THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Normalization and Differential Gene Expression Analysis of Microarray Data

Magnus Åstrand

CHALMERS | GÖTEBORG UNIVERSITY

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University
Göteborg, Sweden, 2008

Normalization and Differential Gene Expression Analysis of Microarray Data
Magnus Åstrand
ISBN 978-91-7385-043-8

Doktorsavhandlingar vid Chalmers Tekniska Högskola
Ny serie Nr 2724
ISSN 0346-718X
Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Printed at the Department of Mathematical Sciences
Göteborg, Sweden, 2008

Normalization and Differential Gene Expression Analysis of Microarray Data
Magnus Åstrand
Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University

Abstract

DNA microarray technologies have the capability of simultaneously measuring the abundance of thousands of mRNA-sequences. Analysis of microarray data involves many different steps such as image analysis, background correction, and normalization, but also more classical statistical analysis such as testing for significant differences between groups of arrays.

The work presented in this thesis is focused on Affymetrix GeneChip arrays and deals with normalization and the problem of finding differentially expressed genes. Normalization of microarray data is essential to allow between-array comparisons. A procedure called Contrast Normalization is proposed and compared with existing methods together with two additional presented methods, Cyclic-Loess and Quantile Normalization. All three presented methods improve on the performance of the existing methods with a slight edge for Quantile Normalization.

The quality of microarray data often varies between arrays. A model called WAME has been proposed, using a global covariance matrix to account for differing variances and array-to-array correlations, and thus WAME defines a weighted analysis for finding differentially expressed genes. This thesis presents two new methods for estimating the covariance matrix. Both methods show superior computer run-time over the existing method. Moreover, the second proposed method greatly reduces the bias of the existing method when used on simulated data with regulated genes, although to a less degree for real data with many regulated genes.

Microarray data frequently shows a dependency between variability and intensity level which is ignored by the majority of moderated t-tests. The WAME model is extended to incorporate this dependency, and two locally moderated t-tests are proposed, Probe level Locally moderated Weighted median-t (PLW), and Locally Moderated Weighted-t (LMW). When compared with 12 existing methods on 5 spike-in data sets, the PLW method produces the most accurate ranking of regulated genes in 4 out of the 5 data sets, whereas LMW consistently performs better than all (globally) moderated t-tests.

Keywords: microarray; differential expression; gene expression; empirical Bayes; locally moderated; moderated statistic; weighted statistic; PLW; LMW

List of publications

This thesis is based on the work contained in the following papers:

Paper I M. Åstrand (2003) Contrast normalization of oligonucleotide arrays, *Journal of Computational Biology*: 10(1), 95-102.

Paper II B.M. Bolstad, R.A. Irizarry, M. Åstrand, and T.P Speed (2003) A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance, *Bioinformatics*: 19(2), 185-193, 2003.

Paper III M. Åstrand, P. Mostad, and M. Rudemo (2007) Improved Covariance Matrix Estimators for Weighted Analysis of Microarray Data, *Journal of Computational Biology*: 14(10), 1353-1367.

Paper IV M. Åstrand, P. Mostad, and M. Rudemo (2007) Empirical Bayes models for multiple probe type arrays at the probe level. *Submitted to BMC Bioinformatics*

Contents

1 Introduction **1**

2 Background **1**
2.1 Affymetrix technology . . . 1
2.1.1 DNA and RNA . . . 1
2.1.2 The arrays . . . 2
2.1.3 Sample preparation . . . 2
2.2 Low-level analysis . . . 3
2.2.1 Affymetrix MAS 5.0 . . . 4
2.2.2 RMA . . . 4
2.2.3 Model-based expression indexes . . . 5
2.3 Finding differentially expressed genes . . . 6
2.3.1 Probe-set level analysis . . . 7
2.3.2 Probe level analysis . . . 8

3 Summary of papers **9**
Paper I . . . 9
Paper II . . . 9
Paper III . . . 10
Paper IV . . . 12

4 Complementary studies **13**
4.1 PLW and LMW combined with GCRMA and MAS5 . . . 13
4.2 Using LMW on two-color array data . . . 14

Appendixes **18**

A Expression indexes **18**
A.1 Affymetrix MAS 5.0 . . . 18
A.2 RMA . . . 19
A.3 GCRMA . . . 20

B Finding differentially expressed genes **21**
B.1 SAM . . . 21
B.2 Efron-t . . . 22
B.3 LIMMA . . . 22
B.4 IBMT . . . 23
B.5 Shrink-t . . . 23
B.6 LPE . . . 24
B.7 WAME . . . 24
B.8 LMW . . . 25
B.9 PLW . . . 26

References **28**

Acknowledgements

Truly, "no man is an island", and indeed these lines would never have been written without the encouragement and guidance I have received from quite a few people. My adviser during this period as a PhD student, the third and presumable the last period, has been Mats Rudemo. Always positive, and with great competence, Mats have together with my co-adviser Petter Mostad given me exactly the support I have needed.

I also would like to send a special thank to Professor Terry Speed, who inspired me to finalize the ideas for normalization of micro-arrays which in the end resulted in my first published paper. At the same time I would like to thank my second co-adviser and boss at AstraZeneca Ziad Taib, who together with Anders Carlquist made this third period as a PhD student possible.

Furthermore, I thank all of you that are not mentioned above: my colleagues at the department, the lunch-gang, and my colleagues at AstraZeneca. The journey would not have been the same without you.

Finally, my warmest appreciation goes to my ever so wonderful wife and life-companion Annika and the rest of my family. Without you the journey would have been meaningless.

Magnus Åstrand
Göteborg
January 2008

1 Introduction

Proteins are the major active elements of cells. They perform many key functions of biological systems and they are the structural building blocks of cells and tissues. The information for producing the proteins required in a cell under a particular condition is contained in the deoxyribonucleic acid (DNA), and the complete DNA sequence of a being, the genome, is organized into chromosomes and genes. The central dogma of molecular biology (Crick, 1958, 1970) describes the information-flow in the replication of DNA and in the making of protein from DNA.

$$\circlearrowright \text{DNA} \rightarrow \text{mRNA} \rightarrow \text{Protein}$$

The production of protein from DNA is divided into two main steps. In step one, known as transcription, single stranded messenger ribonucleic acid (mRNA) is copied from the DNA, and in the second step, known as translation, proteins are produced based on information from the mRNA.

Gene expression analysis is the study of mRNA levels transcribed from DNA. In contrast to DNA which is more or less static over the life-time, and common to all cells of a being, mRNA levels varies over time and between cell types. It also varies within cells under different conditions. For example, the amount of mRNA transcribed from a gene in a healthy being can differ from the amount of mRNA transcribed from the same gene in the corresponding cell type of a sick being. If this is the case we say that the gene is differentially expressed between the two conditions healthy and sick.

Microarrays are technologies for measuring mRNA concentrations in tissue samples. Two of the most commonly used types of arrays are the Affymetrix GeneChip arrays (Lockhart et al., 1996), and spotted DNA arrays (Schena et al., 1995). This dissertation considers mainly the Affymetrix GeneChip type arrays and different analysis procedures and data processing steps required in the analysis of data from such arrays. Specifically, as indicated by the title, the subject of the first half is normalization procedures whereas the second half is focused to the problem of finding differentially expressed genes.

2 Background

2.1 Affymetrix technology

This section introduces the Affymetrix GeneChip technology. Following a brief description of DNA and RNA, the main characteristics of the arrays and procedures for preparing tissue samples for analysis are described.

2.1.1 DNA and RNA

Information about the development and functioning of a being is stored in DNA, for example the instructions needed to construct RNA and proteins. The complete set of DNA, the genome, is organized into chromosomes, and a segment of DNA coding for a protein is called a gene. The genome of a human being, containing a

total of 20,000-25,000 genes (International Human Genome Sequencing Consortium, 2004), is composed of 23 pairs of chromosomes where the last pair consists of the sex-determining X and Y chromosomes.

DNA and RNA are polymers composed of long chains of nucleotides and each nucleotide consists of a base, a sugar and a phosphate. In DNA, there are four bases: adenine (A), cytosine (C), guanine (G) and thymine (T). In RNA the thymine base is replaced by uracil (U). At room temperature DNA is double stranded, consisting of two complementary chains formed as a double helix by hydrogen bounds between complementary bases (Watson and Crick, 1953). The complementary bases A and T binds together while C is the complementary base of G and thus C binds together with G. In contrast to DNA, RNA is single stranded and there are several types of RNA including messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). In the context of this thesis mRNA is the most important type, carrying the information from DNA in the making of proteins with the help of tRNA. More information on DNA, RNA and proteins can be found in numerous text-books about molecular biology, for example Clote and Backofen (2000).

2.1.2 The arrays

The Affymetrix GeneChip arrays are one-color arrays and thus each array measures mRNA-abundance for one tissue sample only. Each gene is represented by a set of probes, the probe-set, with up to 40,000 probe-sets on a single array. A probe-set consists of 10-16 probe-pairs of one perfect match (PM) probe and one mismatch (MM) probe as depicted in Figure 1. The sequence of each probe is 25 bases long and the PM and MM probes have identical sequences of bases except for the middle probe which in the MM probe is set to the complementary base of that in the PM probe. The MM probes are thus designed to measure the background intensity for the corresponding PM probe.

The probes for each gene are selected from a reference sequence representing the gene, see Figure 1. A number of procedures are used to investigate probe candidates with respect to specificity and sensitivity, aiming for probes with a desired intensity range and concentration-dependence, and with a minimal risk of cross-hybridization. Also, the probe selection used for the Affymetrix arrays is biased towards the 3' end of the reference sequence.

Each probe corresponds to a square on the physical array, sized between 5×5 and 20×20 μm depending on array type, with all probes organized into a matrix with equal number of rows and columns. Older arrays had all probes of a probe-set situated together along the rows of the array with the PM probe placed directly on the row above the corresponding MM probe. Newer arrays have the probe-pairs of a probe-set scattered on the array, however the PM probe is still placed above the MM probe.

2.1.3 Sample preparation

When preparing the biopsies or tissue samples for which the mRNA levels are to be analyzed a number of laboratory steps are performed. Briefly the laboratory work done can be described as follows,

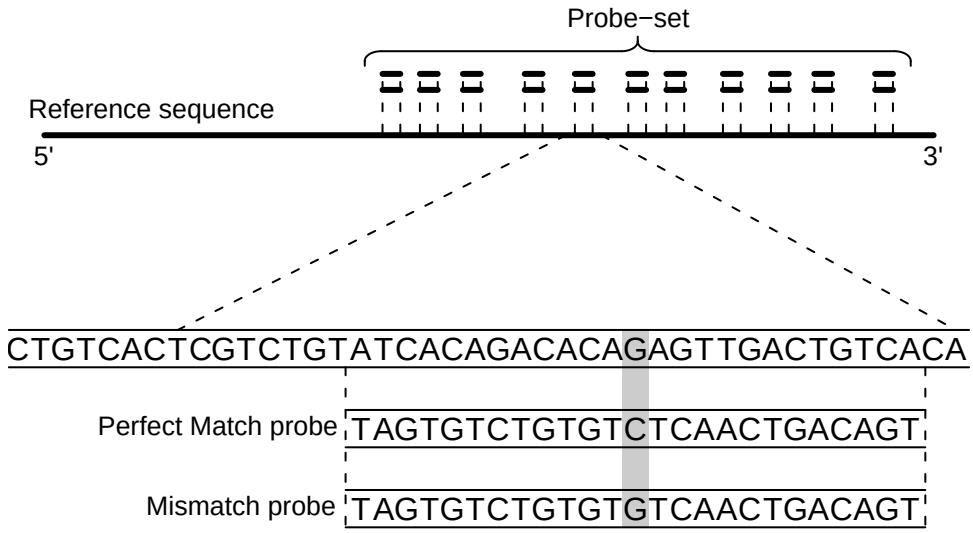

Figure 1: Affymetrix GeneChip arrays have multiple probe-pairs for each gene with probe-pairs consisting of one perfect match (PM) probe and one mismatch (MM) probe. The PM and MM probes are identical except for the middle base as highlighted in the figure.

$$\text{Total RNA} \rightarrow \text{ss cDNA} \rightarrow \text{ds cDNA} \rightarrow \text{biotin-labeled cRNA}$$

The first step is isolating of total ribonucleic acid (RNA). To be able to analyze a sample successfully the minimum amount required is around 1 µg. Total RNA is the starting material for obtaining labeled complementary RNA (cRNA). First, single stranded cDNA is synthesized by reverse transcription using poly-A primers present in the total RNA. Next, the single stranded cDNA is converted into double stranded cDNA. An in vitro transcription (IVT) reaction is then carried out in presence of biotinylated Uridine- and Cytidine-Triphosphate to produce biotin-labeled cRNA. The resulting cRNA is fragmented before hybridization onto the array.

After 16 hours of hybridization at 45 °C is completed non-hybridized cRNA is removed and a series of washing and staining steps are performed. The array is then scanned using a 16-bit scanner resulting in an image file which is the end-result of the laboratory work and the starting point for the analysis and data processing tools discussed in this thesis. More detailed information on all steps required before attaining the image file can be found in the Expression Analysis Technical Manual (Affymetrix, 2004).

2.2 Low-level analysis

Collectively, the processing of raw data from a set of Affymetrix GeneChip arrays into measures of gene expression is called low-level analysis, and the obtained measures

are called expression indexes. This typically involves image analysis, background estimation/correction and normalization. The first step is grid alignment of the image file so that the pixel-intensities of each probe can be identified. Control probes generating very high intensity values are placed along the borders, and in a check-pattern in each corner of the array, and are used when setting the grid. Each square of the grid represents a probe and a probe-intensity is calculated by first removing the border pixels and then computing the 75th percentile of the remaining intensity values. The probe intensities are then saved in a so called CEL-file. The grid alignment is described in detail in the United States Patent 6090555. Very little is written in the literature about alternatives to the procedure described above and the reminder of this thesis deals with analysis methods for raw or processed CEL-file data. In the following sections different procedures of computing expression indexes are described. Starting with two of the most frequently used methods, Affymetrix MAS 5.0 and Robust Multichip Average (RMA), and followed by three model based expression indexes.

2.2.1 Affymetrix MAS 5.0

The Affymetrix MAS 5.0 (MAS5 for brevity) expression index (Affymetrix, 2002, 2004) is the method for computing an expression index that is implemented in the Affymetrix GeneChip Operating Software. MAS5 is a single-array method, in that it can be computed separately for each array, and it consists of 4 steps: Global background correction, local background correction, summarization, and normalization.

In the first step a global background, defined as the 2% intensity quantile is subtracted from all probe intensities, and in the second step an Ideal Mismatch intensity (IM) is subtracted from all PM probes. For probes-pairs where the PM intensity is greater than the MM intensity, IM is equal to the MM intensity, whereas IM is equal to a fraction of the PM intensity otherwise. The IM-intensity was introduced to avoid the problem of many negative values with AvgDiff, the predecessor of MAS5. Special care is also taken in the global background correction to avoid negative values.

In the summarization-step, the MAS5 expression index is computed using the 1-step Tukey biweight M-estimator (Huber, 1981) on the background corrected PM intensities of each probe-set. Normalization is then performed by multiplying the expression index by a scaling factor to obtain a predefined overall intensity, defined by a trimmed mean, excluding the 2% highest and lowest values, of the expression indexes. See Appendix A.1 for further details.

