalization procedure which combines the utility of normalization methods based on all genes and those based on only the MSP titration spots. For low A intensity values, normalization is based on all genes in the corresponding intensity range. For higher A values, particularly in more divergent biological samples, normalization is based primarily on the MSP titration series. In other circumstances as they warrant, other normalization methods may be incorporated into the composite technique. For example, in cases where microarrays are printed without MSP titration spots, very high intensities may be normalized using housekeeping genes and median to low intensities may be normalized using all genes in the corresponding range.

The MSP spots were essential to validate the assumptions behind our various normalization procedures and are necessary for normalizing biologically divergent samples. The construction of the MSP titration series is important, as we observed intensity-dependent dye biases in many experiments. Efforts are still in progress to devise variants of this control set for scale normalization procedures. It is evident that no single control sample or normalization procedure is accurate or adequate for all types of microarray comparisons. However, it is becoming increasingly common for investigators to print microarrays with a large complement of control spots. This flexibility expands the opportunity to customize normalization procedures, depending on the experimental conditions.

The strengths and weaknesses of the normalization techniques and control samples discussed in this paper are summarized in Table 1. Finally, the methods described in the article are implemented in the package R (17), SMA (Statistics for Microarray Analysis), which may be downloaded from <http://www.R-project.org>. Supplementary analyses, figures and datasets are available at <http://www.stat.berkeley.edu/users/terry/zarray/Html/index.html>.

ACKNOWLEDGEMENTS

We would like to acknowledge Matthew J. Callow from the Lawrence Berkeley National Laboratory for providing the data we used to develop the various normalization approaches and Yasushi Okazaki and Yoshihide Hayashizaki of the RIKEN Genomics Sciences Center for graciously providing their normalized mouse cDNA clone set for our use. We would like to thank Elva Diaz, Andrew Finn, Jonathan Scolnick and Tito Serafini for discussions and assistance over the course of this project. We are also grateful to Eric Schadt and Wing Hung Wong for discussions as well as for providing the code for their rank-invariant normalization methods. This work was supported in part by the NIH through grants 5R01MH61665-02 (J.N. and T.P.S.) and 8R01GM59506A (T.P.S.), by funds from the Department of Molecular and Cell Biology and Helen Wills Neuroscience Institute (University of California at Berkeley) and by a PMMB Burroughs-Wellcome postdoctoral fellowship (S.D.).

REFERENCES

1. Taniguchi, M., Miura, K., Iwao, H. and Yamanaka, S. (2001) Quantitative assessment of DNA microarrays—comparison with northern blot analyses. *Genomics*, **71**, 34–39.
2. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepanians, S.B., Shoemaker, D.D., Gachotte, D., Chakraburty, K., Simon, J., Bard, M. and Friend, S.H. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
3. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futsch, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
4. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.L., Yang, L., Marti, G.E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Staudt, L.M. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.
5. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, **96**, 6745–6750.
6. DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A. and Trent, J.M. (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.*, **14**, 457–460.
7. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
8. Perou, C.M., Jeffrey, S.S., van de Rijn, M., Rees, C.A., Eisen, M.B., Ross, D.T., Pergamenschikov, A., Williams, C.F., Zhu, S.X., Lee, J.C., Lashkari, D., Shalon, D., Brown, P.O. and Botstein, D. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA*, **96**, 9212–9217.
9. Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D. and Brown, P.O. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.*, **24**, 227–235.
10. Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002) Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Stat. Sin.*, in press.
11. The Chipping forecast. (1999) *Nature Genet.*, **21** (suppl.).
12. Marton, M.J., DeRisi, J.L., Bennett, H.A., Iyer, V.R., Meyer, M.R., Roberts, C.J., Stoughton, R., Burchard, J., Slade, D., Dai, H., Bassett, D.E. Jr, Hartwell, L.H., Brown, P.O. and Friend, S.H. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.*, **4**, 1293–1301.
13. Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engle, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., Wu, L.F., Altschuler, S.J., Edwards, S., King, J., Tsang, J.S., Schimmack, G., Schelter, J.M., Koch, J., Ziman, M., Marton, M.J., Li, B., Cundiff, P., Wand, T., Castle, J., Krolewski, M., Meyer, M.R., Mao, M., Burchard, J., Kidd, M.J., Dai, H., Phillips, J.W., Linsley, P.S., Stoughton, R., Scherer, S. and Boguski, M.S. (2001) Experimental annotation of the human genome using microarray technology. *Nature*, **409**, 922–927.
14. Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P. and Rubin, E.M. (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res.*, **10**, 2022–2029.
15. Miki, R., Kadota, K., Bono, H., Mizuno, Y., Tomaru, Y., Carminci, P., Itoh, M., Shibata, K., Kawai, J., Konno, H., Watanabe, S., Sato, K., Tokusumi, Y., Kikuchi, N., Ishii, Y., Hamaguchi, Y., Nishizuka, I., Goto, H., Nitanda, H., Satomi, S., Yoshiki, A., Kusabe, M., DeRisi, J.L., Eisen, M.B., Iyer, V.R., Brown, P.O., Muramatsu, M., Shimada, H., Okazaki, Y. and Hayashizaki, Y. (2001) Delineating developmental and metabolic pathways *in vivo* by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc. Natl. Acad. Sci. USA*, **98**, 2199–2204.
16. Buckley, M.J. (2000) *The Spot User's Guide*. CSIRO Mathematical and Information Sciences, North Ryde, NSW 1670, Australia. <http://www.cmis.csiro.au/iap/spot.htm>
17. Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Statist.*, **5**, 299–314.
18. Westfall, P.H. and Young, S.S. (1993) *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*. Wiley, New York.
19. Schadt, E., Li, C., Ellis, B. and Wo, W.H. (1999) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. Department of Statistics, UCLA. Preprint 303, www.stat.ucla.edu
20. Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J. and Wong, W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, **29**, 2549–2557.
21. Yang, Y.H., Dudoit, S., Luu, P. and Speed, T.P. (2001) Normalization for cDNA microarray. In Bittner, M.L., Chen, Y., Dorsel, A.N. and Dougherty, E.R. (eds), *Microarrays: Optical Technologies and Informatics*. SPIE, Society for Optical Engineering, San Jose, CA.
22. Chen, Y., Dougherty, E.R. and Bittner, M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics*, **2**, 364–374.
23. Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.
24. Axon Instruments Inc. (1999) *GenePix 4000A User's Guide*. Axon Instruments, Union City, CA.
25. Kerr, M.K., Martin, M. and Churchill, G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.
26. Finkelstein, D.B., Gollub, J., Ewing, R., Sterky, F., Somerville, S. and Cherry, J.M. (2000) Iterative linear regression by sector: renormalization of cDNA microarray data and cluster analysis weighted by cross homology. In *CAMD*. http://afgc.stanford.edu/afgc_html/site2Stat.htm
27. Sapir, M. and Churchill, G.A. (2000) Estimating the posterior probability of differential gene expression from microarray data. Poster. The Jackson Laboratory. www.jax.org/research/churchill/pubs
28. Kepler, T.B., Crosby, L. and Morgan, K.T. (2000) Normalization and analysis of DNA microarray data by self-consistency and local regression. Santa Fe Institute. 00-99-055, www.santafe.edu/sfi/publications/00wplst.html

Biostatistics (2003), **4**, 2, pp. 249–264
 Printed in Great Britain

Exploration, normalization, and summaries of high density oligonucleotide array probe level data

RAFAEL A. IRIZARRY

Department of Biostatistics, Johns Hopkins University, Baltimore MD 21205, USA
 rafa@jhu.edu

BRIDGET HOBBS

Division of Genetics and Bioinformatics, WEHI, Melbourne, Australia

FRANCOIS COLLIN

Gene Logic Inc., Berkeley, CA, USA

YASMIN D. BEAZER-BARCLAY, KRISTEN J. ANTONELLIS, UWE SCHERF

Gene Logic Inc., Gaithersburg, MD, USA

TERENCE P. SPEED

*Division of Genetics and Bioinformatics, WEHI, Melbourne, Australia. Department of Statistics,
 University of California at Berkeley*

SUMMARY

In this paper we report exploratory analyses of high-density oligonucleotide array data from the Affymetrix GeneChip® system with the objective of improving upon currently used measures of gene expression. Our analyses make use of three data sets: a small experimental study consisting of five MGU74A mouse GeneChip® arrays, part of the data from an extensive spike-in study conducted by Gene Logic and Wyeth's Genetics Institute involving 95 HG-U95A human GeneChip® arrays; and part of a dilution study conducted by Gene Logic involving 75 HG-U95A GeneChip® arrays. We display some familiar features of the perfect match and mismatch probe (*PM* and *MM*) values of these data, and examine the variance–mean relationship with probe-level data from probes believed to be defective, and so delivering noise only. We explain why we need to normalize the arrays to one another using probe level intensities. We then examine the behavior of the *PM* and *MM* using spike-in data and assess three commonly used summary measures: Affymetrix's (i) average difference (AvDiff) and (ii) MAS 5.0 signal, and (iii) the Li and Wong multiplicative model-based expression index (MBEI). The exploratory data analyses of the probe level data motivate a new summary measure that is a robust multi-array average (RMA) of background-adjusted, normalized, and log-transformed *PM* values. We evaluate the four expression summary measures using the dilution study data, assessing their behavior in terms of bias, variance and (for MBEI and RMA) model fit. Finally, we evaluate the algorithms in terms of their ability to detect known levels of differential expression using the spike-in data. We conclude that there is no obvious downside to using RMA and attaching a standard error (SE) to this quantity using a linear model which removes probe-specific affinities.

To whom correspondence should be addressed

R. A. IRIZARRY ET AL.

An R package with the functions used for the analyses in this paper is part of the Bioconductor project and can be downloaded (<http://www.bioconductor.org>). Supplemental material, such as color versions of the figures, is available on the web (<http://www.biostat.jhsph.edu/~ririzarr/affy>).

1. INTRODUCTION

High-density oligonucleotide expression array technology is now widely used in many areas of biomedical research. The system (Lockhart *et al.*, 1996) uses oligonucleotides with length of 25 base pairs that are used to probe genes. Typically, each gene will be represented by 16–20 pairs of oligonucleotides referred to as *probe sets*. The first component of these pairs is referred to as a perfect match (*PM*) probe. Each *PM* probe is paired with a mismatch (*MM*) probe that is created by changing the middle (13th) base with the intention of measuring non-specific binding. The *PM* and *MM* are referred to as a *probe pair*. See the Affymetrix Microarray Suite User Guide (1999) for details. RNA samples are prepared, labeled and hybridized with arrays. Arrays are scanned and images are produced and analysed to obtain an intensity value for each probe. These intensities represent how much hybridization occurred for each oligonucleotide probe. Of interest is finding a way to combine the 16–20 probe pair intensities for a given gene to define a measure of expression that represents the amount of the corresponding mRNA species.

We denote the intensities obtained for each probe as

$$PM_{ijn} \text{ and } MM_{ijn}, i = 1, \dots, I, j = 1, \dots, J_n, \text{ and } n = 1, \dots, N$$

with n representing the different genes, i representing different RNA samples, and j representing the probe pair number (this number is related to the physical position of the oligonucleotide in the gene). The number of genes N usually ranges from 8000 to 20000, the number of arrays I ranges from one to hundreds, and the number of probe pairs within each gene J_n usually ranges from 16 to 20. Throughout the text indices are suppressed when there is no ambiguity.

Section 2 describes the three data sets used in this paper. Section 3 explores various interesting features of the data with the objective of defining an effective measure of gene expression using the probe level data. Section 4 describes normalization. Some expression measures, for example AvDiff and Li and Wong's MBEI, are based on $PM - MM$. Other measures, for example Affymetrix's Average Log Ratio, are based on $\log(PM/MM)$. In Sections 3 and 4 we also explore the behavior of these quantities. Section 5 describes four measures of expression. Section 6 assesses the four expression measures in terms of bias, variance, and model fit. Section 7 examines the ability of the four methods at detecting differentially expressed probe sets. Section 8 presents our conclusions.

2. DESCRIPTION OF DATA

To properly compare summary measures of expression in terms of bias, variance, sensitivity, and specificity, data for which we know the 'truth' is required. In this paper we examine three data sets for which assessments can be performed where specific results are expected. Data set A provides probes for which we can assume the measurements are entirely due to non-specific binding. This permits us to study the variance–mean relationship for intensity measures. Data set B provides the results of a spike-in experiment where gene fragments have been added at known concentrations. These data can be used to assess bias, sensitivity and specificity. Data set C provides the results from a study in which samples were hybridized at different dilutions. This permits us to assess bias and variance in a more 'realistic' scenario than with data set B.

Exploration, normalization, and summaries of high density oligonucleotide array probe level data

Data sets B and C are available from the web at <http://qolotus02.genelogic.com/datasets.nsf/>. In this section we describe them in detail for readers interested in using them. We also explain which specific subsets of the data were used for the analyses presented in this paper.

2.1 *Mouse data set A*

Data set A comes from an experiment where five MG-U74A mouse GeneChip^R arrays were used. These were hybridized with samples of lung tissue mRNA obtained from five mice exposed to different experimental conditions. About 1/5 of the probe pairs in the MG-U74A array were incorrectly sequenced. We therefore assume that the measurements read for most of these probes are entirely due to non-specific binding.

2.2 *Spike-in data sets B*

Data set B consists of experiments where 11 different cRNA fragments were added to the hybridization mixture of the GeneChip^R arrays at different picomolar (pM) concentrations. The 11 control cRNAs were BioB-5, BioB-M, BioB-3, BioC-5, BioC-3, BioDn-5 (all *E. coli*), CreX-5, CreX-3 (phage P1), and DapX-5, DapX-M, DapX-3 (*B. subtilis*) (Hill *et al.*, 2000, 2001; Baugh *et al.*, 2001). The cRNA were chosen to match the target sequence for each of the Affymetrix control probe sets. For example, for DapX (a *B. subtilis* gene), the 5', middle and 3' target sequences (identified by DapX-5, DapX-M, DapX-3) were each synthesized separately and spiked-in at a specific concentration. Thus, for example, on one of the arrays DapX-3 target sequence was added to the total hybridization solution of 200 μ l to give a final concentration of 0.5 pM.

There are two series of spike-in experiments. The experiments were originally carried out for the development of normalization procedures (Hill *et al.*, 2001). In this paper we use the data in a different way, mainly for the comparison of expression measures.

2.2.1 *The varying concentration series data set, B1.* For an individual array, all of the 11 control cRNAs were spiked-in at the same concentration and this concentration was varied across arrays, taking the values 0.0, 0.5, 0.75, 1, 1.5, 2, 3, 5, 12.5, 25, 50, and 150 pM. For example, array 1 had all control cRNAs spiked with 0.0 pM and array 2 had all control cRNAs spiked with 0.5 pM, etc. Of these 12 concentrations, 0, 0.5, 0.75, 1, 1.5, 2, 3 were represented on just one array, 5 and 100 on two arrays, and the rest were in triplicate, i.e. on three arrays for a total of 27 arrays. All arrays have a common background cRNA from an acute myeloid leukemia (AML) tumor cell line. In this paper we use only 12 arrays, one replicate for each of the 12 concentrations. One of the probe set spike-in combinations (CreX-3) failed to respond adequately, and data from that probe set is entirely omitted from the analysis. Thus we analyse data from 10 spiked-in probe-sets.