2.2.2 RMA

Together with MAS5, Robust Multichip Average (RMA) (Irizarry et al., 2003a,b) is probably the most frequently used expression index. Although background correction is computed separately for each array in the first step, the remaining procedures used in RMA borrow information between arrays, and thus RMA is a multiple-array method. Also, RMA differs from MAS5 in that only PM intensities are used, and that normalization is done before summarization.

The background correction is based on a convolution model, assuming that the PM-intensity is a sum of a background intensity (Y) and real signal intensity (X).

With Y following a truncated normal-distribution and X an exponential-distribution, the background corrected PM intensity is defined as the conditional expectation of the real signal intensity given the total intensity,

$$\mathrm{E}\Big[X\Big|X+Y=\text{PM intensity}\Big] \ .$$

In the second step the background corrected PM intensities are normalized using Quantile-normalization, one of the three methods presented in Paper II. Summarization is the last step and is performed separately for each probe-set but across all arrays. A two-way ANOVA model is fitted with array- and probe-effects,

$$\log_2(PM_{ip}) = \theta_i + \psi_p + \text{error} \ ,$$

for $i = 1, \ldots,$ number of arrays and $p = 1, \ldots,$ number of probes. The median polish algorithm (Tukey, 1977) is used when fitting the model and the RMA expression indexes are then taken as the estimated array effects. See Appendix A.2 for further details.

GCRMA, a modification of the RMA method, was proposed by Wu et al. (2004) using a more sophisticated background correction where background is divided into optical noise and non-specific binding. See Appendix A.3 for further details.

2.2.3 Model-based expression indexes

The expression index method suggested in Li and Wong (2001a,b) is a model based expression index using the MM probe intensities for background correction. Here normalization is the first step and is performed on PM- and MM-probe intensities. A baseline array having median overall brightness (as measured by the median probe intensity) is selected, and used when normalizing all other arrays. Normalization is performed using a fitted curve in the scatter-plot of intensities with the baseline array on the y-axis, and the array to be normalized on the x-axis. A piecewise linear running median line is fitted to a subset of data-points, obtained by iteratively excluding data-points with large absolute rank differences in the two arrays.

Following the normalization, PM-MM differences are calculated and for a probe-set having P probe-pairs, with data from I arrays, the multiplicative model

$$PM_{ip} - MM_{ip} = \theta_i \psi_p + \epsilon_{ip}$$

is used. Here θ_i is the expression index of array i, ψ_p is a probe-sensitivity index of probe p, and ϵ_{ip} are independent normal distributed random variables with zero mean and variance σ^2. The model is fitted iteratively by holding the ψ_p's (or θ_i's) fixed and assumed known, and estimating the θ_i's (or ψ_p's), using the constraint $\sum_p \psi_p^2 = P$. Each time an outlier-detection is done, searching for array and probe outliers, as well as single probe outliers on a single array. When the model fitting has converged, the expression indexes reported are the estimated values of $\theta_1, \ldots, \theta_I$.

Another model-based method for computing expression indexes is the BGX method (Hein et al., 2005; Hein and Richardson, 2006). A fully Bayesian hierarchical model is suggested with the aim of integrating all processing steps of the raw CEL-file data into one framework. For a single array analysis, with PM_{gp} and MM_{gp} denoting the

PM and MM intensity, respectively, of the p'th probe-pair of the g'th probe-set, the hierarchical model used is

$$\begin{aligned} PM_{gp}|S_{gp}, H_{gp}, \tau^2 &\sim N(S_{gp} + H_{gp}, \tau^2) \ , \\ MM_{gp}|S_{gp}, H_{gp}, \tau^2, \phi &\sim N(\phi S_{gp} + H_{gp}, \tau^2) \ , \\ \log(S_{gp}+1)|\mu_g, \sigma_g^2 &\sim TN(\mu_g, \sigma_g^2) \ , \\ \log(H_{gp}+1)|\lambda, \eta^2 &\sim TN(\lambda, \eta^2) \ . \end{aligned}$$

Here S_{gp} and H_{gp} are referred to as true signal and nonspecific hybridization, respectively, and $\phi \in [0,1]$ is the fraction of the true signal that also hybridizes to the MM probes. $TN(\mu, \sigma^2)$ is the normal distribution truncated at zero, corresponding to the positive part of the un-truncated normal distribution with mean μ and variance σ^2, denoted by $N(\mu, \sigma^2)$. Prior distributions, roughly non-informative, are specified for the parameters τ^2, ϕ , μ_g, λ, and η^2, while an empirical Bayes approach is used to set the prior for σ_g^2, see Hein et al. (2005) for details. The median of the truncated normal distribution for $\log(S_{gp}+1)$,

$$\theta_g \quad = \quad \mu_g - \sigma_g \Phi^{-1}\Big(\frac{\mu_g}{2\sigma_g}\Big)$$

is defined as the expression index for probe-set g, and using MCMC simulations a complete posterior distribution for θ_g is obtained. The model is then generalized to the case of multiple arrays and for each probe-set g and array i the posterior distribution of θ_{ig} is computed by means of MCMC simulations.

A similar model is the multi-mgMOS model(Liu et al., 2005), a modification of the gMOS model suggested in Milo et al. (2003). With PM_{igp} and MM_{igp} denoting the PM and MM intensity, respectively, of the p'th probe-pair of the g'th probe-set on array i, multi-mgMOS is a hierarchical model given by

$$\begin{aligned} PM_{igp}|b_{gp} &\sim \Gamma(a_{ig} + \alpha_{ig}, b_{gp}) \ , \\ MM_{igp}|b_{gp} &\sim \Gamma(a_{ig} + \phi\alpha_{ig}, b_{gp}) \ , \\ b_{gp} &\sim \Gamma(c_g, d_g) \ , \end{aligned}$$

where $\Gamma(a,b)$ is the Γ-distribution with shape parameter a and scale parameter b. As for the BGX-model the parameter $\phi \in [0,1]$ describes how much of the true signal that also hybridizes to the MM probes, and an empirically derived prior for ϕ is used. It is also assumed that the true binding signal SS_{igp} is Γ-distributed with parameters α_{ig} and b_{gp}. Here true binding signal is the PM signal that would be obtained without any background signal present. The b_{gp}'s are integrated out, and the model is fitted using maximum likelihood conditional on ϕ and a maximum a posteriori estimate for ϕ. When the parameters have been estimated $E_{igp} = \mathrm{E}[\log(S_{igp})]$ is computed and the expression index is taken as the median E_{igp} of each array i and probe-set g.

2.3 Finding differentially expressed genes

In this section various methods for finding differentially expressed genes are described. Generally such analyses are performed at the probe-set level. That is, a method for

summarizing the probe-level data into expression indexes is first used, for example one of the methods described in Section 2.2, and an analysis is then performed on the expression indexes obtained. We start by describing general methods which can be applied at the probe-set level of Affymetrix type data using only the expression indexes as input. We then continue with methods applied at the probe-level and other alternative methods. Detailed descriptions for some of the methods are found in Appendix B.

2.3.1 Probe-set level analysis

There exist numerous methods for finding differentially expressed genes. In the early days of microarray analysis, the so called fold-change was frequently used to rank genes with respect to differential expression. Fold-change is usually defined as the ratio of two means, where each mean is calculated on a set of replicated arrays under the same condition. However, the means are often calculated on logged values, and the mean value is then anti logged before calculating the ratio. A list of potentially regulated genes was typically obtained by selecting all probe-sets with a fold-change value above 2 (or 3) together with probe-sets showing a fold-change below 1/2 (or 1/3). Thus, the variability of the replicated values was ignored with the obvious drawback that probe-sets with high fold-change may also be highly variable, and the high fold-change may have occurred by chance only. However, the direct application of Student's t-test calculated separately for each probe-set is not a better option. The number of replicates in many studies is small. As a consequence, variance estimators computed separately for each probe-set are highly variable and very small values can appear just by chance. Thus when used in the denominator of the t-statistic, a very large absolute value is obtained even if the estimated difference in the numerator is small.

The aim of moderated t-tests is to improve on the performance of the ordinary t-test. With empirical Bayes methods (Baldi and Long, 2001; Lönnstedt and Speed, 2002; Smyth, 2004; Kristiansson et al., 2005; Sartor et al., 2006; Sjögren et al., 2007) and the penalized t-test suggested by Opgen-Rhein and Strimmer (2007), the variance estimators calculated separately for all probe-sets are modified in order to produce more stable results. Together with the probe-set specific variance estimators a global estimator is computed. Based on the accuracy and the variability of the gene-specific variance estimators, weights are determined and used to calculate a weighted mean of the global and probe-set specific estimator, respectively. The weighted mean is then used in the denominator in place of the probe-set specific estimator. Other examples of moderated t-tests are the Significance Analysis of Microarrays (SAM) method (Tusher et al., 2001) and the method suggested by Efron et al. (2001), where a constant is added to the probe-set specific sample standard deviation.

The weighted moderated t-test derived in Kristiansson et al. (2005, 2006) and further developed by Sjögren et al. (2007) differs from the other moderated t-test in that weighted means are used in the numerator. A global covariance matrix is used to account for differing variances between arrays as well as array-to-array correlations. It is motivated by the fact that quality often varies between arrays and samples. Also, sources of variations are introduced at several steps in microarray-data, and when sources of variations are shared between arrays, the measurements are expected

to be correlated.

Another aspect of microarray-data is that variability often varies with intensity level. This is ignored by the majority of moderated t-test, but it is utilized in the moderated t-test suggested by Sartor et al. (2006). They build on the moderated t-test suggested by (Lönnstedt and Speed, 2002; Smyth, 2004) with the addition of first fitting a loess curve in the scatter-plot of logged variance estimators against mean intensity. The fitted curve is used when estimating the model parameters. Another example where the variance is modeled as a function of intensity-level is the method suggested by Eaves et al. (2002). They use a weighted average of the probe-set specific variance estimator and a pooled estimate based on the 500 probe-sets with most similar mean expression level. A similar approach is the local-pooled-error method (LPE) suggested by Jain et al. (2003), using a variance function fitted to estimated variances and mean intensities. Comander et al. (2004) pool genes with respect to minimum intensity rather than mean intensity, and Hu and Wright (2007) use a hierarchical model with a linear relationship between variance and intensity-level.

Another way of attacking the problem of dependency between variability and intensity level is to apply a variance stabilizing transformation. The generalized-log family (glog) was introduced by Munson (2001); Huber et al. (2002); Durbin et al. (2002) and is further used by Durbin and Rocke (2003); Geller et al. (2003). Other transformations are the started logarithm transformation (Tukey, 1977) and the log-linear hybrid transformation (Holder et al., 2001). Rocke and Durbin (2003) performed a comparison of the three transformations and concluded that the generalized-log family is "probably the best choice when it is convenient to use it". Also, the glog transformation implicitly defines a background correction, and can thus be used when calculating an expression index (Huber et al., 2003; Zhou and Rocke, 2005).

2.3.2 Probe level analysis

The logit-t method suggested by Lemon et al. (2003) uses PM-probe intensities for finding differentially expressed genes from two groups of replicated arrays. A logit-like transformation adjusting for background intensity followed by a Z-transformation, i.e. shifting and scaling to zero mean and unit variance, is applied to PM-probe intensities. The transformed intensities are then analyzed using Student's t-test to compare the two groups, and the logit-t statistic for a probe-set is then defined as the median t-value among all PM probes in the probe-set.

In the Probability of Positive Log-ratio method (PPLR) described by Liu et al. (2006), a probe-level error measurement is included in the analysis together with an associated expression index. They propose a Bayesian hierarchical model where variances of expression indexes are modeled as a sum of the probe-level error measurement and a gene specific variance component common to all arrays. The normal- and Γ^{-1}-distribution are used as prior distributions for condition specific mean values and variances, respectively, and three different methods for fitting the model and to obtain the posterior distributions of condition specific mean values are described. The method is demonstrated using the expression indexes and error measurements from the multi-mgMOS model (Liu et al., 2005).

With the BGX method (Hein et al., 2005) a complete posterior distribution for

each probe-set and array is obtained, and thus could also be used as input to the PPLR method. However, the single array model for BGX can alternatively be generalized to the case of replicated arrays under different conditions, instead of multiple arrays as described in Section 2.2.3. Thus, posterior distribution of differential expression can be calculated directly from the MCMC simulations.

3 Summary of papers

Paper I: *Contrast normalization of oligonucleotide arrays*

One crucial step in the analysis of microarray data is normalization. Since the overall brightness of the scanned images can differ substantially between arrays, normalization is usually required to allow direct array-to-array comparisons. A very simple way to normalize a set of arrays is to compute a multiplication factor forcing equal overall intensity, for example measured by the mean or median intensity. However, quite often there exist non-linear relationships between the intensities of arrays, and thus more flexible solutions are required.

In this paper a method called Contrast Normalization is proposed. The method can be seen a generalization of the intensity-dependent normalization procedure proposed by Yang et al. (2002), hereafter called loess normalization. The loess normalization is designed for two-color arrays where a direct normalization of the ratios of red and green signal is natural. For Affymetrix type data it is more natural to normalize intensities rather than ratios, and in particular having smooth functions that normalizes the intensities of each array.

A direct application of the loess normalization on a pair of Affymetrix arrays is to fit a smooth loess-curve in the scatter-plot with $X_1 - X_2$ on the y-axis and $\frac{1}{2}(X_1 + X_2)$ on the x-axis, which can be seen as a change of basis from the original values X_1 and X_2. Here X_1 and X_2 represents the logged intensity on array 1 and 2, respectively. Thus, the mean logged intensity across arrays is used to model the difference of logged intensities. For the case of more than 2 arrays, this principle is generalized in the proposed method by using a set of orthogonal contrasts (of logged intensities) which all are modeled by the mean logged intensity across arrays.

However, a direct application of the loess normalization results in a non-continuous normalization in that intensities being equal on one array prior to normalization may not be equal after normalization. To solve this problem the fitted curves are used to define a mapping of data points, to a set of ideal data points that would not require any normalization. Thus, the mapping is defined in the alternative basis of mean logged-intensity and the set of contrasts. By viewing this mapping in the original basis of array intensities smooth functions are obtained that normalizes the intensities of each array.

Paper II: *A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance*

In this paper three methods for normalizing probe-level data are presented and evaluated using two publicly available data sets, a dilution data set and a spike-in dataset.

The presented methods, Cyclic-loess, Contrast normalization, and Quantile normalization are compared to the method implemented in the Affymetrix software computing a scaling factor for each array using a trimmed mean of intensities, and to the non-linear method proposed by Li and Wong (2001a,b) using a baseline array to which all other arrays are normalized.

Cyclic-loess is a generalization of the intensity-dependent normalization procedure proposed by Yang et al. (2002), hereafter called loess normalization. With only two arrays Cyclic-loess is a direct application of the loess normalization. With more than two arrays Cyclic-loess iteratively performs loess normalization on all pairs of arrays, until all pairs of arrays are sufficiently normalized. Contrast normalization is another generalization of the loess normalization and is further described in Paper I.

The goal of the Quantile normalization is to produce identical empirical distributions of intensities on all arrays analyzed. Let G_i denote the empirical distribution of intensities on the i'th array, and let F denote the empirical distribution of the averaged sample quantiles. Identical empirical distributions of intensities are then achieved by normalizing the intensities on the i'th array by the composite function $F^{-1} \circ G_i$.