2.2.2 *Latin square series data set, B2.* In this series each of the 11 control cRNAs were spiked-in at a different concentration on each array (apart from replicates). The 12 concentrations used were 0.5, 1, 1.5, 2, 3, 5, 12.5, 25, 37.5, 50, 75, and 100 pM, and these were arranged in a 12×12 cyclic Latin square, with each concentration appearing once in each row and column. The 12 combinations of concentrations used on the arrays were taken from the first 11 entries of the 12 rows of this Latin square. Of the 12 combinations used, 11 were done on three arrays and one on just one array. All of these arrays had the same AML background as in data set B1.

The analysis in this paper makes use of data from six arrays that are a pair of triplicates. The spike-in concentrations for each of the 11 control RNAs on the two sets of triplicates is shown in Table 1.

R. A. IRIZARRY ET AL.

Table 1. Concentrations and observed ranks of each spiked-in gene in a comparison of two sets of triplicates from the Latin square series spike-in data set

Probe set	Concentration		Expected Rank	Observed Rank			
	Set of triplicates 1	set of triplicates 2		AvDiff	MAS 5.0	Li & Wong	RMA
BioB-5	100.0	0.5	1	6	2	1	1
BioB-3	0.5	25.0	2	16	1	3	2
BioC-5	2.0	75.0	4	74	6	2	3
BioB-M	1.0	37.5	4	30	3	7	5
BioDn-3	1.5	50.0	5	44	5	6	4
DapX-3	35.7	3.0	6	239	24	24	7
CreX-3	50.0	5.0	7	333	73	36	9
CreX-5	12.5	2.0	8	3276	33	3128	8
BioC-3	25.0	100.0	9	2709	8579	681	6431
DapX-5	5.0	1.5	10	4598	102	12203	10
DapX-M	3.0	1.0	11	165	19	13	6

Notice that relative concentrations of the spike-ins are three fold or more, which permits us to check the sensitivity of expression indices.

2.3 Dilution data set C

Two sources of cRNA, A (human liver tissue) and B (central nervous system cell line), were hybridized to human array (HG-U95A) in a range of proportions and dilutions. In this publication, we study data from arrays hybridized to source A starting with 1.25 μg cRNA, and rising through 2.5, 5.0, 7.5, 10.0 to 20.0 μg . There were five replicate arrays for each tissue: that is, each generated cRNA was hybridized on five HG-U95 GeneChip^R arrays. Five scanners were used in this study. Each array replicate was processed in a different scanner.

3. FEATURES OF PROBE LEVEL DATA

Figure 1(a) shows histograms of log ratio, $\log_2(PM/MM)$, stratified by quantiles of abundance, $\log_2 \sqrt{PM \times MM}$, with gray scale representing height of histogram (light is high and dark is low) for one array from data set A. The histograms have been scaled so that the mode of each histogram is represented with the same gray scale. This figure shows that, in general, MM grows with PM . Furthermore, for larger values of abundance the differences have a bimodal distribution with the second mode occurring for negative differences. The same bimodal effect is seen when we stratify by $\log_2(PM)$, thus it is not an artifact of conditioning on sums. In Figures 1(b)–1(e), four histograms with a broader stratification clearly show this effect. The figure also displays (in darker grays) the histograms of the defective probes where the bimodal distribution is also seen. Notice, there are many probe pairs with $MM \gg PM$. Finally, notice that for about 1/3 of the probes $MM > PM$. The number of probe pairs within probe sets for which $MM > PM$ varies from 0 to 14. The distribution across probe sets is the following:

# of times $MM > PM$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
# of probe sets	7401	481	628	819	1123	1461	1759	1906	1555	1200	760	345	152	50	14

All these effects have been seen in many arrays.

The defective probes are used to assess the variance–mean relationship. Intensities obtained from probe j in arrays $i = 1, \dots, I$, PM_{ijn} , are expected to have the same mean and variance. If standard deviations (SDs) $\sqrt{\{(I-1)^{-1} \sum (PM_{ijn} - PM_{\cdot jn})^2\}}$ and averages $PM_{\cdot jn} = I^{-1} \sum_i PM_{ijn}$ are computed

Exploration, normalization, and summaries of high density oligonucleotide array probe level data

Fig. 1. (a) Histograms of log ratio $\log_2(PM/MM)$, stratified by quantiles of abundance, $\log_2 \sqrt{PM \times MM}$, with gray scale representing height of histogram (light grays are high and dark grays are low) for one array from the mouse data set. The histograms have been scaled so that the mode of each histogram is represented with the same gray scale. (b) Histogram of log ratios for first quartile of abundance with the histogram for the defective probes represented by a darker gray. (c) Like (b) for abundance values between first and third quartile. (d) Like (b) for abundance values in the last quartile excluding the highest 5 percent. (e) Like (b) for the highest 5% of abundance.

R. A. IRIZARRY ET AL.

for a random sample of 2000 defective probe sets, the SD increases from roughly 50 to 5000, a factor of 100-fold, as the average increases on its entire range. After a log transformation of the PM intensities there is only a 1.5-fold increase.

4. NORMALIZATION

In many of the applications of high-density oligonucleotide arrays, the goal is to learn how RNA populations differ in expression in response to genetic and environmental differences. For example, large expression of a particular gene or genes may cause an illness resulting in variation between diseased and normal tissue. These sources of variation are referred to as *interesting variation*. Observed expression levels also include variation introduced during the sample preparation, manufacture of the arrays, and the processing of the arrays (labeling, hybridization, and scanning). These are referred to as sources of 'obscuring variation'. See (Hartemink *et al.*, 2001) for a more detailed discussion. The obscuring sources of variation can have many different effects on data. Unless arrays are appropriately *normalized*, comparing data from different arrays can lead to misleading results.

Dudoit *et al.* (2001) describe the need for normalization procedures for cDNA microarray data. Similar issues are present with GeneChip^R arrays. Figures 2(a) and 2(b) show box plots of $\log_2(PM)$ and $PM - MM$ for data set C. The different gray scales represent the six different sets of five replicates processed on scanners 1 to 5. The scanner effect is clearly seen in Figure 2. For example, note that the $\log_2(PM)$ boxplot intensities obtained using scanner/fluidic station 1 were in general higher than those obtained from scanner/fluidic station 5. For the replicate arrays we expect no genes to be differentially expressed. This figure shows direct array to array comparison of PM values warrants normalization. Figure 2(b) boxplot shows that further normalization is needed for the $PM - MM$ as well.

Figures 3(a) and 3(b) show log ratios, $M = \log_2(y/x)$ versus abundance $A = \log_2 \sqrt{x \times y}$, (MVA) plots for $x = PM_1$, $y = PM_2$ and $x = PM_1 - MM_1$, $y = PM_2 - MM_2$ for two arrays (denoted with 1 and 2) in which the BioDn-3 gene has been spiked at 5 pM and 2 pM respectively. These plots have been used by, for example, Dudoit *et al.* (2002) to explore intensity related biases. Because the same RNA background was hybridized to arrays 1 and 2, we do not expect any of the non-spiked-in genes to be differentially expressed and therefore these plots to scatter around 0. It is clear from Figure 3 that these data need normalization.

For cDNA arrays the normalization procedure presented in Dudoit *et al.* (2002) has worked well in practice. For each array, a loess curve is fitted to the MVA plot of intensities of the red and green labels and the residuals are considered the normalized log ratios. However, this approach is not appropriate for GeneChip^R arrays because only one sample is hybridized to each array instead of two (red and green). A procedure that normalizes each array against all others is needed.

Various methods have been proposed for normalizing GeneChip^R arrays. Bolstad *et al.* (2002) present a review of these methods and find *quantile normalization* to perform best. The goal of quantile normalization is to make the distribution of probe intensities the same for arrays $i = 1, \dots, I$. The normalization maps probe level data from all arrays, $i = 1, \dots, I$, so that an I -dimensional quantile-quantile plot follows the I -dimensional identity line. A possible problem with this approach is that we risk removing some of the signal in the tails. However, empirical evidence suggest this is not a problem in practice: see Bolstad *et al.* (2002) for details.

In Figures 3(c) and 3(d) the MVA plots of the normalized arrays are shown. Notice how the normalization has removed the bias seen in Figures 3(a) and 3(b). The large points represent the 20 spiked-in probes and the small black dots represent a random sample of non-spiked-in probes. Notice that in all plots, normalization helps identify the spiked-in probes as differentially expressed. The benefits of this normalization at the probe level are also seen in Figures 2(c) and 2(d).

Exploration, normalization, and summaries of high density oligonucleotide array probe level data

Fig. 2. Boxplots of $\log_2(PM)$ and $PM - MM$ for the 30 arrays from data set C. Because $PM - MM$ values are usually between -2000 and $10\,000$, a reduced range is used to get a better view of the interquartile range. The bottom row are the after quantile normalization boxplots. The y-axis scale can be deduced from the plot titles.

5. MEASURES OF EXPRESSION

Various measures of expression have been proposed: for example see Li and Wong (2001), Naef *et al.* (2001), and Holder *et al.* (2001). The most commonly used (at the time this paper was written) is AvDiff,

R. A. IRIZARRY ET AL.

Fig. 3. MVA plots (described in text) of $\log_2(PM)$ and $\log_2(PM - MM)$ for two arrays in which the BioDn-3 gene has been spiked at 5 pM and 2 pM respectively. The large points represent the 20 spiked-in probes and the small black dots represent a random sample of non-spiked-in probes. (a) and (c) are before normalization and (b) and (d) are after quantile normalization.

the Affymetrix default. For each probe set n on each array i , AvDiff is defined by

$$\text{AvDiff} = \frac{1}{\#A} \sum_A (PM_j - MM_j)$$

Exploration, normalization, and summaries of high density oligonucleotide array probe level data

with A the subset of probes for which $d_j = PM_j - MM_j$ are within 3 SDs away from the average of $d_{(2)}, \dots, d_{(J-1)}$ with $d_{(j)}$ the j th smallest difference. $\#A$ represents the cardinality of A . Many of the other expression measures are versions of AvDiff with different ways of removing outliers and different ways of dealing with small values.

We have observed that linear scale measures, such as AvDiff, are not optimal. Li and Wong (2001) observed this and proposed an alternative model based expression index. For each probe set n , Li and Wong's measure is defined as the maximum likelihood estimates of the $\theta_i, i = 1, \dots, I$ obtained from fitting

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \epsilon_{ij} \quad (1)$$

with ϕ_j representing probe-specific affinities and the ϵ_{ijn} are assumed to be independent normally distributed errors. The estimation procedure includes rules for outlier removal.

Affymetrix also appears to have noticed that the linear scale is not appropriate and, in the new version of their analysis algorithm MAS 5.0, are now using a log scale measure. Specifically the MAS 5.0 signal (measure) is defined as

$$\text{signal} = \text{Tukey Biweight}\{\log(PM_j - CT_j)\}$$

with CT_j a quantity derived from the MM s that is never bigger than its PM pair. See Hubbell (2001) for more details.

Each of these measures rely upon the difference $PM - MM$ with the intention of correcting for non-specific binding. However, the exploratory analysis presented in Section 3 suggests that the MM may be detecting signal as well as non-specific binding. Some researchers (Naef *et al.*, 2001) propose expression measures based only on the PM . In Figure 4 we show the $PM, MM, PM/MM$ and $PM - MM$ values for each of the 20 probes representing BioB-5 in the 12 spiked-in arrays, from data set B1, plotted against spike-in concentration. The 20 different probe pairs are represented with different symbols and line types. As expected, the PM values are growing in proportion to the concentration. Notice also that the lines representing the 20 probes are close to being parallel, showing there is a strong additive (in the log scale) probe-specific effect. As evident in Figure 4(c), the additive probe-specific effect is also detected by the MM motivating their subtraction from the PM . However, in Figure 4(d) the parallel lines are still seen in $PM - MM$, demonstrating that subtracting is not enough to remove the probe effect. The fact that parallel lines are not as obvious in Figure 4(c) shows that dividing by MM removes, to some degree, the probe effect. However, the MM also grow with concentrations, because they detect signal as well as non-specific binding, hence the signal in PM/MM is attenuated. Notice, in particular, that PM/MM is unable to distinguish between concentrations of 25 and 150. Since subtracting probe-specific MM adds noise with no obvious gain in bias and because PM/MM results in a biased signal, in this paper we propose an alternative measure to those based on $PM - MM$ or PM/MM .

Figure 4(a) shows that on a log scale (i) the PM s grow roughly linearly with respect to concentrations, (ii) the variances are roughly constant and (iii) the probe-specific affinity is approximately additive. This suggests an additive linear model for the probe set data and the average $J^{-1} \sum_{j=1}^J \log(PM_{ij})$ as a log scale measure of expression. However, this measure does not account for non-specific binding. Because, in Figure 4, the log-scale slope of the PM is less than 1, particularly for small concentrations, the PM values should be adjusted to account for non-specific binding. To see this consider a hypothetical case with two arrays where the signal of a probe set is twice as big in one of the arrays, but an additive signal of 100 units occurs due to non-specific binding and/or background noise in both arrays. In this case the observed difference in the signals would be about $\log_2(100 + 2s) - \log_2(100 + s)$ instead of $\log_2(2s) - \log_2(s)$. For small values of s the incorrect difference would be close to 0 instead of 1.

Figure 5 shows histograms of $\log_2(MM)$ for an array in which no probe-set was spiked along with the three arrays in which BioB-5 was spiked-in at concentrations of 0.5, 0.75, and 1 pM (from data set

R. A. IRIZARRY ET AL.

Fig. 4. PM , MM , PM/MM , and $PM - MM$ values for each of the 20 probes representing BioB-5 (with the exception of CreX-3, all other spike-in genes behaved similarly to BioB-5) in the 12 spiked-in arrays from the varying concentration experiment plotted against concentration. The different probes are represented by the different line types and symbols. The horizontal line represents the median of the 20 BioB-5 probes for the non-spiked-in array. The dashed lines are the 25th and 75th quantiles.

Exploration, normalization, and summaries of high density oligonucleotide array probe level data

Fig. 5. Histograms of $\log_2(MM)$ for a array in which no probe-set was spiked along with the three arrays in which BioB-5 was spiked-in at concentrations of 0.5, 0.75, and 1 pM. The observed PM values for the 20 probes associated with BioB-5 are marked with crosses and the average with an arrow. The black curve represents the log normal distribution obtained from left-of-the-mode data.

R. A. IRIZARRY ET AL.

B1). All arrays in all data sets had similar shaped $\log_2(MM)$ histograms. Furthermore, the $\log_2(MM)$ histograms for the spiked-in probe set had similar histograms as well. The MM s to the left of the mode of the histogram can be approximated with the left-hand tail of a log-normal distribution. This suggests that the MM s are a mixture of probes for which (i) the intensities are largely due to non-specific binding and background noise and (ii) the intensities include transcript signal just like the PM s. The mode of the histogram is a natural estimate of the mean background level. The observed PM values for the 20 probes associated with BioB-5 are marked with crosses and the average with an arrow. All the average PM values are close to 100. Thus, judging solely on the average, a difference would be hard to detect. However, distance of the average PM from the average background noise does in fact increase with concentration.