For each group of 5 arrays with identical dilution concentrations (or mixture proportions) variance of probe-set summaries was computed for each probe-set and normalization. All three of the presented methods, together with the non-linear method, reduced the variability of probe-set summaries to a greater degree than using a global scaling factor. Also, pairwise comparisons were performed by computing the absolute distance from the x-axis to a smooth fitted curve in M versus A plots. When averaging this distance across all pairwise comparisons the quantile method gave the smallest distance between arrays and the distance was fairly constant across intensity levels.

The methods were also compared with respect to bias using the spike-in data set. For each of the 11 spiked probe-sets a linear regression model was fitted to RMA expression indexes and with $\log_2$ spike-in concentration as the regressor. The ideal result would be to have slope estimates close to 1. Although with slope estimates consistently below 1, all three presented methods performed comparably well. The non-linear method performed poorer while using a global scaling factor resulted in slightly higher but also more variable slope estimates.

The concept as well as the algorithm of the Quantile normalization method is very simple, and in terms of speed Quantile normalization is superior over the other methods based on curve fitting. In summary, with favorably performance in terms of speed, variance, and bias, it is recommended that Quantile normalization should be used in preference to the other methods.

Paper III: *Improved Covariance Matrix Estimators for Weighted Analysis of Microarray Data*

For microarray data, many sources of variations are introduced that may affect the final measurements obtained. For example, variability is introduced during the laboratory work and when obtaining the biological samples, e.g. taking biopsies. This means that data quality within an experiment often varies between arrays. Moreover, when sources of variations are shared between arrays we can expect correlations

between the measurements. To accommodate these differences in data quality, i.e. differences in variances and the possibility of correlations between arrays, the empirical Bayes WAME model was proposed by Kristiansson et al. (2005), and further developed in Kristiansson et al. (2006) and Sjögren et al. (2007).

WAME makes use of a global covariance matrix which is scaled between genes by a gene-specific parameter assumed to be inverse-gamma distributed. The global covariance matrix is estimated under the temporary assumption of no regulated genes. The assumption is then relaxed and the remaining model parameters are estimated. However, due to the temporary assumption, the covariance matrix estimator is biased when regulated genes exist. Also, the computational procedure used is very computationally intensive resulting in long computer run times. In this paper two new methods for estimating the covariance matrix are proposed with the aim of reducing or eliminating these two drawbacks.

The first method, hereafter called method I, is based on the same temporary assumption of no regulated genes as used in the method proposed by Kristiansson et al. (2005). Under this assumption the WAME model describes a multivariate t-distribution with zero mean, and unknown covariance matrix (Σ) and degrees of freedom (m). Method I is then obtained as a direct application of the EM algorithm (Dempster et al., 1977). Method I is compared with the procedure proposed by Kristiansson et al. (2005). With respect to precision and bias the methods performed equally well, but method I was superior in terms of computer time.

In the second method, hereafter called method II, the WAME model is extended with a prior distribution for μ_g, the mean intensity profile across arrays. However, a linear transformation derived from the design and the one-row contrast matrices is first applied to data. With x_g denoting the q sized vector of *transformed* log-intensities the following model is used. For $g = 1, \ldots, G$ let

$$\begin{aligned} x_g|c_g &\sim \mathrm{N}_q(\mu_g, c_g\Sigma) \ , \\ c_g &\sim \Gamma^{-1}(\tfrac{1}{2}m, \tfrac{1}{2}m\nu) \ , \\ \mu_g &= (0, \ldots, 0, \delta_g)^T \ , \\ \delta_g &\sim \begin{cases} \equiv 0 & \text{with prob. } \psi_0 \ , \\ F(\beta) & \text{with prob. } 1-\psi_0 \ . \end{cases} \end{aligned}$$

Here δ_g is the linear combination of the design parameters that is to be estimated (usually logged fold change between two conditions), and $F(\beta)$ is a continuous distribution, parameterized by β and with density function $f(x|\beta)$. The difference from the original WAME-model is the structure of μ_g and the distributional assumption on δ_g.

Method II is composed of two steps. First, the last dimension of x is dropped and estimates of the hyperparameters m and ν together with an estimate of Σ_A (the sub-matrix of the first $q-1$ rows and columns of Σ) are computed using method I. The hyperparameters m and ν are then treated as known and equal to the estimated values. In step 2 estimates of Σ and β are computed by replacing the continuous distribution $F(\beta)$ by a discrete version $\tilde{F}(\beta)$. As for method I the EM algorithm is used, treating the c_g's and δ_g's as missing data.

Methods I and II are then compared on simulated data with and without regulated genes, and on real data with many regulated genes. On simulated data with regulated genes, the bias observed for the estimator of Σ when using method I is greatly reduced by method II. Moreover, in the case of no regulated genes the methods showed comparable variability and no bias. However, when used on real data with many regulated genes method II appears to be biased although to less extent than method I, and thus the problem of bias with the procedure proposed by Kristiansson et al. (2005) is only partly resolved. On the other hand, the proposed methods make it possible to apply weighted analysis of microarray data using the WAME-model to large data sets with reasonable computer run times.

Paper IV: *Empirical Bayes models for multiple probe type arrays at the probe level*

The Affymetrix type arrays differ from other arrays in that each gene is represented by multiple probes. The standard way of dealing with the multiple-probes is to derive a summary measurement, an expression index, for each probe-set (gene) and array (sample), for example, using one of the methods described in Section 2.2. Generally, such methods include background correction, normalization, and summarization. Differential gene expression analysis is then performed using the expression indexes obtained.

In this paper a procedure for finding differentially expressed genes excluding the step of summarization is proposed, and thus performing inference at the level of background corrected and normalized perfect match probe data. Generally, data at this level show a clear relationship between variability and intensity level even on log-scale. Also, such relationships often exist for data at the level of expression indexes.

The WAME-model proposed in Kristiansson et al. (2005, 2006) is extended to incorporate the variability to intensity-level dependency by modeling the scale parameter of the prior distribution for probe specific variances as a smooth function of intensity-level. A cubic spline is used to parameterize the function and the model is fitted using maximum likelihood by means of the EM-algorithm.

When applying the extended WAME-model at the probe-level, weighted moderated t-tests are computed for each PM probe. The t-statistics obtained for each probe-set are then summarized into a score by the median t-statistic. This score is the value used for ranking probe-sets with respect to differential expression in the proposed method Probe level Locally moderated Weighted median-t method (PLW).

The second proposed method, Locally Moderated Weighted-t (LMW), is a more general method intended for single probe type arrays or summary measures of multiple probe type arrays. LMW uses the same model and estimation-procedure as PLW but excludes the final median summarization since only one t-statistic is obtained for each probe-set. The proposed methods are compared with existing methods on five publicly available spike-in data sets.

It is shown that both methods perform very well compared to existing methods. The proposed method PLW has the most accurate ranking of regulated genes in four out of the five examined data sets. With LMW consistently performing better than all (globally) moderated t-tests it is also shown that introducing an intensity-level

dependent scale parameter for the prior distribution of the gene-specific variances improves the performance of the moderated t-test. Also, with LMW having more accurate ranking than the locally moderated t-test IBMT on all 5 data sets, it seems that weighted analysis is as important as local moderation. But most strikingly, with the PLW method performing overall better than all compared methods it appears that the probe-level inference approach is preferable over the standard approach using gene expression indexes for inference.

4 Complementary studies

4.1 PLW and LMW combined with GCRMA and MAS5

The PLW method suggested in Paper IV uses probe-level data in the form of background corrected and normalized PM intensities. In Paper IV the default background correction and normalization of the RMA method is used, and PLW is compared with LMW and other general methods for finding differentially expressed genes using RMA expression indexes as input.

This section presents results from two additional comparisons of PLW and LMW to other methods for finding differentially expressed genes. The comparisons are done using the same 5 spike-in data-sets as was used in Paper IV, and the comparisons are summarized by ROC-curve-AUC up to 100 false positives. The methods compared with PLW and LMW include

- observed fold change (FC),
- ordinary t-test,
- the SAM method (Tusher et al., 2001) in the R-package `samr`,
- Efron's penalized t-test (Efron et al., 2001) in the R-package `st`,
- the moderated t-test LIMMA in the R-package `limma` (Smyth, 2004),
- the weighted moderated t-test WAME (Kristiansson et al., 2005, 2006) in the R-package `WAME.EM` available at `www.math.chalmers.se/~astrandm/wame_em/`,
- the moderated t-test IBMT suggested by Sartor et al. (2006) using the R-code available at `http://eh3.uc.edu/r/ibmtR.R`,
- the Shrink-t method (Opgen-Rhein and Strimmer, 2007) in the R-package `st`,
- the Local-pooled-error test (Jain et al., 2003) in the R-package `LPE`.

If not otherwise specified, the R-packages are available at `www.bioconductor.org/` or at `www.r-project.org/`, and PLW and LMW are both implemented in the R-package `plw` available at `www.math.chalmers.se/~astrandm/plw/`. The methods listed above are further described in Appendix B.

In the first comparison the model based background correction named GCRMA suggested in Wu et al. (2004), is used instead of the default background correction of RMA. Functions implemented in the R-package `gcrma` was used with the `fast`

option set to `FALSE`, and thus the empirical Bayes approach of GCRMA was used to calculate background corrected intensities, see Appendix A.3 for further details. The probe-level method PLW is compared with 10 probe-set level methods, including LMW, applied to expression indexes obtained using the GCRMA method. The result is summarized in the upper part of Table 1 and overall the results are very similar to the results presented in Paper IV. The IBMT method performs slightly better when applied to GCRMA expression indexes whereas WAME performs slightly worse. The ordering of the top-three methods is unchanged with PLW ranked as number one, although the advantage of PLW over the other methods is not as pronounced as in Paper IV.

In the second comparison logged MAS5 expression indexes are used and LMW is compared with 10 other probe-set level methods. The result is summarized in the lower part of Table 1. Since MAS5 expression indexes show a very clear dependency between variability and intensity level, and since the variability decreases with intensity it comes as no surprise that all three methods taking this dependency into account consistently performs better than all other methods. The LMW method has the most accurate ranking of genes in 4 out of the 5 data-sets, and performs better than the IBMT method on all 5 data-sets. Since the main difference between LMW and IBMT is that LMW performs a weighted analysis based on the WAME model proposed by (Kristiansson et al., 2005), and since WAME overall performs better than LIMMA, weighted analysis should be used in preference to analysis using un-weighted analysis. However, the effect of using local moderation on MAS5 expression indexes is greater than the effect of using a weighted analysis, and thus local moderation appears to be even more important.

4.2 Using LMW on two-color array data

This section presents a case study where LMW is used on data from two-color spotted cDNA microarrays. When using two-color microarrays, mRNA from two sources are hybridized on each array. The mRNA of one source is labeled with Cy5 and the mRNA of the other source is labeled with Cy3. When scanning the arrays the signal is divided into red and green signal, corresponding to the mRNA labeled with Cy5 and Cy3, respectively. Analysis is then performed on normalized logged ratios of the red signal over the green signal, measuring the relative mRNA abundance of the two sources.

The data-set studied here is from an experiment comparing 8 ApoAI knockout mice with 8 normal mice (Callow et al., 2000) using a set of $n = 16$ arrays, and is available at `http://bioinf.wehi.edu.au/limmaGUI/DataSets.html`. Liver tissue was obtained from each mouse and the extracted mRNA was labeled with Cy5. The mRNA was then hybridized together with a Cy3 labeled reference mRNA mixture, obtained by pooling mRNA from the 8 normal mice. Data was pre-processed as described in (Callow et al., 2000) and the analysis presented here is based on the 6068 genes (out of 6226) having no missing values.

Let x_{ig} denote the logged R/G-ratio for gene g on array i. Assume that array 1-8 is the control group with mRNA from normal mice and put $x_g = (x_{1g}, \ldots, x_{ng})^T$.

	Method	Affymetrix U95	Affymetrix 133A	Golden Spike	Gene Logic Tonsil	Gene Logic AML
GCRMA	PLW	97(1)	92(8)	38(2)	87(1)	87(1)
	LMW	95(2)	93(1)	21(10)	84(2)	79(4)
	LPE	95(4)	91(10)	40(1)	82(4)	86(2)
	IBMT	95(3)	93(2)	27(4)	81(7)	76(6)
	Efron-t	94(5)	93(4)	26(5)	82(5)	79(5)
	WAME	94(8)	93(3)	26(6)	83(3)	75(8)
	LIMMA	94(7)	93(5)	27(3)	80(8)	73(9)
	FC	93(10)	93(7)	25(7)	81(6)	86(3)
	SAM	94(6)	93(6)	24(9)	79(9)	76(7)
	Shrink-t	94(9)	92(9)	25(8)	78(10)	70(10)
	t-test	86(11)	84(11)	15(11)	64(11)	53(11)
MAS5	LMW	89(1)	87(1)	54(1)	79(1)	70(2)
	IBMT	87(2)	87(2)	52(2)	77(3)	69(3)
	LPE	84(3)	84(3)	49(3)	78(2)	79(1)
	WAME	71(6)	81(5)	15(6)	69(4)	54(8)
	LIMMA	71(7)	81(6)	17(5)	67(6)	54(6)
	SAM	74(4)	81(4)	1(8)	67(5)	54(9)
	Shrink-t	71(8)	80(7)	10(7)	67(7)	54(7)
	t-test	73(5)	76(8)	38(4)	60(9)	47(10)
	Efron-t	65(9)	72(9)	1(9)	66(8)	57(4)
	FC	56(10)	61(10)	0(10)	58(10)	55(5)
	# of genes	12626	22029	11475	12626	12626
	# of spikes	16	42	1331	11	11
	# of groups	20	14	2	12	10

Table 1: Area under ROC curves up to 100 false positives rounded to nearest integer value with an optimum of 100. Numbers within parenthesis are within data set ranks for the methods compared. Methods are ordered with respect to mean rank across data sets. The upper part of the table show result when using GCRMA background correction (Wu et al., 2004) together with Quantile normalization and (for all methods but PLW) the default summarization method of the RMA expression index. The lower part of the table shows result when using MAS5 expression indexes.

Analysis using the R-package `limma` is performed using a design matrix D,

$$D^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

to specify the expected profile $\mu_g = \mathrm{E}[x_g]$ for gene g across arrays as $\mu_g = D\gamma_g$. Here γ_g is a gene-specific parameter vector of length 2 where the second element represents the difference between the control group and the group of ApoAI knockout mice. Each vector x_g is then modeled as

$$\begin{aligned} x_g|c_g &\sim \mathrm{N}_n(\mu_g, c_g I) \ , \\ c_g &\sim \Gamma^{-1}(\tfrac{1}{2}m, \tfrac{1}{2}m\nu) \ , \end{aligned} \tag{1}$$

where I is the $n \times n$ identity matrix, $\mathrm{N}_n(\mu, \Sigma)$ denotes an n-dimensional normal distribution with mean μ and covariance matrix Σ, and $\Gamma^{-1}(\alpha, \beta)$ is the inverse-gamma distribution with shape-parameter α and scale-parameter β. For the ApoAI data-set the estimated value of m is 3.85 and the estimated value of ν is 0.0499.

Panel A of Figure 2 shows a histogram of logged ratios of sample error variances for all genes (s_g^2) divided by the estimated value of ν, together with the marginal distribution of $\log(s_g^2/\nu)$ according to model (1). As seen in the Figure, the fitted marginal distribution deviates from the empirical distribution. Panel B of Figure 2 displays a scatter-plot of $\log(s_g^2)$ against average of $\frac{1}{2}(\log(R)+\log(G))$, together with a smooth fitted curve, demonstrating that genes with high intensity level are less variable compared to low intensity genes, thus contributing to the heavy lower tail of the empirical distribution of $\log(s_g^2/\nu)$ in panel A.