Figure 5 motivates a background plus signal model of the form $PM_{ijn} = bg_{ijn} + s_{ijn}$. Here bg_{ijn} represents background signal in array i caused by optical noise and non-specific binding. We assume each array has a common mean background level, $E(bg_{ijn}) = \beta_i$. We want to adjust the PM intensities to remove the background effect. A naive approach is to consider $PM_{ijn} - \hat{\beta}_i$, with $\log_2(\hat{\beta}_i)$ the mode of the $\log_2(MM)$ distribution. An estimate of this distribution can be obtained using a density kernel estimate. In practice, a problem with this measure is that for a small percentage of probes $PM_{ijn} \leq \hat{\beta}_i$ and log transforming $PM_{ijn} - \hat{\beta}_i$ becomes a problem. An alternative background correction is to consider $B(PM_{ijn}) - E(s_{ijn}|PM_{ijn})$. If we impose a strictly positive distribution on s_{ijn} , then $B(PM_{ijn}) > 0$. To obtain a computationally feasible $B(\cdot)$ we consider the closed-form transformation obtained when assuming s_{ijn} is exponential and bg_{ijn} is normal. Although the data suggest that this model can be improved, the results obtained using $B(\cdot)$ work well in practice, as is demonstrated in the next section.

To obtain an expression measure we assume that for each probe set n , the background-adjusted, normalized, and log-transformed PM intensities, denoted with Y , follow a linear additive model

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \epsilon_{ijn}, i = 1, \dots, I, j = 1, \dots, J, n = 1, \dots, n \quad (2)$$

with α_j a probe affinity effect, μ_i representing the log scale expression level for array i , and ϵ_{ij} representing an independent identically distributed error term with mean 0. For identifiability of the parameters we assume that $\sum_j \alpha_j = 0$ for all probe sets. This assumption is saying that Affymetrix technology has chosen probes with intensities that on average are representative of the associated genes expression. The estimate of μ_i gives the expression measures for probe set n on array i .

To summarize, in this paper we consider a new expression measure that (i) background-corrects the arrays using the transformation $B(\cdot)$, (ii) normalizes the arrays using quantile normalization, and (iii) for each probe set n , fits a linear model (2) to the background-corrected, normalized and log (base 2) transformed probe intensities denoted here with $Y_{ij}, i = 1, \dots, I, j = 1, \dots, J$. To protect against outlier probes we use a robust procedure, such as median polish (Holder *et al.*, 2001), to estimate model parameters. We use the estimate of μ_i as the log scale measure of expression which we refer to as robust multi-array average (RMA).

6. BIAS, VARIANCE, AND GOODNESS OF FIT COMPARISONS

Plots of log observed expression versus known concentration (not shown) demonstrate that the expression measures perform similarly in detecting the spiked-in probe sets. However, for the highest concentration, AvDiff and MBEI sometimes underestimate the predicted value from the known concentrations. This results from the attenuation caused by subtracting MM . We also notice that RMA is less noisy than all other measures at lower concentrations.

It is possible that the control genes used in data set B1 provide a stronger than usual signal. Therefore, a comparison based on all probe sets of the HG-U95A arrays is conducted using data set C. For these data

Exploration, normalization, and summaries of high density oligonucleotide array probe level data

Fig. 6. Data set C boxplots. (a) Averages over replicates for each gene in (b). (b) Loess curves fitted to standard deviation versus average expression scatter-plots.

the amount of hybridization of probe sets representing expressed genes is expected to double when the amount of RNA hybridized to the array is double. Furthermore, the difference in gene expression across replicate arrays should be small.

For each of the four measures, we denote the expression values with E_{ik} , $i = 1, \dots, 6$, $k = 1, \dots, 5$ with i representing the dilution concentration level and k the replicate (which also identifies scanner). The averages are denoted with $E_i = (1/5) \sum_{k=1}^5 E_{ik}$ and the SDs with $SD_i = (1/4) \sum_{k=1}^5 (E_{ik} - E_i)^2$. Figure 6(a) shows boxplots of the E_i for each dilution concentration i . Notice that all measures have

R. A. IRIZARRY ET AL.

roughly the same ability to detect signal. Figure 6(b) shows loess curves fitted to the scatter plot (on the log scale) of SD_i vs E_i . Clearly, RMA has the smallest SD across replicates. The advantage of RMA is especially noticeable in the low expression values where the SD is 10 times smaller than the other measures.

Li and Wong's method provides not only an estimate of θ_i but a nominal SE for this estimate, denoted here with $\hat{\sigma}_i$. Under (2) one can obtain a naive nominal estimate for the SE of $\hat{\mu}$ using an analysis of variance approach. Because there are five replicates one can also obtain an observed SE of any estimate by simply considering the SD_i defined above. If the model is close to the actual mechanism giving rise to the data, the nominal and observed SE should agree. Plots of nominal to observed SE log ratios versus expression (not shown) show that in general, RMA is closer to 0 than Li and Wong's MBEI showing that the observed and nominal standard error methods are, in general, closer when using (2) instead of (1).

7. DETECTION OF DIFFERENTIAL EXPRESSION

Data set B2 was used to assess how well the different expression measures perform at detecting differentially expressed probe sets. For each of the six arrays studied expression measures $E_{11n}, E_{12n}, E_{13n}, E_{21n}, E_{22n}, E_{23n}$ were obtained in their respective scale (log for MAS 5.0 and RMA) for each probe set $n = 1, \dots, N$. We then computed the averages over triplicates $E_{i-n} = (1/3) \sum_{k=1}^3 E_{ikn}, i = 1, 2, n = 1, \dots, N$. For the probe sets representing spike-in RNAs the observed ratios or 'fold changes' (E_{2-n}/E_{1-n} for AvDiff and MBEI or $2^{E_{1-n}-E_{2-n}}$ for MAS 5.0 and RMA) should coincide with the true ratio of the spike-in concentrations shown in Table 1. Recall that apart from the spiked-in probe sets, the background samples hybridized to the six arrays are the same. We therefore expect only the 11 probe sets shown in Table 1 to be differentially expressed. In the left side of Figure 7 MVA plots of the average expressions obtained are shown. Probe sets with negative expression measures were left out for AvDiff and Li and Wong's MBEI. Notice that all measures separate 10 out of the 11 spiked-in probe sets from the cloud of points. However, the cloud of points for probe sets with small total intensity has a much larger spread for AvDiff, MBEI, and MAS 5.0 than for RMA. For this reason, many of the probe sets with high differential expressions for AvDiff, MBEI, and MAS 5.0 are not actually the spiked-in probe-sets. The smaller spread of RMA results in better detection of differentially expressed probe-sets. In the right side of Figure 7, quantile-quantile plots of the observed ratios are shown. RMA is the only measure to perfectly differentiate the spiked-in probe sets (with the exception BioC-3, which no measure was able to detect) from the rest. Table 1 shows the observed rank of the spiked-in probe sets.

8. CONCLUSION

In this paper we have developed a novel measure of gene expression and compared it to other standard measures. Through the analyses of three data sets, we have shown that expression is better measured using log-transformed PM values, after carrying out a global background adjustment and across-array normalization. We studied the performance of a version of the Affymetrix summary measures AvDiff and MAS 5.0, the Li and Wong model-based expression index, and the new measure RMA. We evaluated the four expression summary measures using spike-in and dilution study data, assessing their behavior in terms of bias, variance, the ability to detect known differential expression levels, and (for MBEI and RMA) model fit. We conclude that there is no obvious downside to summarizing the expression level of a probe set with RMA, and attaching an SE to this quantity using a linear model that removes probe-specific affinities. The greater sensitivity and specificity of RMA in detection of differential expression provides a useful improvement for researchers using the GeneChip^R technology. We expect marginal though worthwhile gains to be achievable by using a more carefully designed and tested background correction procedure.

Exploration, normalization, and summaries of high density oligonucleotide array probe level data

Fig. 7. MVA (described in text) and qq-plots indicating the positions of differentially expressed genes ranked by their absolute log relative expression values.

R. A. IRIZARRY ET AL.

ACKNOWLEDGEMENTS

The work of Rafael A. Irizarry was supported by the PGA U01 HL66583. We would like to thank Skip Garcia, Tom Cappola and Joshua M. Hare from Johns Hopkins University for the mouse data and Gene Brown's group at Wyeth/Genetics Institute for helpful suggestions in the design of the spike-in experiment. We would like to thank Rehannah Borup and Eric Hoffman from the Children's National Medical Center Microarray Center for help obtaining the raw mouse data. We would like to thank Laurent Gautier from the Technical University of Denmark, Ben Bolstad from UC Berkeley and Magnus strand from Astra Zeneca Mölndal for developing and coding up the normalization routines. Finally, we thank Earl Hubbell (Affymetrix), Cheng Li (Harvard), the Associate Editor, and the referee for suggestions that have improved this paper.

REFERENCES

- AFFYMETRIX (1999). *Affymetrix Microarray Suite User Guide, version 4 edition*. Santa Clara, CA: Affymetrix.
- BAUGH, L., HILL, A., BROWN, E. AND HUNTER, C. P. (2001). Quantitative analysis of mRNA amplification by *in vitro* transcription. *Nucleic Acids Research* **29**, 1–9.
- BOLSTAD, B., IRIZARRY, R., STRAND, M. AND SPEED, T. (2002). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, to appear.
- DUDOIT, S., YANG, Y. H., CALLOW, M. J. AND SPEED, T. P. (2001). Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.
- HARTEMINK, A. J., GIFFORD, D. K., JAAKOLA, T. S. AND YOUNG, R. A. (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization. *SPIE BIOS*.
- HILL, A., HUNTER, C., TSUNG, B., TUCKER-KELLOGG, G. AND BROWN, E. (2000). Genomic analysis of gene expression in *c. elegans*. *Science* **290**, 809–812.
- HILL, A. A., BROWN, E. L., WHITLEY, M. Z., TUCKER-KELLOGG, G., HUNTER, C. P. AND SLONIM, D. K. (2001). Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genomebiology* **2**, 1–13.
- HOLDER, D., RAUBERTAS, R. F., PIKOUNIS, V. B., SVETNIK, V. AND SOPER, K. (2001). Statistical analysis of high density oligonucleotide arrays: a SAFER approach. *Proceedings of the ASA Annual Meeting 2001*. Atlanta, GA.
- HUBBELL, E. (2001). Estimating signal with next generation Affymetrix software. *Gene Logic Workshop on Low Level Analysis of Affymetrix GeneChip® data*.
<http://www.stat.berkeley.edu/users/terry/zarray/Affy/GLWorkshop/genelogic2001.html>.
- LI, C. AND WONG, W. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science U S A* **98**, 31–36.
- LOCKHART, D. J., DONG, H., BYRNE, M. C., FOLLETTIE, M. T., GALLO, M. V., CHEE, M. S., MITTMANN, M., WANG, C., KOBAYASHI, M., HORTON, H. AND BROWN, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**, 1675–1680.
- NAEF, F., LIM, D. A., PATIL, N. AND MAGNASCO, M. O. (2001). From features to expression: High density oligonucleotide array analysis revisited. *Tech Report* **1**, 1–9.

[Received June 3, 2002; revised July 8, 2002; accepted for publication July 22, 2002]

The Annals of Statistics
 2006, Vol. 34, No. 5, 2387–2412
 DOI: 10.1214/009053606000000759
 © Institute of Mathematical Statistics, 2006

A MULTIVARIATE EMPIRICAL BAYES STATISTIC FOR REPLICATED MICROARRAY TIME COURSE DATA

BY YU CHUAN TAI¹ AND TERENCE P. SPEED²

*University of California, Berkeley and Walter and Eliza Hall Institute
 of Medical Research, Australia*

In this paper we derive one- and two-sample multivariate empirical Bayes statistics (the *MB*-statistics) to rank genes in order of interest from longitudinal replicated developmental microarray time course experiments. We first use conjugate priors to develop our one-sample multivariate empirical Bayes framework for the null hypothesis that the expected temporal profile stays at 0. This leads to our one-sample *MB*-statistic and a one-sample $\tilde{T}^2$ -statistic, a variant of the one-sample Hotelling T^2 -statistic. Both the *MB*-statistic and $\tilde{T}^2$ -statistic can be used to rank genes in the order of evidence of nonzero mean, incorporating the correlation structure across time points, moderation and replication. We also derive the corresponding *MB*-statistics and $\tilde{T}^2$ -statistics for the one-sample problem where the null hypothesis states that the expected temporal profile is constant, and for the two-sample problem where the null hypothesis is that two expected temporal profiles are the same.

1. Introduction. Microarray time course experiments differ from other microarray experiments in that gene expression values at different time points can be correlated. This may happen when the design is longitudinal, that is, where the mRNA samples at successive time points are taken from the same *units*. Such longitudinal experiments make it possible to monitor and study the temporal *changes* within units of biological processes of interest for thousands of genes simultaneously. Two major categories of time course experiments are those involving periodic and developmental phenomena. Periodic time courses typically concern natural biological processes such as cell cycles or circadian rhythms, where the temporal profiles follow regular patterns [7, 8, 25, 26]. On the other hand, developmental time course experiments measure gene expression levels at a series of times in a developmental process, or after applying a treatment such as a drug to the organism, tissue or cells [9, 32, 34]. In this case, we typically have few prior expectations concerning the temporal patterns of gene expression. The gene

---

Received August 2004; revised August 2005.

¹Supported in part by NIH Grant HG00047.

²Supported in part by NIH Grant R01 Lmo7609-01.

AMS 2000 subject classifications. Primary 62M10; secondary 62C12, 92D10.

Key words and phrases. Microarray time course, longitudinal, multivariate empirical Bayes, moderation, gene ranking, replication.

Y. C. TAI AND T. P. SPEED

ranking methods we develop in this paper are mainly for longitudinal replicated developmental time course data.

A typical microarray time course dataset consists of expression measurements of G genes across k time points, under one or more biological conditions (e.g., wildtype versus mutant). The number of genes G (1,000–40,000) is very much larger than the number of time points k , which can be 5–10 for shorter and 11–20 for longer time courses. Many such experiments are unreplicated due to cost or other limitations, and when replicates are done, the number n is typically quite small, say, 2–5. We refer the reader to [29] for a fuller review of microarray time course experiments.

One of the statistical challenges here is to identify genes of interest. In what we call the one-sample problem, these are genes whose patterns of expression change over time, perhaps in some specific way. In the two-sample problem we seek genes whose temporal patterns differ across two biological conditions. Such genes are of interest to biologists because they are often involved in the biological processes motivating the experiment. The challenge arises from the fact that there are very few time points, and very few replicates per gene. The series are usually so short that we cannot consider using standard time series methods as described in [10], such as Fourier analysis, ARMA models or wavelets. The methods proposed in this paper are for the one- and two-sample problems with longitudinal replicated microarray time course experiments of the developmental kind.