There are two differences between model (1) and the model used in LMW, presented in Paper IV and described in Appendix B.8. A covariance matrix Σ is used in place of the identity matrix I and ν is modeled as a smooth function of $\bar{\mu}_g$, the average of expected intensities. Applying the LMW method to logged R/G-ratios from a set of two-color microarrays requires a slight modification of the model presented in Paper IV. Let y_{ig} represent the average log-signal of the red and green signal, respectively, for gene g on array i, thus

$$y_{ig} = \frac{\log(R) + \log(G)}{2} = \log(\sqrt{R \cdot G}) \quad \text{whereas} \quad x_{ig} = \log(R/G) \ .$$

For notational simplicity the subscripts i and g are suppressed for R and G, denoting the red and green intensity, respectively. With $y_g = (y_{1g}, \ldots, y_{ng})^T$, let $\theta_g = \mathrm{E}[y_g]$ and $\bar{\theta}_g$ denote the average of the vector θ_g. For two-color microarrays the scale-parameter ν should be modeled as a function of $\bar{\theta}_g$ instead of $\bar{\mu}_g$, and thus each vector x_g is modeled as

$$\begin{aligned} x_g|c_g &\sim \mathrm{N}_n(\mu_g, c_g \Sigma) \ , \\ c_g &\sim \Gamma^{-1}(\tfrac{1}{2}m, \tfrac{1}{2}m\nu(\bar{\theta}_g)) \ , \end{aligned} \tag{2}$$

where Σ is an $n \times n$ covariance matrix and $\nu(\,\cdot\,)$ is a smooth function. With $\bar{\theta}_g$ replaced by $\bar{y}_g$ model (2) is fitted as described in Paper IV.

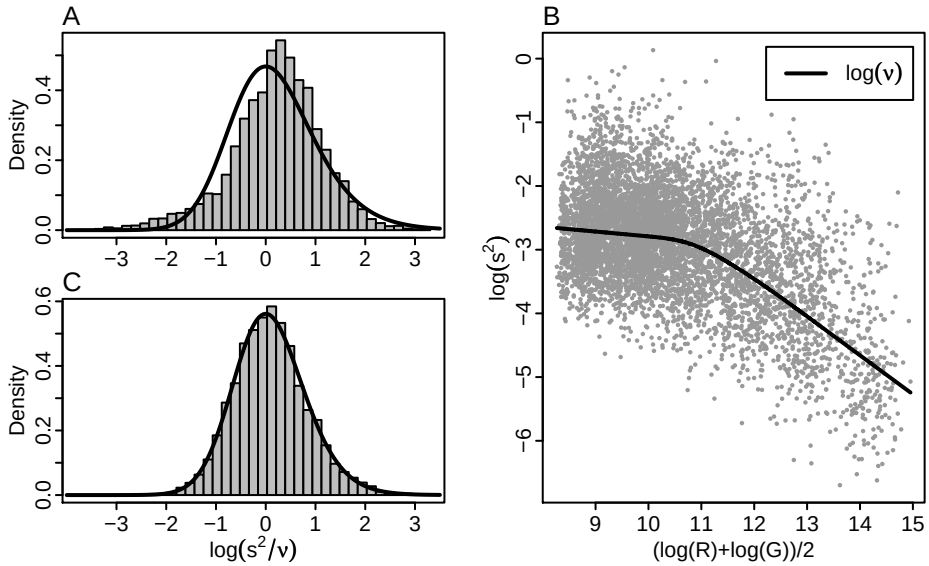

Figure 2: A) Histogram of $\log(s_g^2/\nu)$ for the ApoAI data set. s_g^2 is the sample variance with 14 degrees of freedom for gene g, ν is the estimated scale-parameter, and the black curve is the fitted marginal density of $\log(s_g^2/\nu)$, all three calculated using model (1). B) Scatter-plot of $\log(s_g^2)$ versus average red and green log-intensity level. s_g^2 is here calculated according to model (2) as an adjusted weighted residual sum of squares (see end of Appendix B.8). The black curve is the log of the intensity dependent scale parameter of model (2). C) As A), but s_g^2 calculated as in B), ν equal to the black curve in B), and the black curve is the fitted marginal density of $\log(s_g^2/\nu(\bar{\theta}_g))$.

The log of the estimated function $\nu(\,\cdot\,)$ is displayed in panel B of Figure 2, and panel C shows histogram of $\log(s_g^2/\nu(\bar{y}_g))$ together with the marginal distribution according to model (2). Comparing panel A with panel C, shows that model (2), with an intensity dependent scale-parameter, fits data much better than model (1) having a global scale-parameter. The estimator of m for model (2) is equal to 6.03. Thus, when computing the moderated t-statistic the weights of the prior and the sample error variance are 30% and 70%, respectively. For model (1) where the estimator of m is 3.85, the weight for the prior is 21%.

It should be mentioned that a direct modeling of green and red logged-signals using the model suggested in Paper IV, with a design matrix specifying a paired comparison, is equivalent to the modification of the LMW-model described above.

Appendixes

A Expression indexes

This appendix is a detailed description of three methods for computing expression indexes.

A.1 Affymetrix MAS 5.0

The Affymetrix MAS 5.0 (MAS5 for brevity) expression index (Affymetrix, 2002, 2004) is the method for computing an expression index that is implemented in the Affymetrix GeneChip Operating Software. Background correction is here done in two steps. First all probe intensities are corrected for global background intensity. The array is divided into K squares of equal size and within each square k the 2% sample quantile (b_k) is computed. Let $d_k(i,j)$ be the distance between the coordinate (i,j) and the center of the k'th square. The probe intensity at (i,j) is then corrected by subtracting

$$B(i,j) \quad = \quad \frac{\sum_{k=1}^{K} w_k(i,j)\ b_k}{\sum_{k=1}^{K} w_k(i,j)} \quad \text{where} \quad w_k(i,j) = \frac{1}{d_k(i,j) + smooth}$$

where $smooth > 0$ is a constant. With s_k equal to the standard deviation of intensity values of square k below b_k, $S(i,j)$ is a weighted average of $s_1, \ldots, s_K$ computed the same way as $B(i,j)$. To avoid negative and too small values when subtracting $B(i,j)$, the minimum corrected intensity at (i,j) is $NoiseFrac \cdot S(i,j)$, for a certain constant $NoiseFrac$.

Secondly, an Ideal Mismatch intensity (IM) acting as local background intensity is computed. For a probe-set having P probe-pairs, let PM_p and MM_p denote the PM and MM intensity, respectively, of the p'th probe-pair. Given constants δ and λ, IM_p is calculated according to,

$$IM_p = \begin{cases} MM_p & \text{if} \quad PM_p > MM_p\ , \\ PM_p \cdot 2^{-SB} & \text{if} \quad PM_p \leq MM_p \quad \text{and} \quad SB > \delta\ , \\ PM_p \cdot 2^{-\delta\lambda/(\lambda+\delta-SB)} & \text{if} \quad PM_p \leq MM_p \quad \text{and} \quad SB \leq \delta\ . \end{cases}$$

Here SB is a robust location estimator calculated on logged ratios of PM_p over MM_p,

$$SB = \mathrm{TB}\Big(\log_2(MM_1/PM_1), \ldots, \log_2(MM_P/PM_P) \Big)\ ,$$

where TB denotes the 1-step Tukey biweight M-estimator (Huber, 1981).

The same type of M-estimator is applied to logged $PM_p - IM_p$ differences and in detail the MAS5 expression index is defined as

$$2^{\mathrm{TB}(V_1, \ldots, V_P)} \quad \text{where} \quad V_p = \max\big(\log_2(PM_p - IM_p), -20 \big)\ .$$

Normalization is the last step and is performed by computing a scaling factor (sf) for the array which all expression indexes are multiplied by. A trimmed mean,

excluding the 2% highest and lowest values, of the expression indexes is calculated, and given a target intensity (Sc) the scaling factor sf is Sc divided by the trimmed mean value. Default values for the parameters used in the MAS5 algorithm are K=16, $smooth$=100, $NoiseFrac$=0.5, δ=0.03, λ=10, and Sc=500.

A.2 RMA

The Robust Multichip Average method (RMA) (Irizarry et al., 2003a,b) uses a convolution model when correcting for background. The intensities from all PM probes of an array are modeled as a sum of a background intensity (Y) and real signal intensity (X). Formally,

$$\begin{aligned} S &= X + Y \ , \\ X &\sim \exp(\alpha) \ , \\ Y &\sim TN(\mu, \sigma^2) \ , \end{aligned}$$

where S is the PM intensity, and $\exp(\alpha)$ is the exponential-distribution with mean α^{-1}. $TN(\mu, \sigma^2)$ is the normal distribution truncated at zero, corresponding to the positive part of the un-truncated normal distribution with mean μ and variance σ^2. A step-wise procedure is used when estimating the parameters. First, μ is estimated by the mode of a kernel density fitted to the observed PM intensities. Then the intensities below the estimated μ are used to estimate σ^2, and the intensities above the estimated μ are used to estimate α. The parameters are then set equal to the estimated values and assumed known. Background corrected PM intensities are computed according to

$$\mathrm{E}[X|S=s] = s - \mu - \alpha\sigma^2 + \sigma \frac{\phi\left(\frac{s-\mu-\alpha\sigma^2}{\sigma}\right) - \phi\left(\frac{\mu+\alpha\sigma^2}{\sigma}\right)}{\Phi\left(\frac{s-\mu-\alpha\sigma^2}{\sigma}\right) + \Phi\left(\frac{\mu+\alpha\sigma^2}{\sigma}\right) - 1} \ ,$$

where s is the observed PM intensity, ϕ and Φ are the density and cumulative distribution function of the standard normal distribution, respectively.

Normalization in the RMA method is performed on background corrected PM intensities using the quantile normalization method described in paper II. The method produces identical empirical distributions of intensities on all arrays analyzed. With G_i denoting the empirical distribution of intensities on the i'th array, and with F denoting the empirical distribution of the averaged sample quantiles, the intensities on the i'th array is normalized by the composite function $F^{-1} \circ G_i$.

In the final step of the RMA method expression indexes are computed. Separately for each probe-set, the normalized PM intensities are modeled using a two-way ANOVA. Formally, for $i = 1, \ldots, I$, $p = 1, \ldots, P$

$$\log_2(PM_{ip}) = \theta_i + \psi_p + \epsilon_{ip} \ ,$$

where I is the number of arrays, P the number probes, θ_i and ψ_p are the array and probe effect, respectively, and ϵ_{ip} are iid random variables with zero mean. Using the constraint $\psi_1 + \cdots + \psi_P = 0$ the model is fitted using the median polish algorithm (Tukey, 1977). The expression indexes are then taken as the estimated array effects $\hat{\theta}_1, \ldots, \hat{\theta}_I$, and thus the RMA expression index is on $\log_2$-scale.

A.3 GCRMA

The GCRMA method proposed by Wu et al. (2004) and the RMA method are based on the same procedures for normalization and summarization, but use different background corrections. In GCRMA background signal is divided into optical noise and non-specific binding. Using an affinity model, the non-specific binding (α) is modeled as a sum of position-dependent base effects,

$$\alpha = \sum_{k=1}^{25} \sum_{j \in \{A,T,G,C\}} \mu_{jk} I(b_k = j) \ , \tag{3}$$

where $k = 1, \ldots, 25$ indicates the position along the probe, j denotes the base letter, b_k represents the base at position k of the probe, $I(A)$ is the indicator function for the event A, and μ_{jk} represents the contribution to affinity of base j in position k. For fixed j, $\mu_{j1}, \ldots, \mu_{j25}$ is modeled using a spline with 5 degrees of freedom. A similar model was used by Naef and Magnasco (2003) using a polynomial of degree 3 for μ_{jk} instead of a spline. Model (3) is fitted to log intensities obtained by hybridizing yeast control RNA on to an array measuring human genes, thus the intensities obtained are likely to reflect optical noise as well as non-specific binding but no real signal. The parameters μ_{jk} are then treated as known and held fixed at the estimated values in the analysis of other data sets.

For a general data set, with PM and MM denoting the intensity of the PM and MM probe of a probe-pair, respectively, background correction in GCRMA is based on the assumption that

$$\begin{aligned} PM &= O + N_{PM} + S \ , \\ MM &= O + N_{MM} \ , \end{aligned}$$

where O is an array specific constant representing optical noise, N represents non-specific binding, and S is the true signal proportional to mRNA abundance. N_{PM} and N_{MM} are assumed to follow a bivariate normal distribution with means equal to μ_{pm} and μ_{mm}, respectively, variances equal to σ^2 and correlation equal to ρ. The optical noise O is estimated by 1 + the minimum intensity of each array, and the value of ρ was estimated using the same data used for estimating the parameters μ_{jk} of model (3) and is held fixed at the estimated value equal to 0.7. The mean values μ_{pm} and μ_{mm} are determined by a smooth function h as $h(\alpha_{PM})$ and $h(\alpha_{MM})$, respectively. Here α_{PM} and α_{MM} are calculated using model (3) and the base-sequences of the PM and MM probe, respectively, and h is a loess curve fitted to $\log(MM - \hat{O})$ with α_{MM} as regressor. The median absolute value of negative residuals for the fitted curve h is used to estimate σ^2.

With the array-specific parameters O, h, σ^2, and ρ estimated as described above, two methods for computing a background corrected PM intensity are proposed by Wu et al. (2004). The first method computes a maximum likelihood estimator of S as the background adjusted PM intensity,

$$\hat{S} \ = \ \min\left(PM - \hat{O} - \exp\left\{\rho \log(MM - \hat{O}) + \mu_{pm} - \rho\mu_{mm} - (1-\rho^2)\sigma^2\right\}, m\right) \ .$$

where $m = 6$ is the minimum allowed value for S. The second method uses an empirical Bayes approach, treating S as a random variable and defines the logged

background adjusted PM intensity as

$$\tilde{s} \quad = \quad \mathrm{E}\Big[\log(S)|S>0, PM, MM\Big] \ .$$

With m now defined as the smallest value of S with positive probability, the prior distribution for $\log(S)$ is set to uniform on the interval $[\log(m), \log(2^{16})]$, and assuming independence between the vector (N_{PM}, N_{MM}) and S, the logged background adjusted PM intensity $\tilde{s}$ is computed using numerical integration.

B Finding differentially expressed genes

This appendix is a detailed description of 11 procedures for ranking genes with respect to differential expression. The described methods are evaluated in Section 4.1 together with the methods PLW and LMW proposed in Paper IV. Although some of the methods can be applied to general linear comparisons of design-parameters, the descriptions here are restricted to a comparison of two groups of n_1 and n_2 arrays, respectively, with no missing values. With this restriction all but the last four methods are based on group-mean-differences and the corresponding sample standard deviation calculated separately for each gene or probe-set. Formally, let x_{ijg} denote the measured gene expression level (generally on log-scale) of gene g for the i'th array of group j for a total of G genes and $n_1 + n_2$ arrays, and put

$$D_g = \bar{x}_{1g} - \bar{x}_{2g} \quad \text{and} \quad s_g^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1-1)s_{1g}^2 + (n_2-1)s_{2g}^2}{n_1+n_2-2} \ , \tag{4}$$

where $\bar{x}_{jg}$ is the sample mean and s_{jg} the sample standard deviation of the measured gene expression levels for group j and gene g. Given the two summary statistics in (4) for each gene the statistic used when ranking genes using the so called fold change (FC) and the ordinary t-test are obtained directly according to

$$\text{FC}: \quad D_g \ ,$$

$$\text{t-test}: \quad \frac{D_g}{s_g} \ .$$

B.1 SAM

In the Significance Analysis of Microarrays method (SAM) suggested by Tusher et al. (2001) genes are ranked with respect to the absolute value of

$$d_g = \frac{D_g}{s_g + s_0} \ .$$

The principle used when setting the value of s_0 is that the variability of d_g should be independent of the level of s_g. This is achieved by computing the variability of d_g as a function of s_g in windows across the data. The Median Absolute Distance (MAD) is used the estimate the variability and 100 windows with equal number of data points are used by default. With mad_k equal to the estimated variability of d_g within the k'th window, s_0 is chosen so that the coefficient of variation of $mad_1, \ldots, mad_{100}$ is as small as possible.

B.2 Efron-t

The moderated t-statistic suggested by Efron et al. (2001), here denoted by Efron-t, uses the same statistic as the SAM method described in the previous section, but with a different procedure for setting the parameter s_0. In Efron-t s_0 is equal to the 90% quantile of $s_1, \ldots, s_G$. A case study of a paired comparison of 8 arrays is used to empirically motivate the choice of the 90% quantile.

B.3 LIMMA

The empirical Bayesian t-test implemented in the R-package `limma` (Smyth, 2004), based on the methods presented by Lönnstedt and Speed (2002), assume prior knowledge on the unknown gene-specific variances in terms of a Γ^{-1}-distribution. For the summary statistics in (4) this can be described as

$$\begin{aligned} D_g|\sigma_g^2 &\sim N(\delta_g, \sigma_g^2) \ , \\ s_g^2|\sigma_g^2 &\sim \tfrac{\sigma_g^2}{df}\chi_{df}^2 \ , \\ \sigma_g^2 &\sim \Gamma^{-1}(\tfrac{1}{2}m, \tfrac{1}{2}m\nu) \ , \end{aligned} \tag{5}$$

where δ_g represents the logged fold-change between the groups, N is the normal distribution, χ_{df}^2 is the χ^2-distribution with $df = n_1 + n_2 - 2$ degrees of freedom, and $\Gamma^{-1}(\alpha, \beta)$ is the inverse-gamma distribution with density function

$$f(x) = \frac{\beta^\alpha x^{-(\alpha+1)}}{\Gamma(\alpha)} \exp\{-\frac{\beta}{x}\} \quad , \quad x > 0 \ .$$

As specified in (5), the Γ^{-1}-prior for the unknown gene-specific variances is equivalent to a prior estimator equal to ν with m degrees of freedom. The marginal distribution of $\log(s_g^2)$ is a shifted Fisher's z-distribution (Johnson et al. 1995, page 78) with

$$\mathrm{E}\Big[\log(s_g^2)\Big] = \log(m\nu/df) + \psi(\tfrac{1}{2}df) - \psi(\tfrac{1}{2}m) \ , \tag{6}$$

$$\mathrm{var}\Big(\log(s_g^2)\Big) = \psi'(\tfrac{1}{2}m) + \psi'(\tfrac{1}{2}df) \ , \tag{7}$$

where ψ and ψ' are the digamma and trigamma functions, respectively. The hyperparameters m and ν are estimated by equating the sample mean and squared sample standard deviation of $\log(s_g^2)$ with the theoretical mean and variance, respectively. The parameters m and ν are then treated as known and set equal to the estimated values, and the moderated t-test described in Smyth (2004), here denoted by LIMMA, is defined as

$$\tilde{t}_g = \frac{D_g}{\sqrt{\dfrac{m\nu + df\, s_g^2}{m + df}}} \ ,$$

and under H_0, gene g is unregulated, it is shown that $\tilde{t}_g$ is t-distributed with $m + df$ degrees of freedom.

B.4 IBMT

The Intensity Based Moderated T-test (IBMT) builds on the empirical Bayesian t-test LIMMA described in the previous section. Let $\bar{x}_g$ denote the grand mean intensity of gene g calculated across all $n_1 + n_2$ arrays. The difference between LIMMA and IBMT is that the global prior variance estimator ν in (5) is modeled as a smooth function of $\bar{x}_g$, and thus is gene-specific. The model parameters are estimated as follows. A loess-curve is fitted in the scatter-plot of $\log(s_g^2)$ versus $\bar{x}_g$. With $f(\,\cdot\,)$ denoting the fitted curve, the parameter m is estimated by equating the theoretical variance of $\log(s_g^2)$ in (7) with the mean residual sum of squares

$$MSS = \frac{1}{G-1}\sum_{g=1}^{G}\left(\log(s_g^2) - f(\bar{x}_g)\right)^2 \ .$$

With m set equal to the estimated value in (6), the gene-specific prior variance estimator ν_g is found by setting (6) equal to $f(\bar{x}_g)$ and solving for ν_g,

$$\nu_g = \exp\left\{f(\bar{x}_g) - \log(m/df) - \psi(\tfrac{1}{2}df) + \psi(\tfrac{1}{2}m)\right\} \ .$$

The IBMT statistic is defined as

$$\tilde{t}_g = \frac{D_g}{\sqrt{\dfrac{m\nu_g + df\, s_g^2}{m + df}}} \ ,$$

and thus except for the addition of prior variance estimators ν_g being variance specific, the IBMT statistic is identical to the moderated t-test LIMMA (Smyth, 2004) described in the previous section.

B.5 Shrink-t

The moderated t-statistic suggested by Opgen-Rhein and Strimmer (2007) is based on the James-Stein ensemble shrinkage estimation rule (Gruber, 1998). Applied to the gene-specific variance estimators $s_1^2, \ldots, s_G^2$ the rule results in adjusted estimators defined as

$$\tilde{s}_g^2 = \hat{\lambda} s_0^2 + (1 - \hat{\lambda}) s_g^2$$

where $\hat{\lambda}$ is the estimated pooling parameter

$$\hat{\lambda} = \min\left(1, \frac{\sum_{g=1}^{G} \widehat{\mathrm{var}(s_g^2)}}{\sum_{g=1}^{G}(s_g^2 - s_0^2)^2}\right) \ .$$

The target estimator s_0^2 is the median of $s_1^2, \ldots, s_G^2$, and $\widehat{\mathrm{var}(s_g^2)}$ is estimated by

$$\frac{(n_1+n_2)^3}{(n_1+n_2-1)^3}\sum_{j=1}^{2}\sum_{i=1}^{n_j}\left(\frac{(x_{ijg} - \bar{x}_{jg})^2}{n_1 n_2} - \frac{n_1+n_2-2}{(n_1+n_2)^2}s_g^2\right)^2 \ .$$

The shrink-t statistic is then defined as

$$\tilde{t}_g = \frac{D_g}{\sqrt{\tilde{s}_g^2}} \ .$$

B.6 LPE

Jain et al. (2003) propose the Local-Pooled-Error (LPE) method for estimating variances and use group-median-differences instead of differences of group-wise mean values. For each of the two groups the error variance is evaluated as follows. For all possible pairs of arrays, with X_1 and X_2 denoting the expression indexes of the first and second array of the pairs, respectively, the difference in gene expression M, as $X_1 - X_2$ as well as $X_2 - X_1$, and the average $A = \frac{1}{2}(X_1 + X_2)$ is calculated. The sample variance of expression indexes as a function of intensity level is evaluated using the variability of M in windows with equal number of data-points across the vector A. With nw equal to the number of data-points and with variances and medians computed within each window, a local regression model is fitted to the sample variance of M times $\frac{1}{2}(nw - \frac{1}{2})/(nw - 1)$ with the median of A as regressor. The group-specific error variance as a function of intensity level is then taken as the fitted curve.

With med_{jg} denoting the median measured gene expression level for group $j = 1, 2$, the LPE statistic is then defined as

$$z_g = \frac{med_{1g} - med_{2g}}{\sqrt{\frac{\pi}{2}\left[\frac{\sigma_1^2(med_{1g})}{n_1} + \frac{\sigma_2^2(med_{2g})}{n_2}\right]}} ,$$

where $\sigma_j^2(\,\cdot\,)$ is the error variance for group $j = 1, 2$. The estimator of the variance of group-median-differences in the denominator of z_g is based on the asymptotic variance for the median of a normal distributed sample. With sample variance equal to τ^2, the asymptotic variance for the median is $\frac{1}{2}\pi\tau^2$ divided by the sample size (Mood et al., 1998).

B.7 WAME

The empirical Bayes WAME-model (Kristiansson et al., 2005, 2006; Sjögren et al., 2007) uses a global covariance structure to model dependencies and differing variances between arrays. With x_g equal to the vector of measured gene expression levels of gene g across a set of n arrays, the WAME-model is defined as

$$x_g|c_g \sim \mathrm{N}_n(\mu_g, c_g\Sigma) ,$$

$$c_g \sim \Gamma^{-1}(\tfrac{1}{2}m, \tfrac{1}{2}m\nu).$$

Here $\mathrm{N}_n(\mu, \Sigma)$ denotes an n-dimensional normal distribution with mean μ and covariance matrix Σ and $\Gamma^{-1}(\alpha, \beta)$ is the inverse-gamma distribution with density function $f(x) = \beta^\alpha \exp\{-\frac{\beta}{x}\}x^{-(\alpha+1)}/\Gamma(\alpha)$ for $x > 0$. The expression profile μ_g is defined in terms of a global design matrix D and gene-specific parameter vector γ_g. A contrast matrix C is used to specify the linear combination δ_g of the parameter vector that is of interest. In summary,

$$\mu_g = D\gamma_g \quad \text{and} \quad \delta_g = C\gamma_g .$$

The special case considered in this appendix is a two group comparison using one-color arrays. Let γ_{jg} denote the underlying expression level of gene g for group

j, and $\delta_g = \gamma_{2g} - \gamma_{1g}$ represent the logged fold-change between groups 2 and 1. For this design the 1×2 contrast matrix $C = [-1 \;\; 1]$ can be used together with an $n \times 2$ design matrix D. For example, with $n_1 = 3$ and with $n_2 = 4$ the design matrix would be

$$D^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} .$$

With the WAME-model presented as in Paper IV, a transformation matrix M is derived using D and C. Put

$$A_0 = I - D(D^T D)^{-1} D^T \quad \text{and} \quad B = D(D^T D)^{-1} C^T \tag{8}$$

and let A be an $n \times (n-2)$ matrix of full rank whose column space equals that of A_0. The $n \times (n-1)$ transformation matrix M is then defined and used to compute vectors z_g of *transformed* expression levels according to

$$M = [A; B] \quad \text{and} \quad z_g = M^T x_g \tag{9}$$

giving the reduced model

$$\begin{aligned} z_g | c_g &\sim \mathrm{N}_q((0, \ldots, 0, \delta_g)^T, c_g \Sigma_z) \; , \\ c_g &\sim \Gamma^{-1}(\tfrac{1}{2}m, \tfrac{1}{2}m\nu) \; , \end{aligned} \tag{10}$$

where $q = n - 1$ and $\Sigma_z = M^t \Sigma M$.

Given estimators of m, ν, and Σ_z, for example obtained using one of the methods presented in Paper III, these parameters are treated as known, and a weighted moderated t-test is derived. The unbiased minimum variance estimator of δ_g is

$$\hat{\delta}_g = (\lambda^T \Sigma_z^{-1} \lambda)^{-1} \lambda^T \Sigma_z^{-1} z_g \; , \tag{11}$$

where λ is the vector $(0, \ldots, 0, 1)^T$ of length q. The weighted moderated t-statistic is then defined as

$$\tilde{t}_g = \sqrt{\frac{q + m - 1}{(\lambda^T \Sigma_z^{-1} \lambda)^{-1}}} \frac{\hat{\delta}_g}{\sqrt{m\nu + \mathrm{RSS}_g}} \; ,$$

and for $\delta_g = 0$ it is shown that $\tilde{t}_g$ follows a t-distribution with $q + m - 1$ degrees of freedom. Here

$$\mathrm{RSS}_g = z_g^T \left(\Sigma_z^{-1} - \Sigma_z^{-1} \lambda (\lambda^T \Sigma_z^{-1} \lambda)^{-1} \lambda^T \Sigma_z^{-1} \right) z_g \tag{12}$$

is the weighted residual sum of squares. See Kristiansson et al. (2006) for details.

B.8 LMW

The WAME-model, suggested by Kristiansson et al. (2005), is in Paper IV extended to incorporate the dependency between variability and intensity level that often exists in microarray data, even for log-transformed data. Let μ_g denote the profile for gene

g across the n arrays with mean intensity level $\bar{\mu}_g$, and let x_g represent the vector of measured gene expression levels (generally log-transformed) of gene g across the n arrays. To account for the dependency between variability and intensity-level the scale-parameter of the Γ^{-1}-distribution depends on the mean intensity level $\bar{\mu}_g$ for the gene through the smooth function ν. Formally,

$$\begin{aligned} x_g|c_g &\sim \mathrm{N}_n(\mu_g, c_g\Sigma) \ , \\ c_g &\sim \Gamma^{-1}(\tfrac{1}{2}m, \tfrac{1}{2}m \cdot \nu(\bar{\mu}_g)) \ , \end{aligned} \tag{13}$$

where Σ is an $n \times n$ covariance matrix, m is a real-valued parameter, and $\nu(\,\cdot\,)$ is a smooth real-valued function. N_n denotes an n-dimensional normal distribution, and $\Gamma^{-1}(a, b)$ denotes the inverse-gamma distribution with shape parameter a and scale parameter b. A cubic spline is used to parameterize the function $\nu(\,\cdot\,)$. Given a set of K interior spline-knots

$$\nu(x) = \exp\{H(x)^T\beta\} \ ,$$

where β is a parameter vector of length $2K-1$ and $H : \mathbb{R} \to \mathbb{R}^{2K-1}$ is a set of B-spline basis functions, see chapter 5 of Hastie et al. (2001).

A transformation matrix M, transformed vectors z_g, and a reduced model are derived as described in (8) to (10) in the previous section with $\nu(\bar{\mu}_g)$ replacing ν in the reduced model (10). Estimators of m, Σ_z, and β are computed using the EM-algorithm (Dempster et al., 1977) and these parameters are then treated as known (see Paper IV for details).

With $\lambda = (0, \ldots, 0, 1)^T$ of length $q = n - 1$, the Locally Moderated Weighted-t statistic (LMW) suggested in Paper IV is then defined as

$$\tilde{t}_g = \sqrt{\frac{q+m-1}{(\lambda^T\Sigma_z^{-1}\lambda)^{-1}}}\,\frac{\hat{\delta}_g}{\sqrt{m\hat{\nu}_g + \mathrm{RSS}_g}} \ , \tag{14}$$

with $\hat{\delta}_g$ defined as in (11), RSS_g computed according to (12), and where

$$\hat{\nu}_g = \exp\{H(\bar{x}_g)^T\beta\} \ .$$

B.9 PLW

The Probe level Locally moderated Weighted median-t (PLW) method suggested in Paper IV, is specially designed for Affymetrix arrays, or other multiple probe arrays. PLW uses the same model as the LMW method, also suggested in Paper IV and described in the previous section. However, model (13) is in PLW applied to perfect match (PM) probe intensities whereas PLW applies the model to probe-set summaries.