The gene ranking problem for such microarray experiments is relatively new. Few methods have been proposed to deal specifically with these problems. The most widely used method for identifying temporally changing genes in replicated microarray experiments is to carry out multiple pairwise comparisons across times, using statistics developed for comparing two independent samples, [2, 6, 13, 19, 20, 23, 24, 33]. These methods are not entirely appropriate as they do not incorporate the fact that longitudinal microarray time course samples are correlated. A simple and intuitive approach to our problem is to use classical or mixed ANOVA models; see Chapter 6 of [11] for a discussion of the latter for analyzing longitudinal data and [22] for a modified approach based on the former for use in the microarray context. However, a number of questions are not adequately addressed by the classical ANOVA methods, or the variants of [22]. As with the pairwise comparisons, the F -statistic assumes that gene expression measurements at different times are independent. The classical ANOVA models also assume normality of the gene expression measurements, which may not be a great concern when these are on the log scale. More importantly, standard F -statistics in this context will generally lead to more false positives and false negatives than is desirable, due to poorly estimated error variances in the denominator. This issue can be dealt with using the idea of moderation; see, for example, [2, 14, 20, 24, 33]. Moderation in our longitudinal context means the smoothing of gene-specific sample variance–covariance matrices toward a common matrix. When we do this, fewer

REPLICATED MICROARRAY TIME COURSE DATA

genes which are not differentially expressed over time but have very small replicate variances are falsely identified as being differentially expressed, and fewer genes which are differentially expressed over time but have large replicate variances (e.g., due to outliers) are missed by the F -statistic.

A moderated gene-specific score based on the Wald statistic for the longitudinal one-sample problem was proposed in [15]. However, their method is only applicable to the situation when the number of replicates is greater than the number of time points. In [3] the expression profiles for each gene and each of two biological conditions were represented by continuous curves fitted using B-splines. A global difference between the two continuous curves and an ad hoc likelihood-based p -value was calculated for each gene. B-splines were also used in [17] to identify genes with different temporal profiles in the two-sample case. Recently, B-splines were again adopted by [27] to model the population mean, constructing the F -statistics for both longitudinal and cross-sectional data with one or more biological conditions. A major feature of this paper was a careful treatment of the multiple testing issue in this context. A novel HMM approach which incorporates the dependency in gene expression measurements across times was proposed in [36] for data with two or more biological conditions. This is one example of using HMM to identify differentially expressed genes across at least two biological conditions in this context.

The multivariate empirical Bayes model proposed in this paper was motivated by the analogous univariate model proposed in [20] for identifying differentially expressed genes in two-color comparative microarray experiments, and the more recent extensions by Smyth [24]. The B -statistic in [20] and [24], and the univariate moderated t -statistic $\tilde{t}_g$ in [24], consider just one parameter or contrast at a time in the null hypotheses. They are not for null hypotheses with two or more parameters or contrasts of interest simultaneously. However, a partly-moderated F -statistic was introduced in [24], which moderates the error variance in the denominator of the ordinary F . This partly-moderated F -statistic is useful for the simultaneous comparison of multiple uncorrelated contrasts, but as mentioned above, it is not appropriate for longitudinal experiments. Both the MB -statistics and the $\tilde{T}^2$ -statistics derived in this paper provide a degree of moderation, while retaining the temporal correlation structure. They can be used with both single- and two-channel microarray experiments.

This paper is organized as follows. After briefly introducing our notation, we formally state the null and alternative hypotheses for the gene ranking problem in Section 2. In Section 3 we present moderated versions of the standard likelihood-ratio and Hotelling T^2 -statistics. We formally build up our multivariate empirical Bayes model and derive the MB - and $\tilde{T}^2$ -statistics in Section 4. A brief description of a case study is presented in Section 5. Section 6 reports results from a simulation study in which we compare the one-sample MB -statistic (the $\tilde{T}^2$ -statistic) of Section 4.3 with other statistics. We discuss our models and give directions for future work in Section 7.

Y. C. TAI AND T. P. SPEED

Now we introduce some notation for one-sample problems. For two-sample problems, the notation is similar and easily conceived. For each gene g , $g = 1, \dots, G$, we have n_g independent time series, and we model these as i.i.d. $k \times 1$ random vectors $\mathbf{X}_{g1}, \dots, \mathbf{X}_{gn_g}$ with gene-specific means $\boldsymbol{\mu}_g$ and covariance matrices $\boldsymbol{\Sigma}_g$. Since relative or absolute gene expression measurements are approximately normal on the log scale, we make the multivariate normality assumption on data $\mathbf{X}_{g1}, \dots, \mathbf{X}_{gn_g}$. Our results are to be judged on their practical usefulness, not on the precise fit of our data to a multivariate normal distribution. As will be seen shortly, our final formulae involve the multivariate t distribution. Thus, a measure of robustness is built in, and our approach will probably be about as effective for elliptically distributed random vectors. We use the natural conjugate priors for $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$, that is, an inverse Wishart prior for $\boldsymbol{\Sigma}_g$ and a dependent multivariate normal prior for $\boldsymbol{\mu}_g$. To simplify the notation, the subscript g will be dropped for the rest of this paper. The statistical models presented in the rest of this paper are for an arbitrary single gene g .

The details in this paper differ in two ways from the standard conjugate priors. First, we also have an indicator I such that $I = 1$ when the alternative K is true and $I = 0$ when the null H is true, with the priors for $\boldsymbol{\mu}$ differing in these two cases. Second, when the null hypothesis states that a gene's mean expression level is constant, in order to get a simple closed form expression for the posterior odds, we assume that the gene-specific covariance matrix $\boldsymbol{\Sigma}$ commutes with the $k \times k$ projection matrix $\mathbf{P} = k^{-1}\mathbf{1}_k\mathbf{1}_k'$, that is, $\mathbf{P}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\mathbf{P}$. In this case the $k \times k$ inverse Wishart prior for $\boldsymbol{\Sigma}$ is replaced by a $(k-1) \times (k-1)$ inverse Wishart prior for a part of $\boldsymbol{\Sigma}$ and an inverse gamma prior for the remainder. These two-part priors are independent; see Section 4.3 for details.

2. Hypothesis testing. Our gene ranking problem will be formally stated as a hypothesis testing problem. In this paper we only seek a statistic for ranking genes in the order of evidence against the null hypothesis; we do not hope to obtain raw or adjusted p -values as in [27].

Following the notation in [4], the null hypothesis is denoted by H , while the alternative hypothesis is denoted by K . The null hypothesis corresponding to a gene's mean expression level being 0 is $H: \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} > 0$; the alternative is $K: \boldsymbol{\mu} \neq \mathbf{0}, \boldsymbol{\Sigma} > 0$. An easy extension is $H: \boldsymbol{\mu} = \boldsymbol{\mu}_0, \boldsymbol{\Sigma} > 0$ versus $K: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \boldsymbol{\Sigma} > 0$, where $\boldsymbol{\mu}_0$ is known. Later we will consider the null hypothesis that a gene's expression is constant, against the alternative that it is not: $H: \boldsymbol{\mu} = \mu_0\mathbf{1}, \boldsymbol{\Sigma} > 0$, where μ_0 is a scalar representing the expected value of the gene's expression level at any time point under H , and $\mathbf{1}$ is the $k \times 1$ constant vector of 1s; $K: \boldsymbol{\mu} \neq \mu_0\mathbf{1}, \boldsymbol{\Sigma} > 0$. Finally, we will consider the null hypothesis that a gene's mean expression levels are the same under two different biological conditions, versus the alternative that they are not: $H: \boldsymbol{\mu}_Z = \boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Z = \boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma} > 0$; $K: \boldsymbol{\mu}_Z \neq \boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Z = \boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma} > 0$.

REPLICATED MICROARRAY TIME COURSE DATA

3. The moderated LR -statistic.

3.1. *One-sample or paired two-sample problem.* A likelihood-ratio statistic can be used directly to test the null hypothesis H against the alternative hypothesis K when $n > k$. According to standard multivariate results (e.g., [21]), under the alternative hypothesis that there are no constraints on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, their maximum likelihood estimates are

$$\begin{aligned}\hat{\boldsymbol{\mu}}_K &= \bar{\mathbf{X}}, \\ \hat{\boldsymbol{\Sigma}}_K &= \frac{n-1}{n} \mathbf{S},\end{aligned}$$

where $\mathbf{S} = (n-1)^{-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$ is the sample variance–covariance matrix. Also as in [21], the maximum likelihood estimate for the unconstrained $\boldsymbol{\Sigma}$ under the null that $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ is

$$\hat{\boldsymbol{\Sigma}}_H = \frac{n-1}{n} \mathbf{S} + \mathbf{d}\mathbf{d}',$$

where $\mathbf{d} = \hat{\boldsymbol{\mu}}_K - \hat{\boldsymbol{\mu}}_H$, the difference between the maximum likelihood estimates for $\boldsymbol{\mu}_0$ (if unknown) under H and K . If $\boldsymbol{\mu}_0$ is known, then $\mathbf{d} = \hat{\boldsymbol{\mu}}_K - \boldsymbol{\mu}_0$. The likelihood ratio statistic for testing any such H against the above K is

$$\begin{aligned}(3.1) \quad LR &= 2(\mathbf{l}_K^{\max} - \mathbf{l}_H^{\max}) \\ &= n \log \left(1 + \frac{n}{n-1} \mathbf{d}' \mathbf{S}^{-1} \mathbf{d} \right).\end{aligned}$$

If the null hypothesis states that $\boldsymbol{\mu}_0 = \mathbf{0}$, then $\mathbf{d}$ reduces to $\bar{\mathbf{X}}$, hence, the likelihood-ratio statistic is equation (3.1) with $\mathbf{d}$ replaced by $\bar{\mathbf{X}}$. Similarly, if $\boldsymbol{\mu}_0$ is known, then $\mathbf{d} = \bar{\mathbf{X}} - \boldsymbol{\mu}_0$. On the other hand, if the null hypothesis states that $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_0 \mathbf{1}$, then the maximum likelihood estimate for $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}}_H = \left(\frac{\mathbf{1}' \mathbf{S}^{-1} \bar{\mathbf{X}}}{\mathbf{1}' \mathbf{S}^{-1} \mathbf{1}} \right) \mathbf{1}.$$

The statistic $n\mathbf{d}'\mathbf{S}^{-1}\mathbf{d}$ is the one-sample Hotelling T^2 -statistic, and by Section 5.3.1b in [21], it is distributed as a Hotelling $T^2(k-1, n-1)$ under H , that is, $((n-k+1)n\mathbf{d}'\mathbf{S}^{-1}\mathbf{d})/((n-1)(k-1))$ has an F -distribution with degrees of freedom $(k-1, n-k+1)$.

In the microarray time course context, the number of replicates n is typically smaller than the number of time points k , and so $\mathbf{S}$ has less than full rank. Furthermore, as discussed in [29], we wish to moderate the sample variance–covariance

Y. C. TAI AND T. P. SPEED

matrix. Our moderated $\mathbf{S}$ will take the form

$$\tilde{\mathbf{S}} = \frac{\nu \mathbf{A} + (n-1)\mathbf{S}}{\nu + n - 1},$$

where $\nu > 0$ controls the degree of moderation, and $\mathbf{A} > 0$ is the common $k \times k$ matrix toward which $\mathbf{S}$ is smoothed. In Section 4.1 we give the theoretical reason for choosing this moderated variance–covariance matrix $\tilde{\mathbf{S}}$ and explain how we estimate ν and $\mathbf{A}$. Replacing $\mathbf{S}$ with $\tilde{\mathbf{S}}$ in the LR -statistic, our moderated LR -statistic is

$$(3.2) \quad \tilde{LR} = 2(l_K^{\max} - l_H^{\max}) = n \log \left(1 + \frac{n}{n-1} \tilde{\mathbf{d}}' \tilde{\mathbf{S}}^{-1} \tilde{\mathbf{d}} \right).$$

When all the genes have an equal number of replicates n , equation (3.2) is a monotonic increasing function of $n \tilde{\mathbf{d}}' \tilde{\mathbf{S}}^{-1} \tilde{\mathbf{d}}$. We define the quadratic form $n \tilde{\mathbf{d}}' \tilde{\mathbf{S}}^{-1} \tilde{\mathbf{d}} = \|n^{1/2} \tilde{\mathbf{S}}^{-1/2} \tilde{\mathbf{d}}\|^2$ to be the moderated one-sample Hotelling T^2 -statistic. In the case of the null $H: \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} > 0$, this is identical to the $\tilde{T}^2$ -statistic we derive in Section 4.1. The one-sample moderated LR -statistic and the moderated Hotelling T^2 -statistic are hybrids of likelihood and Bayesian statistics, since $\tilde{\mathbf{S}}$ is estimated using the multivariate empirical Bayes procedure we describe below.

3.2. Unpaired two-sample problem. Similarly, in the unpaired two-sample case, the moderated LR -statistic can be written as a function of the moderated two-sample Hotelling T^2 -statistic

$$\tilde{LR} = (m+n) \log \left(1 + \frac{1}{m+n-2} \frac{mn}{m+n} \mathbf{d}' \tilde{\mathbf{S}}^{-1} \mathbf{d} \right),$$

where $\mathbf{d} = \bar{\mathbf{Z}} - \bar{\mathbf{Y}}$ is the difference between sample averages and $\mathbf{S} = (m+n-2)^{-1}((m-1)\mathbf{S}_Z + (n-1)\mathbf{S}_Y)$ is the pooled sample variance–covariance matrix, and

$$\tilde{\mathbf{S}} = \frac{(m+n-2)\mathbf{S} + \nu \mathbf{A}}{m+n-2+\nu}.$$

The term $(m+n)^{-1}mn \mathbf{d}' \tilde{\mathbf{S}}^{-1} \mathbf{d}$ is our moderated two-sample Hotelling T^2 -statistic. We use the same approach to estimate $\tilde{\mathbf{S}}$ here as that for our two-sample multivariate empirical Bayes model described in Section 4.2.

4. The multivariate empirical Bayes model.

4.1. One-sample or paired two-sample problem.

4.1.1. Models and priors. The data $\mathbf{X}_1, \dots, \mathbf{X}_n$ are multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, denoted by $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We define an indicator

REPLICATED MICROARRAY TIME COURSE DATA

random variable I to reflect the status of the gene,

$$I = \begin{cases} 1, & \text{if } K \text{ is true,} \\ 0, & \text{if } H \text{ is true.} \end{cases}$$

We suppose that I has a Bernoulli distribution with success probability p , $0 < p < 1$. Now we build up our multivariate hierarchical Bayesian model by first assigning independent and identical inverse Wishart priors to the gene-specific covariance matrices Σ :

$$(4.1) \quad \Sigma \sim \text{Inv-Wishart}_\nu((\nu\Lambda)^{-1}),$$

where $\nu > 0$ and $\nu\Lambda > 0$ are the degrees of freedom and scale matrix, respectively. Given Σ , we assign multivariate normal priors for the gene-specific mean μ for the two cases ($I = 1$) and ($I = 0$):

$$\begin{aligned} \mu|\Sigma, I = 1 &\sim N_k(\mathbf{0}, \eta^{-1}\Sigma), \\ \mu|\Sigma, I = 0 &\equiv 0, \end{aligned}$$

where $\eta > 0$ is a scale parameter.