In detail, let y_{ip} be the background corrected and normalized log-intensity on array i for PM probe p and put $y_p = (y_{1p}, \ldots, y_{np})^T$. The locally moderated t-statistics in (14) is computed for all PM probes as described in the previous section with y_p replacing the vectors x_g of measured gene expression levels. As described in Section 2.1.2, the PM probes are divided into G (disjoint) probe-sets $\mathcal{G}_1, \ldots, \mathcal{G}_G$.

With $\tilde{t}_p$ representing the t-statistic for probe p, the PLW statistic for the probe-set $\mathcal{G}$ is then defined as

$$\text{PLW}_{\mathcal{G}} = \text{median}\Big\{\tilde{t}_p : p \in \mathcal{G}\Big\} \ .$$

References

Affymetrix. *Statistical Algorithms Description Document*. Affymetrix, Santa Clara , California, 2002. URL `http://www.affymetrix.com`.

Affymetrix. *Expression Analysis Technical Manual*. Affymetrix, Santa Clara , California, 2004. URL `http://www.affymetrix.com`.

P. Baldi and A. D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, 2001.

M. J. Callow, S. Dudoit, E. L. 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.

P. Clote and R. Backofen. *Computational Molecular Biology: An Introduction*. John Wiley & Sons, August 2000. ISBN 978-0-471-87252-8.

J. Comander, S. Natarajan, M. Gimbrone, and G. Garcia-Cardena. Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation. *BMC Genomics*, 5(1):17, 2004.

F. Crick. On protein synthesis. *The Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

F. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, August 1970.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39:1–38, 1977.

B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18:S105–S110, 2002.

B. P. Durbin and D. M. Rocke. Estimation of transformation parameters for microarray data. *Bioinformatics*, 19(11):1360–1367, 2003.

I. A. Eaves, L. S. Wicker, G. Ghandour, P. A. Lyons, L. B. Peterson, J. A. Todd, and R. J. Glynne. Combining Mouse Congenic Strains and Microarray Gene Expression Analyses to Study a Complex Trait: The NOD Model of Type 1 Diabetes. *Genome Res.*, 12(2):232–243, 2002.

B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96:1151–1160, December 2001.

S. C. Geller, J. P. Gregg, P. Hagerman, and D. M. Rocke. Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, 19(14):1817–1823, 2003.

M. H. J Gruber. *Improving Efficiency by Shrinkage*. Marcel Dekker, Inc, New York, 1998.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, volume 1. Springer, first edition, 2001. ISBN 0-387-95284-5.

A. M. Hein and S. Richardson. A powerful method for detecting differentially expressed genes from genechip arrays that does not require replicates. *BMC Bioinformatics*, 7(1):353, 2006.

A. M. Hein, S. Richardson, H. C. Causton, G. K. Ambler, and P. J. Green. BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data. *Biostatistics*, 6(3):349–373, 2005.

D. Holder, R. F. Raubertas, V. B. Pikounis, V. Svetnik, and K. Soper. Statistical analysis of high density oligonucleotide arrars: a safer approach. *Gene Logic Workshop of Low Level Analysis of Affymetrix GeneChip Data*, 2001.

J. Hu and F. A. Wright. Assessing differential gene expression with small sample sizes in oligonucleotide arrays using a mean-variance model. *Biometrics*, 63(1): 41–49, 2007.

P. J. Huber. *Robust Statistics*. John Wiley & Sons, 1981. ISBN 0471725242.

W. Huber, A. von Heydebreck, H. Sueltmann, A. Poustka, and M. Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Stat. Appl. Genet. Mol. Biol.*, 2(1):article 3, 2003.

W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18:S96–104, 2002.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human gspot. *Nature*, 431(7011):931–945, 2004.

R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucl. Acids Res.*, 31(4):e15, 2003a.

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, 2003b.

N. Jain, J. Thatte, T. Braciale, K. Ley, M. O'Connell, and J. K. Lee. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics*, 19(15):1945–1951, 2003.

N. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 2. John Wiley and Sons, second edition, 1995. ISBN 978-0-471-58494-0.

E. Kristiansson, A. Sjögren, M. Rudemo, and O. Nerman. Weighted analysis of paired microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 4(1):article 30, 2005.

E. Kristiansson, A. Sjögren, M. Rudemo, and O. Nerman. Quality optimised analysis of general paired microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 5(1): article 10, 2006.

W. Lemon, S. Liyanarachchi, and M. You. A high performance test of differential gene expression for oligonucleotide arrays. *Genome Biology*, 4(10):R67, 2003.

C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36, 2001a.

C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, 2(8): research0032.1–research0032.11, 2001b.

X. Liu, M. Milo, N. D. Lawrence, and M. Rattray. A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18): 3637–3644, 2005.

X. Liu, M. Milo, N. D. Lawrence, and M. Rattray. Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics*, 22 (17):2107–2113, 2006.

D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Norton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680, 1996.

I. Lönnstedt and T. P. Speed. Replicated microarray data. *Statistica Sinica*, 12(1): 31–46, 2002.

M Milo, A. Fazeli, M. Niranjan, and N. D. Lawrence. A probabilistic model for the extraction of expression levels from oligonucleotide arrays. *Biochem. Soc. Trans.*, 31:1510–1512, 2003.

A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, New York, 1998.

P. Munson. A 'consistency' test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations. *Gene Logic Workshop of Low Level Analysis of Affymetrix GeneChip Data*, 2001.

F. Naef and M. O. Magnasco. Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Phys. Rev. E*, 68(1):011906, Jul 2003.

R. Opgen-Rhein and K. Strimmer. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat. Appl. Genet. Mol. Biol.*, 6(1): article 9, 2007.

D. M. Rocke and B. Durbin. Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics*, 19(8):966–972, 2003.

M. A. Sartor, C. R. Tomlinson, S. C. Wesselkamper, S. Sivaganesan, G. D. Leikauf, and M. Medvedovic. Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Genomics*, 7 (1):article 538, 2006.

M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270 (5235):467–470, October 1995.

A. Sjögren, E. Kristiansson, M. Rudemo, and O. Nerman. Weighted analysis of general microarray experiments. *BMC Bioinformatics*, 8(1):article 387, 2007.

G. K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3(1):article 3, 2004.

J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(9):5116–5121, 2001.

J. Watson and F. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171:737–738, April 1953.

Z. Wu, R. Irizarry, R. Gentleman, F. M. Murillo, and F. Spencer. A model based background adjustment for oligonucleotide expression arrays. Technical report, Johns Hopkins University, Department of Biostatistics, http://www.bepress.com/jhubiostat/paper1/, 2004.

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. *Nucl. Acids Res.*, 30(4):e15–, 2002.

L Zhou and D. M. Rocke. An expression index for affymetrix genechips based on the generalized logarithm. *Bioinformatics*, 21(21):3983–3989, 2005.

Paper I

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 10, Number 1, 2003

Pp. 95–102

Contrast Normalization of Oligonucleotide Arrays

MAGNUS ÅSTRAND

ABSTRACT

Affymetrix high-density oligonucleotide array is a tool that has the capacity to simultaneously measure the abundance of thousands of mRNA sequences in biological samples. In order to allow direct array-to-array comparisons, normalization is a necessity. When deciding on an appropriate normalization procedure there are a couple questions that need to be addressed, e.g., on which level should the normalization be performed: On the level of feature intensities or on the level of expression indexes? Should all features/expression indexes be used or can we choose a subset of features likely to be unregulated? Another question is how to actually perform the normalization: normalize using the overall mean intensity or use a smooth normalization curve? Most of the currently used normalization methods are linear; e.g., the normalization method implemented in the Affymetrix software GeneChip is based on the overall mean intensity. However, along with alternative methods of summarizing feature intensities into an expression index, nonlinear methods have recently started to appear. For many of these alternative methods, the natural choice is to normalize on the level of feature intensities, either using all feature intensities or only perfect match intensities. In this report, a nonlinear normalization procedure aimed for normalizing feature intensities is proposed.

Key words: oligonucleotide array, normalize, curve-fitting, orthogonal, loess.

1. INTRODUCTION

THE USE OF MICROARRAYS TO MEASURE ABUNDANCE of mRNA sequences in biological samples has emerged the last couple of years. One technology commonly used in this context is Affymetrix oligonucleotide arrays. The starting point of this technology is a sample of cells or tissue from which the researcher isolates RNA from which complementary DNA (cDNA) is generated. Then follows transcription from the cDNA to complementary RNA (cRNA), which after fragmentation is put to hybridize on the array. After the hybridization, excess cRNA is washed off, and the final step before scanning the array is staining. The result, after the researchers efforts, is the scanned intensity image, which is the starting point of the low-level analysis of microarrays such as image analysis, feature extraction, and normalization.

Image analysis and feature extraction are in themselves a great challenge. The aim is to select pixels representing each feature and summarize them into a feature intensity. Since each feature is represented by approximately 8×8 pixels, this a great reduction of the data. However, this issue will not be addressed further here; instead, procedures on the level of feature intensity or higher will be discussed.

Statistical and Mathematical Science, AstraZeneca R & D Mölndal, S-431 83 Mölndal, Sweden.

When analyzing data from oligonucleotide arrays, normalizing is a necessity to allow direct array-to-array comparisons. This is because the overall brightness of the scanned image can differ substantially from one array to the next. An even better understanding of this problem is attained by scatter plots with the feature intensities of one array on the y-axis and the feature intensities of another array on the x-axis (Fig. 1). The main sources of this variation in feature intensity level between arrays are the different steps prior to obtaining the intensity image together with the quality of the arrays.

The most commonly used normalization procedure is probably the one implemented in the Affymetrix software GeneChip. This procedure is based on the Affymetrix expression index *average difference* (AD), which is the average difference between the perfect match intensity (PM) and the mismatch intensity (MM). For each array, a trimmed mean of all probe's AD is calculated, and normalization factors for AD are determined by ratios of such means or by a ratio to a target mean AD. Hence, this is a linear procedure on the level of expression index where all probes are used.

Recently, alternatives to the Affymetrix expression index AD have started to appear, e.g., the model-based expression index (MBEI) introduced by Li and Wong (2001a) and an index based on log(PM-BG), suggested by Irizarry *et al.* (2002), where BG is a global estimate of the background intensity of the array. For such alternative expression indexes, it is more natural to normalize on a lower level of the data, i.e., on the level of feature intensities using only PM or PM and MM together.

Such a normalization procedure is described by Li and Wong (2001b). In this procedure, a baseline array is selected to which the other arrays are normalized by fitting a smooth curve. Prior to fitting the curve, a subset of features with small absolute rank differences is selected. The argument for using this kind of subset selection is that we can expect features belonging to an unregulated gene to have similar intensity ranks on two arrays. The curve is then fitted using these features only. In contrast to the normalization procedure in GeneChip, this is a nonlinear procedure on the level of feature intensities that uses a subset of features. Another procedure suggested by Bolstad *et al.* (2002) uses the distribution of all feature intensities. An *average distribution* is derived by first computing the quintiles of each array separately, and then the quintiles are averaged across the arrays. In relation to the method in Li and Wong (2001b), this procedure uses all features. It is also substantially simpler.

In this report, a method for normalizing using smooth curves is proposed. It is a method meant for normalizing the feature intensities, i.e., the PM and MM intensities. But the method can just as well be applied to PM-MM or an expression index derived from the feature intensities. The method proposed by Li and Wong (2001b) uses smooth curves fitted in scatter plots with the baseline array on the y-axis and the array to be normalized on the x-axis. Another solution is to fit a smooth curve in scatter plots with the feature intensity differences on the y-axis and the intensity means on the x-axis (often the intensities are logged before computing the differences and means). This is the basis of the method proposed in this report. We will start by describing the proposed method, termed *contrast normalization* (CN), and then discuss it together with the other methods mentioned. We will also have a look at how these methods perform.

2. RESULTS

2.1. Contrast normalization

Suppose we have a set of k arrays that are to be normalized; each array is represented by n feature intensities. Let the nxk matrix Y denote the intensities of these arrays. Hence, the element in row i and column j of Y is the unnormalized feature intensity of feature i on array j.

2.1.1. Change of basis. In the first step, these intensities are logged and transformed using the matrix M:

$$Z = [x, y_1, \ldots, y_{k-1}] = \log(Y) \cdot M'. \tag{1}$$

Here, M is an orthonormal kxk matrix; i.e., the rows of M are mutually orthogonal unit vectors. Moreover, the first row of M is always the 1-vector times $\sqrt{1/k}$, and then it follows that the other rows are a set of orthonormal contrast. Matrixes such as M will be called *transformation matrixes* hereafter. Note that with

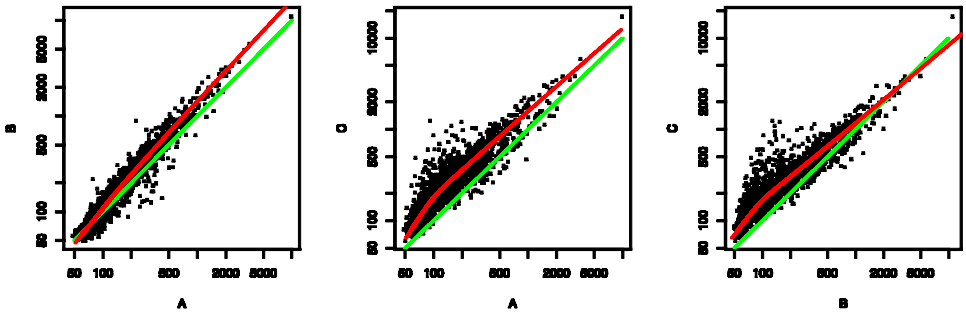

FIG. 1. Scatter plots of 3 arrays, A, B, and C, prior to normalizing. The red curve is the fitted normalizing curve fitted using the alternative basis shown in Fig. 2.

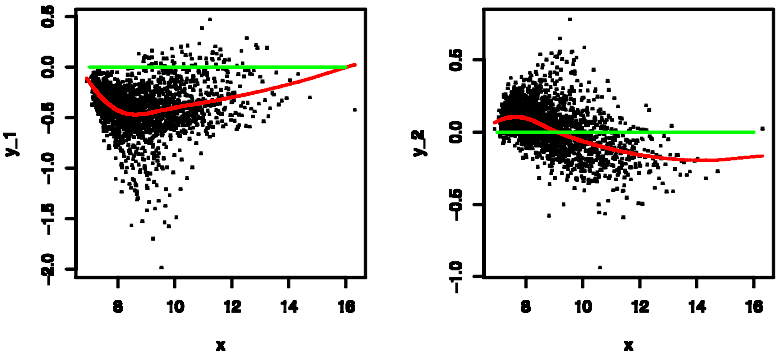

FIG. 2. Contrast plots. Scatter plots of the 2 contrasts against the mean for 3 arrays, A, B, and C, prior to normalizing. The red curve is the fitted normalizing curve, and the green line is the reference line.

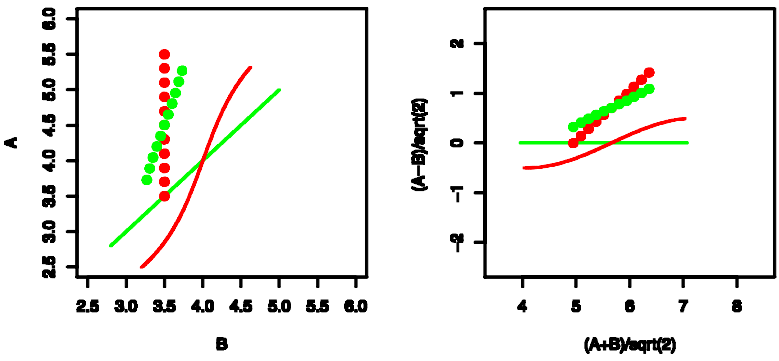

FIG. 3. Normalizing prior change to original basis. This figure illustrates how feature intensities from two arrays are normalized using the alternative basis. The graphs show logged intensities in original basis (left graph) and alternative basis (right graph). Red line is the fitted normalizing curve (fitted using the alternative basis in the right graph), and green line the reference line. The graphs show a couple of features which all have intensities equal to 3.5 on array B, and intensities ranging from 3.5 to 5.5 on array A prior to normalizing (red dots). The intensities are normalized using the alternative basis in the right graph yielding the normalized intensities shown as green dots. Now the features have intensities ranging from 3.25 to 3.75 on array B.

this specification M is unique for k equals 2, but this is not the case for $k > 2$. When k equals 2, we get $M = M_2$, and when k equals 4 we can use $M = M_4$:

$$M_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \sqrt{\frac{1}{2}} \qquad M_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \frac{1}{2} \tag{2}$$

This use of an orthonormal matrix is just a change of basis, where the rows of M form the new basis, denoted the *alternative basis* from now on. When k equals 2, we see that, besides a constant, the alternative basis corresponds to what is called a Bland and Altman plot of the log feature intensities, i.e., a plot of the difference versus the mean. The change to logarithmic scale before performing the change of basis is used to make the error variances more homogenous.

2.1.2. Fitting the normalizing curve. Using the alternative basis, we then fit the normalizing curve: we use the first column of the transformed intensities in Z, i.e., x, as a predictor for column $2, \ldots, k$ of Z, i.e., $y_1, \ldots, y_{k-1}$. When doing this, it's important to have in mind that the set of orthonormal contrasts is not unique. Thus, the method for fitting the curve should be invariant with respect to choice of contrast. Suppose that $[x, y_{a1}, \ldots]$ and $[x, y_{b1}, \ldots]$ are the intensities obtained when transforming according to (1) using the transformation matrixes M_a and M_b, respectively. If $\hat{y}_{a1}, \ldots$ and $\hat{y}_{b1}, \ldots$ are the corresponding fitted curves, $[x, \hat{y}_{a1}, \ldots] \cdot M_a$ should equal $[x, \hat{y}_{b1}, \ldots] \cdot M_b$.

In order to achieve this, we fit a smooth curve using a local regression model (loess) to each vector y_i. For the curve to be less sensitive for outliers, we use a redescending M estimator with the bisquare weight function as is done in the R-function loess (Chambers and Hastie, 1997), but with one important modification. If $\hat{y}_1, \ldots, \hat{y}_{k-1}$ are the vectors of the fitted values, we take $\hat{\epsilon}$ as the Euclidian distance between the rows of the $n \times (k-1)$ matrixes $[y_1, \ldots, y_{k-1}]$ and $[\hat{y}_1, \ldots, \hat{y}_{k-1}]$.

$$\hat{\epsilon} = \sqrt{\sum_{i=1}^{k-1} (\hat{y}_i - y_i)^2} \tag{3}$$

Thus, in each iteration, the same set of robust weights is used for each of the $k-1$ contrast vectors, and these weights are invariant to the choice of orthonormal contrasts. Further, since the local regression model is fitted using weighted least squares, the fitted curve is invariant to the choice of orthonormal contrasts.

2.1.3. Normalizing the arrays. The normalizing curve can be represented with the matrix $[x, \hat{y}_1, \ldots, \hat{y}_{k-1}]$. These sets of points can be viewed either using the original basis or the alternative basis, the red curves in Figs. 1 and 2, respectively. Hence, we still could choose a baseline array and normalize the others by the fitted normalizing curve, e.g., using the two rightmost graphs in Fig. 1 and normalizing A and B to the baseline array C. If doing so, we have used a normalizing curve that is invariant to the choice of baseline array. But the scale to which we normalize still depends on which baseline array we choose.

Another way of normalizing the arrays using the fitted curve is to simply subtract the fitted values using the alternative basis, i.e., Fig. 2, and then go back to the original basis using the matrix M. In this case, the normalized and unlogged intensities would be

$$\exp\left\{[x, y_1 - \hat{y}_1, \ldots, y_{k-1} - \hat{y}_{k-1}] \cdot M\right\}. \tag{4}$$

But this results in a nonsmooth normalizing procedure, in the sense that intensities being equal on one array prior to normalizing may not be equal after. Figure 3 shows why this is the case.

However, we can still use the normalizing curve with the alternative basis. The matrix $[x, \hat{y}_1, \ldots, \hat{y}_{k-1}]$ is a representation of the normalizing curve, and the matrix $[x, 0, \ldots, 0]$ is a representation of what the curve should be after normalization. Hence, the mapping

$$[x, \hat{y}_1, \ldots, \hat{y}_{k-1}] \longmapsto [x, 0, \ldots, 0] \tag{5}$$

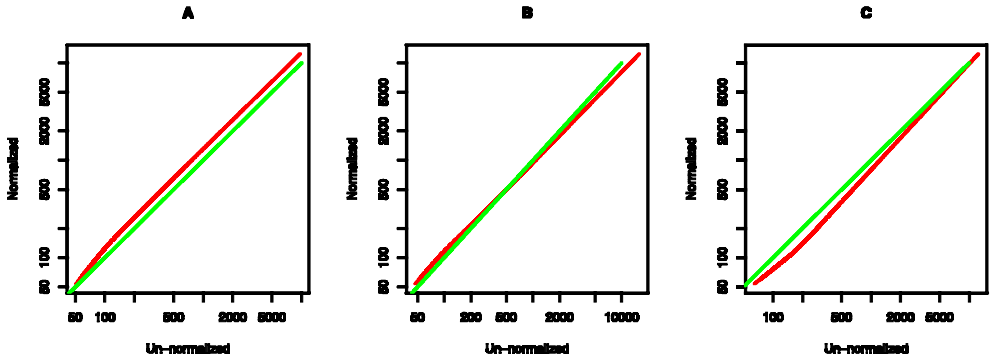

FIG. 4. Normalizing functions. The normalizing functions defined through the mapping of the fitted normalizing curve on to the reference line in Fig. 2 are shown for three arrays, A, B, and C (red line). The green line is the reference line with slope one and zero intercept.

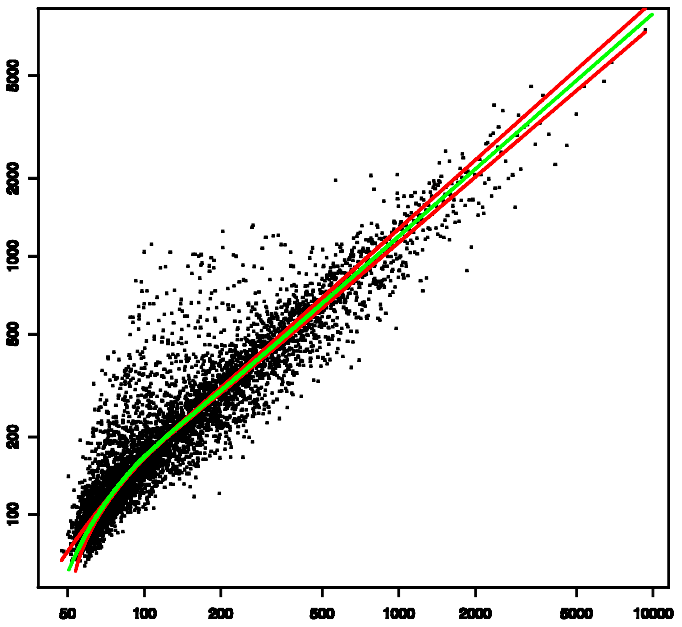

FIG. 5. Scatter plot of 2 arrays, A and B. The two red lines are fitted loess curves from using A (and B) as a predictor for B (and A). The green curve is a loess curve fitted using the alternative basis.

defines a transformation that does the job of evening out the contrast for the alternative basis. Moreover, the mapping

$$\exp\left\{[x, \hat{y}_1, \ldots, \hat{y}_{k-1}] \cdot M\right\} \longmapsto \exp\{[x, 0, \ldots, 0] \cdot M\} \tag{6}$$

defines the same transformation but for the original basis and anti-logged scale. This transformation forms a function $F : R^k \mapsto R^k$ that row-by-row normalizes the matrix of intensities Y. Thus, if $F((x_1, \ldots, x_k)) = (f_1(x_1), \ldots, f_k(x_k))$, f_j is function that normalizes array j. These functions, f_1, f_2, and f_3 for the set of three arrays A, B, and C, are shown in Fig. 4.

One may note that

$$[x, 0, \ldots, 0] \cdot M = \frac{1}{\sqrt{k}} \cdot [x, x, \ldots, x] \tag{7}$$

where $x/\sqrt{k}$ equals $\overline{\log(Y)}$, i.e., the mean across the rows of $\log(Y)$. Hence, this procedure normalizes to a scale determined by $\exp\{\overline{\log(Y)}\}$, i.e., the geometric mean of the arrays.

2.1.4. Adding arrays. Suppose a set of arrays have been normalized and further analyzed, e.g., expression indexes have been computed. Now we have an additional set of arrays that we would like to add to the original ones to use in the same analysis. We would like to do this without affecting the intensities of the original set. This can be done by first normalizing the new set of arrays separately. These arrays are then normalized to a scale determined by their geometrical mean. Thus, we have to transform these to the same scale as the original set, i.e., the scale determined by the geometrical mean of the arrays in the original set.

Let Y_1 and Y_2 be the normalized intensities of the original and new set of arrays, respectively. Also, let x_1 and x_2 be the mean across the rows of $\log(Y_1)$ and $\log(Y_2)$, respectively. To find a transformation that transforms the new arrays with the same scale as the original arrays, we apply the normalizing method treating x_1 and x_2 as log-intensities of two "arrays." If $\hat{y}_1$ is the fitted values for the contrast of these "arrays," we have the mappings

$$\exp\left\{\frac{x_1 + x_2}{2} + \frac{\hat{y}_1}{\sqrt{2}}\right\} \longmapsto \exp\left\{\frac{x_1 + x_2}{2}\right\}, \tag{8}$$

$$\exp\left\{\frac{x_1 + x_2}{2} - \frac{\hat{y}_1}{\sqrt{2}}\right\} \longmapsto \exp\left\{\frac{x_1 + x_2}{2}\right\} \tag{9}$$

that form the functions f_1 and f_2 that would normalize the two "arrays" to a common scale. However, we only want to change the scale of the second one (x_2). To do this, we apply $f_1^{-1} \circ f_2$ formed by the mapping

$$\exp\left\{\frac{x_1 + x_2}{2} - \frac{\hat{y}_1}{\sqrt{2}}\right\} \longmapsto \exp\left\{\frac{x_1 + x_2}{2} + \frac{\hat{y}_1}{\sqrt{2}}\right\} \tag{10}$$

on the intensities of the second "arrays." Hence, the $f_1^{-1} \circ f_2$ is the function that transforms the intensities of the new set of arrays to the scale of the original set.

2.1.5. Software. The contrast normalization as described above is included in the R package affy as the "contrast" option in the "normalize" method (Irizarry *et al.*, 2003). The affy package is available through the open source software project Bioconductor (*www.bioconductor.org/*).

3. DISCUSSION

The usage of curve fitting by normalizing to a baseline array is perhaps the most intuitive way. However, it has one obvious drawback: we have to choose the baseline array, i.e., the array to place on the y-axis. How important this drawback is for the result of the downstream analysis of the raw intensities, i.e., computing expression indexes, is hard to tell. However, by normalizing and computing MBEI (Li and Wong, 2001a) of two arrays, A and B, first using array A as the baseline array, and then a second time using array B as the baseline array, we get an indication that it's not negligible: For each choice of baseline array, the ratios of MBEI, array A to array B, was computed. Of the 8,799 probes, the ratio differed more then 10% (10% greater or smaller) for 1,603 probes (18%), when using array A as baseline instead of array B. The difference was most notable among probes with small indexes. But even among the 2,500 probes with highest expression indexes, the ratio differed more than 10% for 10% of the probes. The two sets of probes filtered out, based on the confidence interval and absolute difference, for each choice of baseline array, differed. There were 469 and 298 probes in the two sets of which 265 were contained in both sets.

TABLE 1. STANDARD DEVIATION COMPARISON[a]

Baseline dataset	*Comparative dataset*					
	UN	*LR*	*CN2*	*CN1*	*QN*	*CN3*
UN	-(-)	89(2.44)	89(2.55)	90(2.51)	90(2.51)	90(2.53)
LR	11(1.42)	-(-)	57(1.18)	57(1.21)	58(1.19)	57(1.22)
CN2	11(1.43)	43(1.14)	-(-)	50(1.04)	53(1.05)	52(1.03)
CN1	10(1.41)	43(1.16)	50(1.04)	-(-)	52(1.02)	53(1.03)
QN	10(1.40)	42(1.14)	47(1.05)	48(1.02)	-(-)	51(1.03)
CN3	10(1.42)	43(1.17)	48(1.02)	47(1.03)	49(1.03)	-(-)

[a]Using two sets of replicated Mu11KsubA arrays (4 replicates in each set), the standard deviation (STD) for each feature across the replicates was computed. The Mu11KsubA array has a total of 262,560 features, which are used for 6,584 probes. The STD for each feature was computed using 6 different sets of intensities: Un-normalized intensities (UN), intensities normalized using linear regression (LR), normalized using QN (QN), and using CN, using all features and a loess span equals 2/3 (CN1), using a subset of 10,000 features and loess span equals 2/3 (CN2) and 0.2 (CN3). The values in upper triangle show the percentage of features of which the comparative dataset had smaller STD then the baseline dataset. For those features, the median STD ratio (baseline dataset to the comparative dataset) is shown within brackets. The values in lower triangle show the percentage of features of which the baseline dataset had smaller STD than the comparative. For those features, the median STD ratio (comparative dataset to the baseline dataset) is shown within brackets.

Moreover, Fig. 5 shows a scatter plot of the raw intensities of the same two arrays together with three curves. The two red curves are the loess curves fitted using the original basis with A as the predictor of B and vice versa. The green curve is the loess curve but fitted using the alternative basis. There is a notable difference between the two red curves with the green curve lying in between. Again, there is a clear indication that the choice of baseline array is not just a theoretical matter.

On the other hand, this approach is simple to apply to a set of k arrays: simply choose a baseline array and normalize the other to that array. Also, if an analysis has been done on a set of arrays, it's easy to add arrays without affecting the analysis of the original set. Just use the baseline array of the original arrays to normalize the new arrays.

When using CN, or the quintile normalization (QN) suggested by Bolstad *et al.* (2002), there is no choice of baseline array; instead all arrays are treated uniformly. But it's not as straightforward to add extra arrays without affecting the result of the analysis of the original set. But the solution in Section 2.1.4 does the job when using CN, and a similar solution for the QN method is to simply use the average distribution of the original set of arrays for the new arrays. Both methods normalize to a scale determined by the geometrical mean across the arrays, in contrast to normalizing using a baseline array where the baseline array determines the scale.