The posterior odds are the probability that the expected time course μ does not stay at 0 (i.e., $I = 1$) over the probability that μ stays at 0 (i.e., $I = 0$), given the data $\mathbf{X}_1, \dots, \mathbf{X}_n$. Following [24]'s notation, we write

$$(4.2) \quad \begin{aligned} \mathbf{O} &= \frac{P(I = 1|\mathbf{X}_1, \dots, \mathbf{X}_n)}{P(I = 0|\mathbf{X}_1, \dots, \mathbf{X}_n)} \\ &= \frac{p}{1-p} \frac{P(\mathbf{X}_1, \dots, \mathbf{X}_n|I = 1)}{P(\mathbf{X}_1, \dots, \mathbf{X}_n|I = 0)}. \end{aligned}$$

The distribution of the data given I can be written as

$$(4.3) \quad P(\mathbf{X}_1, \dots, \mathbf{X}_n|I) = \int P(\bar{\mathbf{X}}|\Sigma, I) P(\mathbf{S}|\Sigma, I) P(\Sigma|I) d\Sigma.$$

4.1.2. Multivariate joint distributions. Once the priors and the models are set, the joint distributions of the data can be determined given I . We omit the standard calculations leading to

$$(4.4) \quad \begin{aligned} &P(\mathbf{X}_1, \dots, \mathbf{X}_n|I = 1) \\ &= \frac{\Gamma_k((n + \nu)/2)}{\Gamma_k((n - 1)/2)\Gamma_k(\nu/2)} \\ &\quad \times (n - 1)^{(1/2)k(n-1)} \nu^{-(1/2)kn} \pi^{-(1/2)k} (n^{-1} + \eta^{-1})^{-(1/2)k} \\ &\quad \times \frac{|\Lambda|^{-(1/2)n} |\mathbf{S}|^{(1/2)(n-k-2)}}{|\mathbf{I}_k + ((n^{-1} + \eta^{-1})\nu\Lambda)^{-1}\bar{\mathbf{X}}\bar{\mathbf{X}}' + (\nu\Lambda/(n - 1))^{-1}\mathbf{S}|^{(1/2)(n+\nu)}}. \end{aligned}$$

Thus, given $I = 1$, the probability density function of the data is a function of $\bar{\mathbf{X}}$ and $\mathbf{S}$ only, which follows a Student–Siegel distribution [1]. Following [1]'s

Y. C. TAI AND T. P. SPEED

notation, this distribution is denoted by $StSi_k(v, \mathbf{0}, (n^{-1} + \eta^{-1})\mathbf{A}, n - 1, (n - 1)^{-1}v\mathbf{A})$. For the case of $I = 0$, we get the same distribution with different parameters, namely, $StSi_k(v, \mathbf{0}, n^{-1}\mathbf{A}, n - 1, (n - 1)^{-1}v\mathbf{A})$.

4.1.3. *MB-statistic and $\tilde{T}^2$ -statistic.* Define our moderated gene-specific sample variance–covariance matrix $\tilde{\mathbf{S}}$ to be the inverse of the posterior mean of $\mathbf{\Sigma}^{-1}$ given $\mathbf{S}$,

$$(4.5) \quad \tilde{\mathbf{S}} = [E(\mathbf{\Sigma}^{-1}|\mathbf{S})]^{-1} = \frac{(n - 1)\mathbf{S} + v\mathbf{A}}{n - 1 + v}.$$

The posterior odds $\mathbf{O}$ we defined earlier can be derived using the distributions of the data given I and is

$$(4.6) \quad \mathbf{O} = \frac{p}{1 - p} \left(\frac{\eta}{n + \eta} \right)^{(1/2)k} \times \left(\frac{n - 1 + v + \tilde{T}^2}{n - 1 + v + (\eta/(n + \eta))\tilde{T}^2} \right)^{(1/2)(n+v)},$$

where $\tilde{T}^2 = \tilde{\mathbf{t}}'\tilde{\mathbf{t}}$ and $\tilde{\mathbf{t}}$ is the moderated multivariate t -statistic defined by

$$(4.7) \quad \tilde{\mathbf{t}} = n^{1/2}\tilde{\mathbf{S}}^{-1/2}\bar{\mathbf{X}}.$$

Following the tradition in genetics, the log base 10 of $\mathbf{O}$ is called the LOD score. To distinguish it from the LOD score (also called the B -statistic) in the univariate model of [20] and [24], the multivariate LOD score in this paper is called the *MB*-statistic,

$$(4.8) \quad MB = \log_{10} \mathbf{O}.$$

When all genes have the same number of replicates n , equation (4.8) is a monotonic increasing function of $\tilde{T}^2$. This shows that the *MB*-statistic is equivalent to the $\tilde{T}^2$ -statistic when n is the same across genes, and therefore, one is encouraged to use the $\tilde{T}^2$ -statistic in this case since it does not require the estimation of η and leads to the same rankings as equation (4.8). We now derive the distribution for $\tilde{T}^2$.

By Gupta and Nagar [16], the Jacobian transformation from $\bar{\mathbf{X}}$ to $\tilde{\mathbf{t}}$ is $J(\bar{\mathbf{X}} \rightarrow \tilde{\mathbf{t}}) = |n^{-1/2}\tilde{\mathbf{S}}^{1/2}|$. Since equation (4.4) is a function of $\bar{\mathbf{X}}$ and $\mathbf{S}$ only, it is the joint probability density function for these two random variables. Substituting for $\bar{\mathbf{X}}$ in terms of $\tilde{\mathbf{t}}$ in equation (4.4), and multiplying the resulting expression by $J(\bar{\mathbf{X}} \rightarrow \tilde{\mathbf{t}})$,

REPLICATED MICROARRAY TIME COURSE DATA

the joint probability density function for $\tilde{\mathbf{t}}$ and $\mathbf{S}$ given $I = 1$ is

$$\begin{aligned}
 & P(\tilde{\mathbf{t}}, \mathbf{S} | I = 1) \\
 &= \pi^{-(1/2)k} \frac{\Gamma((n+v)/2)}{\Gamma((n+v-k)/2)} \left(\frac{n+\eta}{\eta}\right)^{-(1/2)k} (n-1+v)^{-(1/2)k} \\
 (4.9) \quad & \times \left(1 + \frac{1}{n-1+v} \left(\frac{\eta}{n+\eta}\right) \tilde{\mathbf{t}}\tilde{\mathbf{t}}\right)^{-(1/2)(n+v)} \\
 & \times \frac{1}{\beta_k((n-1)/2, v/2)} \\
 & \times \frac{|\mathbf{S}|^{(1/2)(n-k-2)}}{|v\mathbf{\Lambda}/(n-1)|^{(1/2)(n-1)} |\mathbf{I}_k + (v\mathbf{\Lambda}/(n-1))^{-1}\mathbf{S}|^{(1/2)(n+v-1)}}.
 \end{aligned}$$

The above expression is factorized into parts involving $\mathbf{S}$ only and $\tilde{\mathbf{t}}$ only, proving that $\tilde{\mathbf{t}}$ and $\mathbf{S}$ are independent. It is apparent that $\tilde{\mathbf{t}}$ has a multivariate t distribution with degrees of freedom $n+v-k$, scale parameter $n+v-1$, mean vector $\mathbf{0}$ and covariance matrix $\eta^{-1}(n+\eta)\mathbf{I}_k$. This distribution is denoted by $\tilde{\mathbf{t}}|I=1 \sim \mathbf{t}_k(n+v-k, n+v-1, \mathbf{0}, \eta^{-1}(n+\eta)\mathbf{I}_k)$ [16]. It is straightforward to see that $\tilde{\mathbf{t}}|I=0 \sim \mathbf{t}_k(n+v-k, n+v-1, \mathbf{0}, \mathbf{I}_k)$. Given $I=1$, $\mathbf{S}$ is distributed as a generalized type-II beta distribution with parameters $(n-1)/2$, $v/2$, scale matrix $v\mathbf{\Lambda}/(n-1)$ and location matrix $\mathbf{0}$. The distribution is denoted by $GB_k^{II}((n-1)/2, v/2, v\mathbf{\Lambda}/(n-1), \mathbf{0})$ [16]. The marginal distribution of $\mathbf{S}$ does not depend on I so that $P(\mathbf{S}|I=0) = P(\mathbf{S}|I=1)$. This distributional result is used to estimate the hyperparameter $\mathbf{\Lambda}$. The distribution for $\tilde{T}^2$ under the null follows immediately. Under H , $k^{-1}\tilde{T}^2$ has an F distribution with degrees of freedom $(k, n+v-k)$; equivalently, $(n+v-k)^{-1}(n+v-1)\tilde{T}^2$ has the Hotelling T^2 -distribution $T^2(k, n+v-1)$.

The $\tilde{T}^2$ -statistic is identical to the one-sample moderated Hotelling T^2 -statistic in Section 3.1 with the same null hypothesis.

For the easy extension to the above model, $H: \boldsymbol{\mu} = \boldsymbol{\mu}_0, \boldsymbol{\Sigma} > 0$ and $K: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \boldsymbol{\Sigma} > 0$, where $\boldsymbol{\mu}_0$ is known, all the results above hold with $\bar{\mathbf{X}}$ replaced by $\bar{\mathbf{X}} - \boldsymbol{\mu}_0$.

4.1.4. Special cases.

1. $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_k$. By constraining $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_k$, we ignore the correlations among gene expression values at different times, and assume the variances at different times are equal. Suppose that the prior for σ^2 is

$$\sigma^2 \sim \text{inv-gamma}(\tfrac{1}{2}\nu, \tfrac{1}{2}\nu\lambda^2).$$

Define

$$s_j^2 = (n-1)^{-1} \sum_{i=1}^n (\mathbf{X}_{ij} - \bar{\mathbf{X}}_j)^2,$$

Y. C. TAI AND T. P. SPEED

$$\tilde{s}_j^2 = (n - 1 + \nu)^{-1}((n - 1)s_j^2 + \nu\lambda^2)$$

and

$$\tilde{t}_j = n^{1/2}\bar{\mathbf{X}}_j\tilde{s}_j^{-1}, \quad j = 1, \dots, k.$$

In this case, the posterior odds are equivalent to a product of k independent univariate odds,

$$(4.10) \quad \mathbf{O} = \frac{p}{1-p} \left(\frac{\eta}{n+\eta} \right)^{(1/2)k} \prod_{j=1}^k \left(\frac{n-1+\nu+\tilde{t}_j^2}{n-1+\nu+(\eta/(n+\eta))\tilde{t}_j^2} \right)^{(1/2)(n+\nu)},$$

and the MB -statistic is equivalent to the sum of k univariate B -statistics.

2. $n = 1$. When $n = 1$, that is, when there is no replication at all, each gene has its own unknown variability. The moderated multivariate t -statistic becomes $\tilde{\mathbf{t}} = \mathbf{\Lambda}^{-1/2}\mathbf{X}$. The posterior odds are obtained by plugging in $n = 1$ in equation (4.6), and are found to be a function of $\mathbf{X}$ only. Since there is no replication, our hyperparameters must be assigned values, for example, from previous experiments.

3. $k = 1$. When $k = 1$, that is, when there is only one time point, the alternative hypothesis states that there is differential expression at this single time point. Our multivariate model should and does reduce to the univariate model in [20] and [24].

4.1.5. Limiting cases.

1. $\nu \rightarrow \infty$. In this case, the gene-specific variance-covariance matrices are totally ignored. The moderated multivariate t -statistic above becomes $\tilde{\mathbf{t}}_\infty = n^{1/2}\mathbf{\Lambda}^{-1/2}\bar{\mathbf{X}}$, and $\tilde{T}_\infty^2 = \tilde{\mathbf{t}}_\infty'\tilde{\mathbf{t}}_\infty$. The posterior odds become

$$\mathbf{O} = \frac{p}{1-p} \left(\frac{\eta}{n+\eta} \right)^{(1/2)k} \exp \left(\frac{1}{2} \left(\frac{n}{n+\eta} \right) \tilde{T}_\infty^2 \right).$$

2. $\nu \rightarrow 0$. In this case, there is no moderation at all. The posterior odds are just equation (4.6) with ν replaced by 0. If $n < k$, then $\mathbf{S}^{-1/2}$ should be calculated by a g -inverse.

3. $\nu \rightarrow \infty$ and $\mathbf{\Sigma} = \sigma^2\mathbf{I}_k$. Define $\tilde{t}_{\infty j} = n^{1/2}\lambda^{-1}\bar{\mathbf{X}}_j$, $j = 1, \dots, k$. The posterior odds become

$$\mathbf{O} = \frac{p}{1-p} \left(\frac{\eta}{n+\eta} \right)^{(1/2)k} \exp \left(\frac{1}{2} \left(\frac{n}{n+\eta} \right) \sum_{j=1}^k \tilde{t}_{\infty j}^2 \right).$$

4. $\nu \rightarrow 0$ and $\mathbf{\Sigma} = \sigma^2\mathbf{I}_k$. In this case, the posterior odds are just equation (4.10) with ν replaced with 0.

REPLICATED MICROARRAY TIME COURSE DATA

4.1.6. *Hyperparameter estimation.* We have shown that the *MB*-statistic for assessing whether or not a time course has mean $\mathbf{0}$ depends on $(k^2 + k + 6)/2$ hyperparameters: ν , $\mathbf{\Lambda}$, η and p . In practice, we need to estimate these hyperparameters, and plug in our estimates into the formulae for $\tilde{\mathbf{S}}$, $\tilde{\mathbf{t}}$, $\mathbf{O}$, $\dots$ and so on. Slightly abusing our notation, we will use the same symbols for these estimates, relying on context to make it clear whether we are assuming the hyperparameters to be known or not. In our multivariate model, many more hyperparameters need to be estimated, compared to the univariate models in [20] and [24], both of which have four hyperparameters. Closed form estimators for the hyperparameters in the univariate linear model setting are derived in [24], using the marginal sampling distributions of the statistic $\tilde{t}$ and the sample variance s^2 , and are shown to be better than the simple estimators in [20]. Following [24], the aim of this section is to derive estimators for the hyperparameters in our multivariate model. In general, the hyperparameter η associated with the case $I = 1$ is estimated based on only a small subset of genes, while ν and $\mathbf{\Lambda}$ are estimated using the whole gene set. Instead of estimating the proportion of differentially expressed genes p , we plug in a user-defined value, since the choice of p does not affect the rankings of genes based on the *MB*-statistic.

EB estimation of ν and $\mathbf{\Lambda}$. The hyperparameter ν determines the degree of smoothing between $\mathbf{S}$ and $\mathbf{\Lambda}$. The method we use to estimate ν builds on that used to estimate d_0 in Section 6.2 in [24]. However, unlike d_0 in [24], ν is associated with the $k \times k$ matrix $\mathbf{\Sigma}$. Therefore, a method appropriate to this multivariate framework is needed. Let $\hat{\nu}_j$ be the estimated prior degrees of freedom based on the j th diagonal elements of the gene-specific sample variance–covariance matrices (i.e., the replicate variances for the j th time point from the whole gene set) using the method proposed in Section 6.2 in [24]. Our estimation of ν is based on the following two-step strategy. As the first step, set ν as $\hat{\nu} = \max(\text{mean}(\hat{\nu}_j), k + 6)$, $j = 1, \dots, k$. This estimated $\hat{\nu}$ is used to estimate $\mathbf{\Lambda}$. Once $\mathbf{\Lambda}$ is estimated, $\hat{\nu}$ is reset to be $\hat{\nu} = \text{mean}(\hat{\nu}_j)$. In practice, one can even just plug in a user-defined value ν_0 which gives the desired amount of smoothing. In such a case, the first step sets $\hat{\nu} = \max(\nu_0, k + 6)$. This $\hat{\nu}$ is used to estimate $\mathbf{\Lambda}$. After $\mathbf{\Lambda}$ is estimated, $\hat{\nu}$ can be reset to the user-defined value ν_0 .

Our estimate of $\mathbf{\Lambda}$ comes after the first step in the estimation of ν . We showed that, under our model, $\mathbf{S}$ follows the generalized type-II beta distribution with expectation $(\nu - k - 1)^{-1}\nu\mathbf{\Lambda}$. By the weak law of large numbers, $\bar{\mathbf{S}}$ converges in probability to $(\nu - k - 1)^{-1}\nu\mathbf{\Lambda}$. We can thus estimate $\mathbf{\Lambda}$ by $\hat{\nu}^{-1}(\hat{\nu} - k - 1)\bar{\mathbf{S}}$. If $\hat{\nu} \rightarrow \infty$, then $\mathbf{\Lambda}$ is estimated by $\bar{\mathbf{S}}$. The above estimates give quite satisfactory results on real data. A theoretical analysis of the estimation of our hyperparameters will be given later. For the moment, we content ourselves with obtaining reasonable estimates.

Y. C. TAI AND T. P. SPEED

EB estimation of η . The hyperparameter η is related to the moderated multivariate t -statistic $\tilde{\mathbf{t}}$ of nonzero genes. The method we use to estimate η builds on that of estimating v_0 in [24], except that we now need to deal with the multivariate case. Let $\tilde{t}_j$ be the j th element of the moderated multivariate t -statistic $\tilde{\mathbf{t}}$, $j = 1, \dots, k$. As in Section 6.3 in [24], each $\tilde{t}_j$ gives an estimate of η , call it $\hat{\eta}_j$, based on the top $p/2$ portion of genes with the largest $|\tilde{t}_j|$. We set $\hat{\eta}$ to be the mean of the $\hat{\eta}_j$.

4.2. Unpaired two-sample problem. Suppose there are two independent biological conditions Z and Y with sample sizes m and n , respectively. We can also derive the MB -statistic for testing the null $H: \boldsymbol{\mu}_Z = \boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Z = \boldsymbol{\Sigma}_Y > 0$. The null hypothesis turns out to be the same as that in Section 4.1: $H: \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} > 0$, if we write $\boldsymbol{\mu} = \boldsymbol{\mu}_Z - \boldsymbol{\mu}_Y$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_Z = \boldsymbol{\Sigma}_Y$. That is, we solve this two-sample problem using the one-sample approach in Section 4.1. We denote the m i.i.d. time course vectors for biological condition Z by $\mathbf{Z}_1, \dots, \mathbf{Z}_m$, each from a multivariate normal distribution with mean $\boldsymbol{\mu}_Z$ and variance–covariance matrix $\boldsymbol{\Sigma}$. Similarly, those for biological condition Y are denoted by $\mathbf{Y}_1, \dots, \mathbf{Y}_n$, each with mean $\boldsymbol{\mu}_Y$ and variance–covariance matrix $\boldsymbol{\Sigma}$. Since the null hypothesis here is identical to that in Section 4.1, the priors for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are exactly the same as those in Section 4.1, and we omit the details here. In a later paper we will attack this problem by assigning independent priors for $\boldsymbol{\mu}_Y$ and $\boldsymbol{\mu}_Z$ separately.

All the results follow immediately. The moderated multivariate t -statistic $\tilde{\mathbf{t}}$ here is defined as equation (4.7) with n replaced by $(m^{-1} + n^{-1})^{-1}$ and $\bar{\mathbf{X}}$ replaced by $\bar{\mathbf{Z}} - \bar{\mathbf{Y}}$. $\tilde{\mathbf{S}}$ here is the same as that defined in Section 3.2. The posterior odds $\mathbf{O}$ against the null hypothesis that the expected time courses are the same are

$$\mathbf{O} = \frac{p}{1-p} \left(\frac{m^{-1} + n^{-1}}{m^{-1} + n^{-1} + \eta^{-1}} \right)^{(1/2)k} \times \left(\frac{m + n - 2 + v + \tilde{T}^2}{m + n - 2 + v + ((m^{-1} + n^{-1})/(m^{-1} + n^{-1} + \eta^{-1}))\tilde{T}^2} \right)^{(1/2)(m+n+v-1)}.$$

The log base 10 of $\mathbf{O}$ is our two-sample MB -statistic. Again, when all genes have the same sample sizes m and n , the two-sample MB -statistic is equivalent to the $\tilde{T}^2 = \tilde{\mathbf{t}}\tilde{\mathbf{t}}$. Under H , $k^{-1}\tilde{T}^2$ has an F distribution with degrees of freedom $(k, m + n + v - k - 1)$; equivalently, $(m + n + v - k - 1)^{-1}(m + n + v - 2)\tilde{T}^2$ has the Hotelling T^2 distribution $T^2(k, m + n + v - 2)$.

The MB -statistic described in this section involves hyperparameters v , $\boldsymbol{\Lambda}$, η and p . The estimation procedures for these hyperparameters are very similar to those in Section 4.1, except that we have to use the gene-specific pooled sample variance–covariance matrices when estimating v and $\boldsymbol{\Lambda}$, and use the $k \times 1$ moderated multivariate t -statistic $\tilde{\mathbf{t}}$ defined here to estimate η . We omit the details here.

It should be noted that the MB -statistic derived in this section has a slightly different definition: instead of using all the data we observe, we only use the difference in sample averages and the pooled sample variance–covariance matrix. The

REPLICATED MICROARRAY TIME COURSE DATA

$\tilde{T}^2$ -statistic here is identical to the moderated two-sample Hotelling T^2 -statistic in Section 3.2.

4.3. *One-sample problem of constancy.* In this section we derive the posterior odds against the null that a gene's mean expression level stays constant over time. We obtain a closed form solution similar to that in the preceding sections, but only under a constraint on the variance–covariance matrix Σ .

4.3.1. *Transformation.* For each gene, let I be the indicator variable defined in Section 4.1. Let $\mathbf{P} = k^{-1} \mathbf{1}_k \mathbf{1}_k'$ be the $k \times k$ projection matrix onto the rank 1 space of constant vectors, where $\mathbf{1}_k = (1, \dots, 1)$ is a $k \times 1$ vector of 1s. Let $\mathbf{P}^c = \mathbf{I}_k - \mathbf{P}$ be the projection onto the orthogonal complement of $R(\mathbf{P})$. We can write any vector $\mu \in R^k$ as $\mu = \mathbf{P}\mu + \mathbf{P}^c\mu$, and in the case $I = 0$, the second term $\mathbf{P}^c\mu$ vanishes. As in Section 4.1, we build up our multivariate model by first assigning independent inverse Wishart priors to the gene-specific covariance matrices Σ ; see equation (4.1). Given Σ , we next assign multivariate normal priors to the gene-specific mean parameters μ for the case of nonconstant ($I = 1$) and constant genes ($I = 0$), respectively:

$$(4.11) \quad \begin{cases} \mu | \Sigma, I = 1 \sim N(0, \tau^{-1} \mathbf{P} \Sigma \mathbf{P} + \kappa^{-1} \mathbf{P}^c \Sigma \mathbf{P}^c), \\ \mu | \Sigma, I = 0 \sim N(0, \tau^{-1} \mathbf{P} \Sigma \mathbf{P}). \end{cases}$$

Given Σ and $I = 0$, the covariance matrix $\mathbf{P} \Sigma \mathbf{P}$ guarantees that μ is a constant vector, while when $I = 1$, the extra component $\mathbf{P}^c \Sigma \mathbf{P}^c$ adds further variance to μ so that it becomes a nonconstant vector. Again, in order to obtain the full expression for the posterior odds $\mathbf{O}$, we need to derive $P(\mathbf{X}_1, \dots, \mathbf{X}_n | I)$ using equation (4.3). To get a closed-form expression for the posterior odds, we find it necessary to make an additional assumption, namely, that $\mathbf{P} \Sigma = \Sigma \mathbf{P}$. With this assumption, given Σ and $I = 0$, $\bar{\mathbf{X}}$ is a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $(n^{-1} \Sigma + \tau^{-1} \mathbf{P} \Sigma \mathbf{P})$. Similarly, given Σ and $I = 1$, $\bar{\mathbf{X}}$ is a multivariate normal distribution with mean $\mathbf{0}$ and the covariance matrix $(n^{-1} \Sigma + \tau^{-1} \mathbf{P} \Sigma \mathbf{P} + \kappa^{-1} \mathbf{P}^c \Sigma \mathbf{P}^c)$.

For the rest of this section, unless stated otherwise, we assume $\mathbf{P} \Sigma = \Sigma \mathbf{P}$, and we make use of the following lemma, whose proof is omitted.

LEMMA 4.1. *Suppose $\mathbf{T}$ is any $k \times k$ nonsingular matrix whose first row is constant c and the remaining rows have row sums equal to 0. Write $\mathbf{T} = (\mathbf{T}'_0, \mathbf{T}'_1)'$, where $\mathbf{T}'_0$ is the first row of $\mathbf{T}$, and $\mathbf{T}'_1$ is the remainder. Then, for any $\Sigma > 0$ satisfying $\mathbf{P} \Sigma = \Sigma \mathbf{P}$, $\mathbf{T} \Sigma \mathbf{T}' = \tilde{\Sigma}$ is a $k \times k$ block diagonal matrix with the scalar $\tilde{\sigma}^2 > 0$ as the first block and $(k - 1) \times (k - 1)$ matrix $\tilde{\Sigma}_1 > 0$ as the second block; that is,*

$$\mathbf{T} \Sigma \mathbf{T}' = \tilde{\Sigma} = \begin{pmatrix} \tilde{\sigma}^2 & \mathbf{0} \\ \mathbf{0} & \tilde{\Sigma}_1 \end{pmatrix}.$$

Y. C. TAI AND T. P. SPEED

As the first example, let $\mathbf{T}$ be the Helmert matrix, where the ji th element of $\mathbf{T}$ is defined as

$$\begin{cases} t_{ji} = 1/\sqrt{k}, & \text{for } j = 1, i = 1, \dots, k, \\ t_{ji} = 1/\sqrt{j(j-1)}, & \text{for } 2 \leq j \leq k, 1 \leq i \leq j-1, \\ t_{ji} = -(j-1)/\sqrt{j(j-1)}, & \text{for } 2 \leq j \leq k, i = j, \\ t_{ji} = 0, & \text{for } 2 \leq j \leq k-1, j+1 \leq i \leq k. \end{cases}$$

$\mathbf{T}$ can also be the following matrix, where the ji th element of $\mathbf{T}$ is defined as

$$\begin{cases} t_{ji} = 1, & \text{for } j = 1, i = 1, \dots, k, \\ t_{ji} = 1, & \text{for } 2 \leq j \leq k, i = j-1, \\ t_{ji} = -1, & \text{for } 2 \leq j \leq k, i = j, \\ t_{ji} = 0, & \text{otherwise.} \end{cases}$$

For our multivariate empirical Bayes model in this section, we use the Helmert matrix $\mathbf{T}$ to proceed with our calculations. The results can be applied to other $\mathbf{T}$ immediately.

4.3.2. Models and priors. Here $\mathbf{T}$ is partitioned into its first row $\mathbf{T}_0$ ($1 \times k$) and its last $k-1$ rows $\mathbf{T}_1$ ($(k-1) \times k$). Since $\mathbf{X}_1, \dots, \mathbf{X}_n$ are i.i.d. $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the transformed random vectors $\mathbf{TX}_i$ are also multivariate normally distributed with mean $\mathbf{T}\boldsymbol{\mu}$ and covariance matrix $\tilde{\boldsymbol{\Sigma}}$. By Lemma 4.1, the matrix $\tilde{\boldsymbol{\Sigma}}$ is a block diagonal matrix with $\tilde{\sigma}^2$ as the first block, and $\tilde{\boldsymbol{\Sigma}}_1$ as the second block. Defining $\bar{x}_i = k^{-1} \sum_{j=1}^k X_{ij}$, then $\sqrt{k}\bar{x}_i$ and the random vectors $\mathbf{T}_1\mathbf{X}_i$ are independent and normally distributed, with distributions

$$\begin{aligned} \sqrt{k}\bar{x}_i | \mathbf{T}_0\boldsymbol{\mu}, \tilde{\sigma}^2 &\sim N(\mathbf{T}_0\boldsymbol{\mu}, \tilde{\sigma}^2), \\ \mathbf{T}_1\mathbf{X}_i | \mathbf{T}_1\boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}_1 &\sim N(\mathbf{T}_1\boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}_1). \end{aligned}$$

This transformation allows us to separate the gene expression changes into constant and nonconstant changes.

As we have seen in Section 4.1, the joint distributions of data given I can be fully described using the sufficient statistics $\bar{x}$, $\mathbf{T}_1\bar{\mathbf{X}}$, s^2 and $\mathbf{S}_1$, where $\bar{x} = n^{-1} \sum_{i=1}^n \bar{x}_i$, $\mathbf{T}_1\bar{\mathbf{X}} = n^{-1} \sum_{i=1}^n \mathbf{T}_1\mathbf{X}_i$, $s^2 = (n-1)^{-1} \sum_{i=1}^n (\bar{x}_i - \bar{x})^2$ and $\mathbf{S}_1 = (n-1)^{-1} \sum_{i=1}^n (\mathbf{T}_1\mathbf{X}_i - \mathbf{T}_1\bar{\mathbf{X}})(\mathbf{T}_1\mathbf{X}_i - \mathbf{T}_1\bar{\mathbf{X}})'$. The prior for $\tilde{\boldsymbol{\Sigma}}$ is first set through the independent priors for $\tilde{\sigma}^2$ and $\tilde{\boldsymbol{\Sigma}}_1$. We suppose that $\tilde{\sigma}^2$ and $\tilde{\boldsymbol{\Sigma}}_1$ are independently distributed, with an inverse gamma distribution with shape parameter $\xi/2$ and scale parameter $\xi\lambda^2/2$, and an inverse Wishart distribution with degrees of freedom ν and scale matrix $\nu\mathbf{A}_1$, respectively, that is,

$$(4.12) \quad \begin{cases} \tilde{\sigma}^2 \sim \text{inv-gamma}(\frac{1}{2}\xi, \frac{1}{2}\xi\lambda^2), \\ \tilde{\boldsymbol{\Sigma}}_1 \sim \text{inv-Wishart}_\nu((\nu\mathbf{A}_1)^{-1}). \end{cases}$$

REPLICATED MICROARRAY TIME COURSE DATA

The prior for $\mathbf{T}\boldsymbol{\mu}$ has four parts. We assign independent priors to $\mathbf{T}_0\boldsymbol{\mu}$ and $\mathbf{T}_1\boldsymbol{\mu}$ separately for the cases $I = 1$ and $I = 0$. For the case $I = 1$, priors are

$$(4.13) \quad \begin{cases} \mathbf{T}_0\boldsymbol{\mu}|\tilde{\sigma}^2, I = 1 \sim N(\theta, \kappa^{-1}\tilde{\sigma}^2), \\ \mathbf{T}_1\boldsymbol{\mu}|\tilde{\boldsymbol{\Sigma}}_1, I = 1 \sim N(\mathbf{0}, \eta^{-1}\tilde{\boldsymbol{\Sigma}}_1), \end{cases}$$

where $\theta \geq 0$ is the mean, and $\kappa > 0$ and $\eta > 0$ are scale parameters. When $I = 0$, $\mathbf{T}_1\boldsymbol{\mu} = \mathbf{0}$ with probability 1. Thus, the priors in this case are

$$(4.14) \quad \begin{cases} \mathbf{T}_0\boldsymbol{\mu}|\tilde{\sigma}^2, I = 0 \sim N(\theta, \kappa^{-1}\tilde{\sigma}^2), \\ \mathbf{T}_1\boldsymbol{\mu}|\tilde{\boldsymbol{\Sigma}}_1, I = 0 \equiv \mathbf{0}. \end{cases}$$

It is reasonable to assume $P(\mathbf{T}_0\boldsymbol{\mu}|\tilde{\sigma}^2, I = 0) = P(\mathbf{T}_0\boldsymbol{\mu}|\tilde{\sigma}^2, I = 1)$ for large genome-wide arrays since there is no obvious reason why the expected grand mean of the expression levels for nonconstant genes should differ from that of constant genes. For two-color comparative microarray experiments, it is also reasonable to assume $\theta = 0$.

4.3.3. Multivariate joint distributions. The joint distributions can be derived quite readily using a previously established formula, and so we omit the details. Given $I = 1$, $\mathbf{T}_1\bar{\mathbf{X}}$ and $\mathbf{S}_1$ follow the Student–Siegel distribution $StSi_{k-1}(\nu, \mathbf{0}, (n^{-1} + \eta^{-1})\boldsymbol{\Lambda}_1, n - 1, (n - 1)^{-1}\nu\boldsymbol{\Lambda}_1)$. Similarly, the joint distribution of $\mathbf{T}_1\bar{\mathbf{X}}$ and $\mathbf{S}_1$ given $I = 0$ is $StSi_{k-1}(\nu, \mathbf{0}, n^{-1}\boldsymbol{\Lambda}_1, n - 1, (n - 1)^{-1}\nu\boldsymbol{\Lambda}_1)$.

4.3.4. MB-statistic and $\tilde{T}^2$ -statistic. The posterior odds against the null that a gene's mean expression level stays constant over time are equation (4.6) in Section 4.1 with k replaced by $k - 1$, $\tilde{\mathbf{t}}$ expressed by equation (4.7) with $\tilde{\mathbf{S}}$ replaced by $\tilde{\mathbf{S}}_1$ and $\bar{\mathbf{X}}$ replaced by $\mathbf{T}_1\bar{\mathbf{X}}$. $\tilde{\mathbf{S}}_1$ is just equation (4.5) with $\mathbf{S}$ replaced by $\mathbf{S}_1$ and $\boldsymbol{\Lambda}$ replaced by $\boldsymbol{\Lambda}_1$. As in Section 4.1, the MB-statistic is a monotonic increasing function of $\tilde{T}^2 = \tilde{\mathbf{t}}'\tilde{\mathbf{t}}$ when all genes have the same sample size n .

Under H , $(k - 1)^{-1}\tilde{T}^2$ has an F distribution with degrees of freedom $(k - 1, n + \nu - k + 1)$, or, equivalently, $(n + \nu - k + 1)^{-1}(n + \nu - 1)\tilde{T}^2$ has a Hotelling T^2 -distribution $T^2(k - 1, n + \nu - 1)$. The hyperparameter estimation procedures here are similar to those described in Section 4.1, except that all the estimations are performed based on transformed data.

5. Case study. In this section we illustrate our results with a paired two-sample problem, using the *Arabidopsis thaliana* dataset in [35]. Here we only give a very brief description of the data and the results. We refer the reader to [35] and to Chapter 5 of [28] for more thorough discussions.

A. thaliana wildtype (Columbia) and ics1-2 null mutant plants were evenly positioned, intermixed and grown in growth chambers under controlled conditions. When the plants were four weeks old, they were infected with a moderately heavy

Y. C. TAI AND T. P. SPEED

innoculum of the powdery mildew *G. orontii*. Each pair of mRNA samples from wildtype and mutant plants was harvested and collected at six time points post-infection. Plants could not be resampled, so mRNA samples at one time point were from different plants than those of any other time point. We report here on the analysis of three replicate experiments under similar environmental conditions which contribute four biological replicates: one from the first and third experiments (1–3, 3–1) and two from the second experiment (2–1, 2–2). These mRNA samples were hybridized onto Affymetrix *Arabidopsis* ATH1 GeneChips, yielding 22,810 probesets and 48 arrays in our analysis. The array preprocessing were done using the Robust Multi-array Analysis (RMA) algorithm described in [18, 5] which is implemented in the Bioconductor package `affy`.

This study is longitudinal if we treat *experiments* as units, while it is cross-sectional if we treat *plants* as units. We believe it is worthwhile to treat it as a paired longitudinal study, since samples within the same experiment are more similar than those from different experiments. We thus have a paired two-sample problem, with the genes of interest being those whose wildtype and mutant temporal profiles are different. We subtracted the $\log_2$ intensities of the wildtype from those of the paired mutant at each time within each replicate, yielding the $\log_2$ ratios of mutant relative to wildtype. Since the number of time course replicates is the same ($n = 4$) across genes for this dataset, we used the $\tilde{T}^2$ -statistic instead of the MB -statistic to rank genes, so that we did not have to estimate the hyperparameter η .

For comparison purposes, we fitted a linear model to the log ratios for each gene, with time and replicate effects, and calculated the F -statistic for the time effect.

5.1. Results. The extent of moderation from $\hat{\nu} = 5$ was 63%. The left panel of Figure 1 presents three genes falling into different ranges of ranks (rank = 1, 175, 859) by $\tilde{T}^2$, while the ones on the right panel have the same ranks by F . The gene ranked most highly by $\tilde{T}^2$ exhibits much greater differences between the wildtype and mutant temporal profiles than the one ranked most highly by F . The magnitude of the difference, as measured by $\tilde{T}^2$, decreases as the rank goes down. The gene ranked 1 by $\tilde{T}^2$ is well known: pathogenesis-related protein 1 (PR1). Other known pathogenesis-related genes also ranked highly by $\tilde{T}^2$ and were less highly ranked by the F -statistic, as detailed in [35].

To investigate the effect of the amount of moderation on gene ranking, we kept Λ fixed, and re-calculated the $\tilde{T}^2$ -statistic with several different ν values. We then computed the Spearman rank correlation between the different sets of $\tilde{T}^2$ -statistics for all genes and for the top 859 genes only. Table 1 gives the results. The correlations are lower in the two extremes. All of these sets have the same number one gene. This comparison shows that the gene ranks are reasonably stable when the extent of moderation varies within a certain window, and that moderation seems to have more effect on the top genes relative to the whole gene set.

REPLICATED MICROARRAY TIME COURSE DATA

FIG. 1. Genes of ranks 1, 175, 859 by the paired two-sample $\tilde{T}^2$ -statistic in Section 4.1 (left panel) and the F -statistic (right panel). The temporal difference between wildtype and mutant decreases as the rank by $\tilde{T}^2$ goes down. The genes on the left panel all show larger differences between the wildtype and mutant than the corresponding ones on the right panel.

Y. C. TAI AND T. P. SPEED

TABLE 1
Spearman rank correlations between $\tilde{T}^2$ with different ν and the estimated ν for all and the top 859 genes. The percent moderation is defined by $(\nu + n - 1)^{-1}\nu \times 100$

% moderation	Correlation (all)	Correlation (top 859)
97 ($\nu = 100$)	0.97	0.90
80 ($\nu = 12$)	0.99	0.98
40 ($\nu = 2$)	0.99	0.98
25 ($\nu = 1$)	0.98	0.96
0 ($\nu = 0.01$)	0.93	0.90

6. Simulation study.

6.1. *Method.* In this section we report on a small simulation study for the null hypothesis $H: \boldsymbol{\mu} = \mu_0 \mathbf{1}$, $\boldsymbol{\Sigma} > 0$ based on an actual example we have met. We simulate 100 data sets, each with 20,000 genes. The genes are simulated independently, which we regard as an assumption that makes sense to compare methods, but it should be kept in mind that gene expression measures in real data can be quite dependent. In each simulated data set, 400 out of the 20,000 genes are assigned to be nonconstant. That is, $p = 0.02$. Each gene is simulated with three independent replicates ($n = 3$) and eight time points ($k = 8$). The other hyperparameters are the following: $\nu = 13$, $\xi = 3$, $\lambda^2 = 0.3$, $\theta = 0$ (two-color experiments), $\kappa = 0.02$, $\eta = 0.08$, and

$$\mathbf{\Lambda} = \begin{pmatrix} 14.69 & 0.57 & 0.99 & 0.40 & 0.55 & 0.51 & -0.23 \\ 0.57 & 15.36 & 1.22 & 0.84 & 1.19 & 0.91 & 0.86 \\ 0.99 & 1.22 & 14.41 & 2.47 & 1.81 & 1.51 & 1.07 \\ 0.40 & 0.84 & 2.47 & 17.05 & 2.40 & 2.32 & 1.33 \\ 0.55 & 1.19 & 1.81 & 2.40 & 15.63 & 3.31 & 2.75 \\ 0.51 & 0.91 & 1.51 & 2.32 & 3.31 & 13.38 & 3.15 \\ -0.23 & 0.86 & 1.07 & 1.33 & 2.75 & 3.15 & 12.90 \end{pmatrix} \times 10^{-3}.$$

The correlation matrix of $\mathbf{\Lambda}$ is

$$\begin{pmatrix} 1 & 0.04 & 0.07 & 0.03 & 0.04 & 0.04 & -0.02 \\ 0.04 & 1 & 0.08 & 0.05 & 0.08 & 0.06 & 0.06 \\ 0.07 & 0.08 & 1 & 0.16 & 0.12 & 0.11 & 0.08 \\ 0.03 & 0.05 & 0.16 & 1 & 0.15 & 0.15 & 0.09 \\ 0.04 & 0.08 & 0.12 & 0.15 & 1 & 0.23 & 0.20 \\ 0.04 & 0.06 & 0.11 & 0.15 & 0.23 & 1 & 0.24 \\ -0.02 & 0.06 & 0.08 & 0.09 & 0.19 & 0.24 & 1 \end{pmatrix},$$

REPLICATED MICROARRAY TIME COURSE DATA

and we see clear evidence of serial correlation. Note that, in the real world, although often the case, the correlation does not always decrease with time lag. The statistics compared in our study are the following:

- (1) *MB*-statistic, or equivalently, the $\tilde{T}^2$ -statistic;
- (2) *MB*-statistic using first differences: take the differences in gene expression values at consecutive time points within replicates, and use them to test the null hypothesis $H : \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} > \mathbf{0}$ (Section 4.1), where $\boldsymbol{\mu}$ is the mean of the differences;
- (3) *MB*-statistic in the special case $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_k$;
- (4) *MB*-statistic in the limiting case $\nu \rightarrow \infty$;
- (5) *MB*-statistic in the limiting case $\nu \rightarrow 0$;
- (6) ordinary *F*-statistic from an ANOVA model with time and replicate effects;
- (7) partly-moderated *F*-statistic proposed in [24] from an ANOVA model with time and replicate effects;
- (8) one-sample moderated Hotelling T^2 -statistic $\|n^{1/2}\tilde{\mathbf{S}}^{-1/2}\tilde{\mathbf{d}}\|^2$ derived in Section 3, equivalently, the moderated *LR*-statistic, where the degree of moderation and the common matrix toward which each sample covariance matrix moves is estimated by the method given in Section 4.1;
- (9) the variance across time course replicates $(nk - 1)^{-1} \sum_{i=1}^n \sum_{j=1}^k (X_{ij} - \bar{x})^2$.

Here each of the nine statistics incorporates either none (e.g., variance) or one (ordinary *F*-statistic) or more of the following: moderation, correlation structure and replicate variances, and thus can be used to show the importance of the above properties. It is not appropriate to set the prior degrees of freedom ν to be a very small number, since we have the constraint that $\nu \geq k - 1$. We choose ν to be $k + 5 = 13$ because it simulates more stable $\boldsymbol{\Sigma}$'s across genes.

6.2. Results. Table 2 compares the means and standard deviations of the hyperparameter estimates of the diagonal elements of $\boldsymbol{\Lambda}_1$ (λ_j^2), $j = 1, \dots, k - 1$, with their true values. The mean estimate of $\boldsymbol{\Lambda}_1$ is very close to the true $\boldsymbol{\Lambda}_1$, and the standard deviations are very small. The hyperparameter η is always underestimated (mean = 0.026, SD = 0.002), which agrees with Section 8 in [24], where ν_0 was usually over-estimated. The hyperparameter ν is also always underestimated (mean = 7.0, SD = 0.2). In Section 5 we observed that the amount of moderation ν does not greatly affect gene ranking except at the two extremes. One can even choose a user-defined ν which gives reasonable results. Although not well estimated, η only affects the rankings when the number of replicates is different across genes. However, this does not happen often in the real world. Even when that happens, the effect is very small. To investigate the effects of η on gene rankings, we tried a couple of η values from different ranges, while keeping the remaining hyperparameters fixed, and calculated the *MB*-statistics. The rank correlations between rankings of the *MB*-statistics with the user-defined η 's and the estimated η for one set of simulated data are the following: 0.91, 0.94, 0.99, 0.99, 0.99 for $\eta = 2, 1, 0.08$ (true value), 0.05, 0.001, respectively.

Y. C. TAI AND T. P. SPEED

TABLE 2
The means and standard deviations (SD) of the diagonal elements
of the estimated Λ_1

Hyperparameters	True value $\times 10^3$	Mean $\times 10^3$	SD $\times 10^3$
λ_1^2	14.69	14.71	0.16
λ_2^2	15.36	15.37	0.17
λ_3^2	14.41	14.43	0.15
λ_4^2	17.05	17.04	0.19
λ_5^2	15.63	15.63	0.15
λ_6^2	13.38	13.40	0.15
λ_7^2	12.90	12.92	0.17

To examine the relationship between the $\tilde{T}^2$ -statistic and the true deviation from constancy, the $\log_{10}$ transformed $\tilde{T}^2$ -statistic from one simulated dataset is plotted against the Mahalanobis distance between the expected time course vector $\boldsymbol{\mu}$ and its projection onto the rank 1 constant space $\bar{\boldsymbol{\mu}} = \mathbf{P}\boldsymbol{\mu}$ (Figure 2). The squared Mahalanobis distance is defined by $d(\boldsymbol{\mu}, \bar{\boldsymbol{\mu}})^2 = (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})$. Figure 2 clearly

FIG. 2. The $\log_{10} \tilde{T}^2$ statistic versus the true deviation from constancy $d(\boldsymbol{\mu}, \bar{\boldsymbol{\mu}})$ for one simulated dataset. Here 1 denotes nonconstant, and o constant genes.

FIG. 3. Average number of false positives versus number of false negatives of all the nine statistics. The subplot presents the curves for the best seven statistics.

shows that $\log_{10} \tilde{T}^2$ is positively correlated with $d(\mu, \bar{\mu})$, and most of the 400 true nonconstant genes achieve higher $\tilde{T}^2$ -statistics than the constant genes.

Figure 3 plots the average number of false positives against average number of false negatives at different cutoffs. By different cutoffs, we mean choosing the top x genes and calculating the numbers of false positives and false negatives, where x varies across the integers from 400 to 800. The lines in Figure 3 from left to right represent the following: MB -statistic ($\tilde{T}^2$), MB -statistic with first differences (indistinguishable from the MB -statistic), one-sample moderated Hotelling T^2 -statistic (indistinguishable from the MB -statistic), MB -statistic with $\Sigma = \sigma^2 \mathbf{I}_k$, MB -statistic with $\nu \rightarrow \infty$, partly-moderated F -statistic [24], ordinary F -statistic, MB -statistic with $\nu \rightarrow 0$ and variance. The MB -statistic ($\tilde{T}^2$) attains almost the same numbers of false positives and false negatives as MB with first differences and the one-sample moderated Hotelling T^2 -statistic. The effectiveness of moderation is demonstrated by comparing the lines for the MB -statistic, the MB -statistic in the limiting case $\nu \rightarrow \infty$ and the MB -statistic in the limiting case that $\nu \rightarrow 0$. Both of these limiting cases lead to higher aggregate false positives and false negatives. This result supports the view stated in [29] that moderation is useful. In particular, the case $\nu \rightarrow 0$ (no moderation at all) produces much higher numbers of false positives and false negatives. This is likely due to the poor estimation of sample variance-covariance matrices with a small number of replicates. Indeed, the ordinary unmoderated F -statistic which ignores the correlation structure achieves smaller numbers of false positives and false negatives than the unmoderated MB -statistic. A similar situation also arises in the microarray discrimination context; see Section 7 of [12]. The partly-moderated F -statistic [24] which

Y. C. TAI AND T. P. SPEED

ignores the dependence among times behaves like the *MB*-statistic in the special case $\Sigma = \sigma^2 \mathbf{I}_k$. Moreover, it achieves fewer false positives and false negatives than the ordinary *F*-statistic. Figure 3 also demonstrates the importance of incorporating the correlation structure among time points. The *MB*-statistic ($\tilde{T}^2$) and the one-sample moderated Hotelling T^2 -statistic perform better than the partly-moderated *F*-statistic in [24] and the ordinary *F*-statistic; the former incorporates the correlation structure among time points, whereas the latter does not. However, we observe that the amount of moderation given by the partly-moderated *F*-statistic in [24] is usually much less than that given by the *MB*-statistic. When there are a large number of residual degrees of freedom from the linear model, the partly-moderated *F*-statistic [24] behaves very much like the ordinary *F*-statistic. This suggests that the lower number of false positives and number of false negatives from the *MB*-statistic than the partly-moderated *F*-statistic in [24] involve both the incorporation of correlation structures and the amounts of moderation. As expected, the simple variance statistic across replicates, which totally ignores the replicate variances, performs the worst. This demonstrates the importance of incorporating the replicate variances into any statistic.

7. Discussion. In this paper we introduced the *MB*- and $\tilde{T}^2$ -statistics for one- and two-sample longitudinal replicated developmental microarray time course experiments. Our main focus was the one-sample or paired two-sample problem with the null hypothesis $H: \mu = \mathbf{0}, \Sigma > \mathbf{0}$. This *MB*-statistic can be used when there are two biological conditions and the samples are paired across conditions, and it is shown to perform better than the classical *F*-statistic on a problem briefly described in Section 5. In addition, we also derive the *MB*-statistics and $\tilde{T}^2$ -statistics for the two-sample problem with the null $H: \mu_Z = \mu_Y, \Sigma_Z = \Sigma_Y = \Sigma > \mathbf{0}$, and the one-sample problem with the null $H: \mu = \mu_0 \mathbf{1}, \Sigma > \mathbf{0}$ using similar approaches. The latter situation requires a slight assumption on Σ in order to get a simple closed-form solution for the posterior odds against the null. All the *MB*-statistics and $\tilde{T}^2$ -statistics incorporate the correlation structure, replication and moderation. The moderated versions of some standard likelihood-ratio test statistics are also described. When all genes have the same sample size(s), our $\tilde{T}^2$ -statistics are not only equivalent to the *MB*-statistics, but also are identical to their corresponding moderated Hotelling T^2 -statistics, apart from the one-sample problem with the null $H: \mu = \mu_0 \mathbf{1}, \Sigma > \mathbf{0}$, where there is an additional constraint on Σ . In this case the $\tilde{T}^2$ -statistic performed as well as the moderated Hotelling T^2 -statistic in our simulation study, and also on several real datasets we have encountered. We have shown in the simulation study that, with this null, the *MB*-statistic ($\tilde{T}^2$), the *MB*-statistic using first differences and the one-sample moderated Hotelling T^2 -statistic perform best among all the nine statistics compared. This is not entirely surprising given that we simulated data under our model, but the comparisons are still informative. In practice, we regard the *MB*-statistic,

REPLICATED MICROARRAY TIME COURSE DATA

the *MB*-statistic with first differences and the moderated *LR*-statistic as performing equally well, giving very similar (or identical) results. However, to use the *LR*-statistic (the moderated Hotelling T^2 -statistic), we still need to insert moderated sample variance–covariance matrices, and these come from our multivariate empirical Bayes framework. In other words, our models provide a natural way to estimate the gene-specific moderated variance–covariance matrices (Sections 4.1–4.3), while the likelihood-based approach alone does not.

The assumption of $\mathbf{P}\mathbf{\Sigma} = \mathbf{\Sigma}\mathbf{P}$ with the null $H : \boldsymbol{\mu} = \mu_0\mathbf{1}$, $\mathbf{\Sigma} > 0$ allows the mathematical calculations in Section 4.3, and leads to a closed-form formula for the *MB*-statistic. One question which naturally arises to be the impact of this constraint on the rankings of genes. From the practical point of view, the impact of this constraint on gene rankings is very *slight*. The rank correlations between the one-sample *MB*-statistic *with* the commuting assumption and the moderated *LR*-statistic from likelihood-based approach *without* the constraint, from the actual examples we have met, are typically very high (over 0.99). The rank correlations from our simulated data are also over 0.99. On the other hand, using the *MB*-statistic with first differences produces very similar or even identical results to the *MB*-statistic in Section 4.3. Indeed, instead of using the Helmert matrix, if we choose $\mathbf{T}$ to be the second transformation matrix of Section 4.3, we get identical results. Even so, we do not consider the first differences approach to be the solution to this commuting constraint, since the inference drawn is based on reduced, not the original data. In other words, the null hypotheses are not equivalent. The likelihood-based approach with moderation and the first differences approach support the fact that this constraint does not have much effect on the results. In addition, the former is a good way to avoid this assumption, since it performs as well as our *MB*-statistic in Section 4.3.

The statistics proposed in this paper are for one- and two-sample longitudinal data. We should be aware that many experiments in the real world exhibit some features from both longitudinal and cross-sectional experiments (e.g., Section 5).

One thing we plan to investigate in the future is the effect of assuming the same variance–covariance matrix $\mathbf{\Sigma}$ for both $I = 1$ and $I = 0$. Another issue which interests us is the effect of assuming the same $\mathbf{\Sigma}$ across biological conditions in the unpaired two-sample model in Section 4.2. The proposed methods may be extended in several ways, for example, identifying genes of some *specific* pattern, rather than *any* pattern. The statistics for a longitudinal multi-sample problem when there are at least three biological conditions and genes of interest are those with different temporal profiles across conditions derived in [30]. The corresponding statistics for cross-sectional data are also presented in [31].

The proposed methods focus on gene ranking, but not assessing the *significance* using p -values. However, if desired, we believe that generating p -values from a bootstrap analysis should be successful in this context.

We constructed our models using conjugate priors for the multivariate normal likelihoods, so that we got simple closed-form solutions for the posteriors odds.

Y. C. TAI AND T. P. SPEED

Finding a closed-form statistic when the priors on μ and Σ are independent seems to be an open and probably hard problem; that problem probably needs to be dealt with using MCMC.

Acknowledgments. We thank the Editor Jianqing Fan, an Associate Editor and the two referees for their valuable comments on this paper. We are grateful to Mary Wildermuth and her colleagues for sharing their *A. thaliana* dataset prior to publication of Wildermuth et al. [35]. Thanks are also due to Gordon Smyth, Ingrid Lönnstedt, Darlene Goldstein, Greg Hather, Avner Bar-Hen, Alfred Hero, Cavan Reilly and Christina Kendzierski for their valuable comments on an earlier draft, and to Ben Bolstad for his assistance with our simulations. Finally, we would like to acknowledge Mary Wildermuth, John Ngai and their lab members for their helpful discussions on the biological background of time course experiments and access to their data to test the methods developed in this paper.

REFERENCES

- [1] AITCHISON, J. and DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge Univ. Press. MR0408097
- [2] BALDI, P. and LONG, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* **17** 509–519.
- [3] BAR-JOSEPH, Z., GERBER, G., SIMON, I., GIFFORD, D. K. and JAAKKOLA, T. S. (2003). Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proc. Natl. Acad. Sci. USA* **100** 10,146–10,151. MR1998142
- [4] BICKEL, P. J. and DOKSUM, K. A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*, 2nd ed. **1**. Prentice Hall, Upper Saddle River, NJ.
- [5] BOLSTAD, B., IRIZARRY, R., ÅSTRAND, M. and SPEED, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics* **19** 185–193.
- [6] BROBERG, P. (2003). Statistical methods for ranking differentially expressed genes. *Genome Biology* **4** R41.
- [7] CHO, R., CAMPBELL, M., WINZELER, E., STEINMETZ, L., CONWAY, A., WODICKA, L., WOLFSBERG, T., GABRIELIAN, A., LANDSMAN, D., LOCKHART, D. and DAVIS, R. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* **2** 65–73.
- [8] CHO, R., HUANG, M., CAMPBELL, M., DONG, H., STEINMETZ, L., SAPINOSO, L., HAMPTON, G., ELLEDGE, S., DAVIS, R. and LOCKHART, D. (2001). Transcriptional regulation and function during the human cell cycle. *Nature Genetics* **27** 48–54.
- [9] CHU, S., DERISI, J., EISEN, M., MULHOLLAND, J., BOTSTEIN, D., BROWN, P. O. and HERSKOWITZ, I. (1998). The transcriptional program of sporulation in budding yeast. *Science* **282** 699–705.
- [10] DIGGLE, P. J. (1990). *Time Series: A Biostatistical Introduction*. Oxford Univ. Press, New York. MR1055357
- [11] DIGGLE, P. J., HEAGERTY, P., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. Oxford Univ. Press, New York. MR2049007
- [12] DUDOIT, S., FRIDLYAND, J. and SPEED, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.* **97** 77–87. MR1963389

REPLICATED MICROARRAY TIME COURSE DATA

- [13] DUDOIT, S., YANG, Y. H., CALLOW, M. and SPEED, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica* **12** 111–139. MR1894191
- [14] EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571
- [15] GUO, X., QI, H., VERFAILLIE, C. M. and PAN, W. (2003). Statistical significance analysis of longitudinal gene expression data. *Bioinformatics* **19** 1628–1635.
- [16] GUPTA, A. and NAGAR, D. (2000). *Matrix Variate Distributions*. Chapman and Hall/CRC, Boca Raton, FL. MR1738933
- [17] HONG, F. and LI, H. (2006). Functional hierarchical models for identifying genes with different time-course expression profiles. *Biometrics* **62** 534–544.
- [18] IRIZARRY, R. A., BOLSTAD, B. M., COLLIN, F., COPE, L. M., HOBBS, B. and SPEED, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31** e15.
- [19] KENDZIORSKI, C., NEWTON, M., LAN, H. and GOULD, M. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22** 3899–3914.
- [20] LÖNNSTEDT, I. and SPEED, T. P. (2002). Replicated microarray data. *Statist. Sinica* **12** 31–46. MR1894187
- [21] MARDIA, K., KENT, J. and BIBBY, J. (1979). *Multivariate Analysis*. Academic Press, New York. MR0560319
- [22] PARK, T., YI, S.-G., LEE, S., LEE, S. Y., YOO, D.-H., AHN, J.-I. and LEE, Y.-S. (2003). Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics* **19** 694–703.
- [23] REINER, A., YEKUTIELI, D. and BENJAMINI, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19** 368–375.
- [24] SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3** article 3. MR2101454
- [25] SPELLMAN, P. T., SHERLOCK, G., ZHANG, M. Q., IYER, V. R., ANDERS, K., EISEN, M. B., BROWN, P. O., BOTSTEIN, D. and FUTCHER, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* **9** 3273–3297.
- [26] STORCH, K.-F., LIPAN, O., LEYKIN, I., VISWANATHAN, N., DAVIS, F. C., WONG, W. H. and WEITZ, C. J. (2002). Extensive and divergent circadian gene expression in liver and heart. *Nature* **417** 78–83.
- [27] STOREY, J., XIAO, W., LEEK, J. T., TOMPKINS, R. G. and DAVIS, R. W. (2005). Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci. USA* **102** 12,837–12,842.
- [28] TAI, Y. C. (2005). Multivariate empirical Bayes models for replicated microarray time course data. Ph.D. dissertation, Div. Biostatistics, Univ. California, Berkeley.
- [29] TAI, Y. C. and SPEED, T. P. (2005). Statistical analysis of microarray time course data. In *DNA Microarrays* (U. Nuber, ed.) Chapter 20. Chapman and Hall/CRC, New York.
- [30] TAI, Y. C. and SPEED, T. P. (2005). Longitudinal microarray time course *MB*-statistic for multiple biological conditions. Dept. Statistics, Univ. California, Berkeley. In preparation.
- [31] TAI, Y. C. and SPEED, T. P. (2005). Cross-sectional microarray time course *MB*-statistic. Dept. Statistics, Univ. California, Berkeley. In preparation.
- [32] TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S., DMITROVSKY, E., LANDER, E. S. and GOLUB, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96** 2907–2912.

Y. C. TAI AND T. P. SPEED

- [33] TUSHER, V. G., TIBSHIRANI, R. and CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98** 5116–5121.
- [34] WEN, X., FUHRMAN, S., MICHAELS, G. S., CARR, D. B., SMITH, S., BARKER, J. L. and SOMOGYI, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA* **95** 334–339.
- [35] WILDERMUTH, M. C., TAI, Y. C., DEWDNEY, J., DENOUEX, C., HATHER, G., SPEED, T. P. and AUSUBEL, F. M. (2006). Application of $\tilde{T}^2$ statistic to temporal global Arabidopsis expression data reveals known and novel salicylate-impacted processes. To appear.
- [36] YUAN, M. and KENDZIORSKI, C. (2006). Hidden Markov models for microarray time course data in multiple biological conditions (with discussion). *J. Amer. Statist. Assoc.* **101** 1323–1340.

DIVISION OF BIOSTATISTICS
367 EVANS HALL 3860
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720-3860
USA
E-MAIL: yuchuan@stat.berkeley.edu
terry@stat.berkeley.edu