The normalizing procedure described in Li and Wong (2001b) uses a baseline array to which the other arrays are normalized. As mentioned, a curve is fitted using a subset of features. These subsets are derived separately for each of the arrays that are normalized and the baseline array. Hence, there is a risk of using different features when normalizing array 1 as when normalizing array 2. Another way of finding a subset of features is to compare the ranks across all arrays. This could be done using the mean square error (MSE) of the ranks or the range of the ranks. The later was used in the comparisons of Table 1.

In Table 1, the normalized feature intensity standard deviations (STD) over two sets of replicated Mu11KsubA arrays are compared. All methods reduced the standard deviation compared to unnormalized intensities, and CN and QN show somewhat smaller STD's than linear normalization. The different versions of CN (using a subset or all features and different span parameters for the loess model) and QN perform similarly with a slight tendency toward QN and CN using a subset and a small span parameter being the better ones.

REFERENCES

Bolstad, B., Irizarry, R., Åstrand, M., and Speed, T. 2002. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* (To appear).

Chambers, J.M., and Hastie, T.J. 1997. *Statistical models in S.*, Chapman and Hall.

Irizarry, R., Gautier, L., and Cope, L. 2003. An R package for analyses of affymetrix oligonucleotide arrays. In Parmigiani, G., Garrett, E.S., Irizarry, R.A., and Zeger, S.L., eds., *The Analysis of Gene Expression Data: Methods and Software*, Springer, New York. (To appear).

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. 2002. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. (To appear).

Li, C., and Wong, W.H. 2001a. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci.* 98(1), 31–36.

Li, C., and Wong, W.H. 2001b. Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biol.*, 2(8), research 0032.1–0032.11.

Address correspondence to:
Magnus Åstrand
AstraZeneca R & D Mölndal
S-431 83 Mölndal
Sweden

E-mail: magnus.astrand@astrazeneca.com

Paper II

BIOINFORMATICS

Vol. 19 no. 2 2003
Pages 185–193

A comparison of normalization methods for high density oligonucleotide array data based on variance and bias

B. M. Bolstad[1,*], R. A. Irizarry[2], M. Åstrand[3] and T. P. Speed[4,5]

[1]Group in Biostatistics, University of California, Berkeley, CA 94720, USA, [2]Department of Biostatistics, John Hopkins University, Baltimore, MD, USA, [3]AstraZeneca R & D Mölndal, Sweden, [4]Department of Statistics, University of California, Berkeley, CA 94720, USA and [5]Division of Genetics and Bioinformatics, WEHI, Melbourne, Australia

Received on June 13, 2002; revised on September 11, 2002; accepted on September 17, 2002

ABSTRACT

Motivation: When running experiments that involve multiple high density oligonucleotide arrays, it is important to remove sources of variation between arrays of non-biological origin. Normalization is a process for reducing this variation. It is common to see non-linear relations between arrays and the standard normalization provided by Affymetrix does not perform well in these situations.

Results: We present three methods of performing normalization at the probe intensity level. These methods are called complete data methods because they make use of data from all arrays in an experiment to form the normalizing relation. These algorithms are compared to two methods that make use of a baseline array: a one number scaling based algorithm and a method that uses a non-linear normalizing relation by comparing the variability and bias of an expression measure. Two publicly available datasets are used to carry out the comparisons. The simplest and quickest complete data method is found to perform favorably.

Availabilty: Software implementing all three of the complete data normalization methods is available as part of the R package Affy, which is a part of the Bioconductor project http://www.bioconductor.org.

Contact: bolstad@stat.berkeley.edu

Supplementary information: Additional figures may be found at http://www.stat.berkeley.edu/~bolstad/normalize/index.html

INTRODUCTION

The high density oligonucleotide microarray technology, as provided by the Affymetrix GeneChip®, is being used in many areas of biomedical research. As described in Lipshutz *et al.* (1999) and Warrington *et al.* (2000), oligonucleotides of 25 base pairs in length are used to probe genes. There are two types of probes: reference probes that match a target sequence exactly, called the *perfect match* (PM), and partner probes which differ from the reference probes only by a single base in the center of the sequence. These are called the *mismatch* (MM) probes. Typically 16–20 of these probe pairs, each interrogating a different part of the sequence for a gene, make up what is known as a probeset. Some more recent arrays, such as the HG-U133 arrays, use as few as 11 probes in a probeset. The intensity information from the values of each of the probes in a probeset are combined together to get an expression measure, for example, Average Difference (AvgDiff), the Model Based Expression Index (MBEI) of Li and Wong (2001), the MAS 5.0 Statistical algorithm from Affymetrix (2001), and the Robust Multi-chip Average proposed in Irizarry *et al.* (2003).

The need for normalization arises naturally when dealing with experiments involving multiple arrays. There are two broad characterizations that could be used for the type of variation one might expect to see when comparing arrays: interesting variation and obscuring variation. We would classify biological differences, for example large differences in the expression level of particular genes between a diseased and a normal tissue source, as interesting variation. However, observed expression levels also include variation that is introduced during the process of carrying out the experiment, which could be classified as obscuring variation. Examples of this obscuring variation arise due to differences in sample preparation (for instance labeling differences), production of the arrays and the processing of the arrays (for instance scanner differences). The purpose of normalization is to deal with this obscuring variation. A more complete discussion on the sources of this variation can be found in Hartemink *et al.* (2001).

*To whom correspondence should be addressed.

Affymetrix has approached the normalization problem by proposing that intensities should be scaled so that each array has the same average value. The Affymetrix normalization is performed on expression summary values. This approach does not deal particularly well with cases where there are non-linear relationships between arrays. Approaches using non-linear smooth curves have been proposed in Schadt *et al.* (2001, 2002) and Li and Wong (2001). Another approach is to transform the data so that the distribution of probe intensities is the same across a set of arrays. Sidorov *et al.* (2002) propose parametric and non-parametric methods to achieve this. All these approaches depend on the choice of a baseline array.

We propose three different methods of normalizing probe intensity level oligonucleotide data, none of which is dependent on the choice of a baseline array. Normalization is carried out at probe level for all the probes on an array. Typically we do not treat PM and MM separately, but instead consider them all as intensities that need to be normalized. The normalization methods do not account for saturation. We consider this a separate problem to be dealt with in a different manner.

In this paper, we compare the performance of our three proposed complete data methods. These methods are then compared with two methods making use of a baseline array. The first method, which we shall refer to as the scaling method, mimics the Affymetrix approach. The second method, which we call the non-linear method, mimics the approaches of Schadt *et al.* Our assessment of the normalization procedures is based on empirical results demonstrating ability to reduce variance without increasing bias.

NORMALIZATION ALGORITHMS

Complete data methods

The complete data methods combine information from all arrays to form the normalization relation. The first two methods, cyclic loess and contrast, are extensions of accepted normalization methods that have been used successfully with cDNA microarray data. The third method, based on quantiles, is both quicker and simpler than those methods.

Cyclic loess This approach is based upon the idea of the M versus A plot, where M is the difference in log expression values and A is the average of the log expression values, presented in Dudoit *et al.* (2002). However, rather than being applied to two color channels on the same array, as is done in the cDNA case, it is applied to probe intensities from two arrays at a time. An M versus A plot for normalized data should show a point cloud scattered about the $M = 0$ axis.

For any two arrays i, j with probe intensities x_{ki} and x_{kj} where $k = 1, \ldots, p$ represents the probe, we calculate $M_k = \log_2\left(x_{ki}/x_{kj}\right)$ and $A_k = \frac{1}{2}\log_2\left(x_{ki}x_{kj}\right)$. A normalization curve is fitted to this M versus A plot using loess. Loess is a method of local regression (see Cleveland and Devlin 1988 for details). The fits based on the normalization curve are $\hat{M}_k$ and thus the normalization adjustment is $M'_k = M_k - \hat{M}_k$. Adjusted probe intensites are given by $x'_{ki} = 2^{A_k + \frac{M'_K}{2}}$ and $x'_{kj} = 2^{A_K - \frac{M'_k}{2}}$. The preferred method is to compute the normalization curves using rank invariant sets of probes. This paper uses invariants sets since it increases the implementation speed.

To deal with more than two arrays, the method is extended to look at all distinct pairwise combinations. The normalizations are carried out in a pairwise manner as above. We record an adjustment for each of the two arrays in each pair. So after looking at all pairs of arrays for any array k where $1 \leq k \leq n$, we have adjustments for chip k relative to arrays $1, \ldots, k-1, k+1, \ldots, n$. We weight the adjustments equally and apply to the set of arrays. We have found that after only 1 or 2 complete iterations through all pairwise combinations the changes to be applied become small. However, because this method works in a pairwise manner, it is somewhat time consuming.

Contrast based method The contrast based method is another extension of the M versus A method. Full details can be found in Åstrand (2001). The normalization is carried out by placing the data on a log-scale and transforming the basis. In the transformed basis, a series of $n-1$ normalizing curves are fit in a similar manner to the M versus A approach of the cyclic loess method. The data is then adjusted by using a smooth transformation which adjusts the normalization curve so that it lies along the horizontal. Data in the normalized state is obtained by transforming back to the original basis and exponentiating. The contrast based method is faster than the cyclic method. However, the computation of the loess smoothers is still somewhat time consuming.

Quantile normalization The goal of the quantile method is to make the distribution of probe intensities for each array in a set of arrays the same. The method is motivated by the idea that a quantile–quantile plot shows that the distribution of two data vectors is the same if the plot is a straight diagonal line and not the same if it is other than a diagonal line. This concept is extended to n dimensions so that if all n data vectors have the same distribution, then plotting the quantiles in n dimensions gives a straight line along the line given by the unit vector $\left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$. This suggests we could make a set of data have the same distribution if we project the points of our n dimensional quantile plot onto the diagonal.

Let $\boldsymbol{q}_k = (q_{k1}, \ldots, q_{kn})$ for $k = 1, \ldots, p$ be the vector of the kth quantiles for all n arrays $\boldsymbol{q}_k = (q_{k1}, \ldots, q_{kn})$

and $\boldsymbol{d} = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$ be the unit diagonal. To transform from the quantiles so that they all lie along the diagonal, consider the projection of $\boldsymbol{q}$ onto $\boldsymbol{d}$

$$\text{proj}_{\boldsymbol{d}}\boldsymbol{q}_k = \left(\frac{1}{n}\sum_{j=1}^{n} q_{kj}, \ldots, \frac{1}{n}\sum_{j=1}^{n} q_{kj}\right)$$

This implies that we can give each array the same distribution by taking the mean quantile and substituting it as the value of the data item in the original dataset. This motivates the following algorithm for normalizing a set of data vectors by giving them the same distribution:

1. given n arrays of length p, form X of dimension $p \times n$ where each array is a column;
2. sort each column of X to give X_{sort};
3. take the means across rows of X_{sort} and assign this mean to each element in the row to get X'_{sort};
4. get $X_{\text{normalized}}$ by rearranging each column of X'_{sort} to have the same ordering as original X

The quantile normalization method is a specific case of the transformation $x'_i = F^{-1}(G(x_i))$, where we estimate G by the empirical distribution of each array and F using the empirical distribution of the averaged sample quantiles. Extensions of the method could be implemented where F^{-1} and G are more smoothly estimated.

One possible problem with this method is that it forces the values of quantiles to be equal. This would be most problematic in the tails where it is possible that a probe could have the same value across all the arrays. However, in practice, since probeset expression measures are typically computed using the value of multiple probes, we have not found this to be a problem.

Methods using a baseline array

Scaling methods The standard Affymetrix normalization is a scaling method that is carried out on probeset expression measures. To allow consistent comparison with our other methods, we have carried out a similar normalization at the probe level. Our version of this method is to choose a baseline array, in particular, the array having the median of the median intensities. All arrays are then normalized to this 'baseline' via the following method. If x_{base} are the intensities of the baseline array and x_i is any array, then let

$$\beta_i = \frac{\tilde{x}_{\text{base}}}{\tilde{x}_i}$$

where $\tilde{x}_i$ is the trimmed mean intensity (in our analysis we have excluded the highest and lowest 2% of probe intensities). Then the intensities for the normalized array would be

$$x'_i = \beta_i x_i$$

One can also easily implement the scaling algorithm by using probes from a subset of probesets chosen by using some stability criteria. The HG-U133 arrays provide a set of probesets that have been selected for stability across tissue types, and these could be used for establishing a normalization.

Non-linear method The scaling method is equivalent to fitting a linear relationship with zero intercept between the baseline array and each of the arrays to be normalized. This normalizing relation is then used to map from each array to the baseline array. This idea can be extended to use a non-linear relationship to map between each array and the baseline array. Such an approach is detailed in Schadt *et al.* (2002). This method is used in Li and Wong (2001) and implemented in the dChip software http://www.dchip.org. The general approach of these papers is to select a set of approximately rank invariant probes (between the baseline and arrays to be normalized) and fit a non-linear relation, like smoothing splines as in Schadt *et al.* (2002), or a piecewise running median line as in Li and Wong (2001).

The non-linear method used in this paper is as follows. First we select a set of probes for which the ranks are invariant across all the arrays to be normalized. Then we fit loess smoothers to relate the baseline to each of the arrays to be normalized. These loess normalization curves are then used to map probe intensities from the arrays to be normalized to the baseline. This approach is intended to mimic the approach used in dChip. We expect loess smoothers to perform in the same manner as splines or a running median line.

Suppose that $\hat{f}_i(x)$ is the loess smoother mapping from array i to the baseline. Then, in the same notation as above, the normalized array probe intensities are

$$x'_i = \hat{f}_i(x_i)$$

Note that as with the scaling method, the baseline is the array having the median of the median probe intensities.

DATA

We make use of data from two sets of experiments: A dilution/mixture experiment and an experiment using spike-ins. We use these datasets because they allow us to assess bias and variance. The dilution/mixture and spike-in datasets are available directly from GeneLogic (2002) and have been made available for public comparison of analysis methods. This data has been previously described in Irizarry *et al.* (2003).

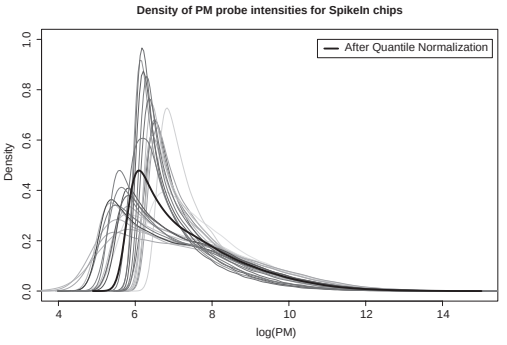

Fig. 1. A plot of the densities for PM for each of the 27 spike-in datasets, with distribution after quantile normalization superimposed.

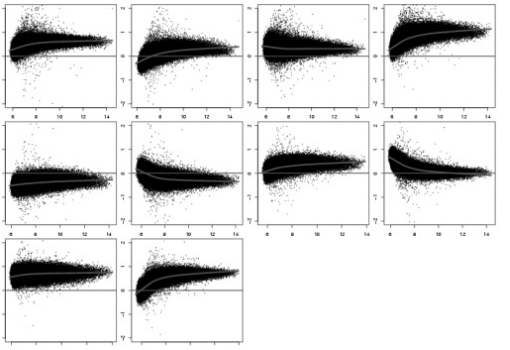

Fig. 2. 10 pairwise M versus A plots using liver (at concentration 10) dilution series data for unadjusted data.

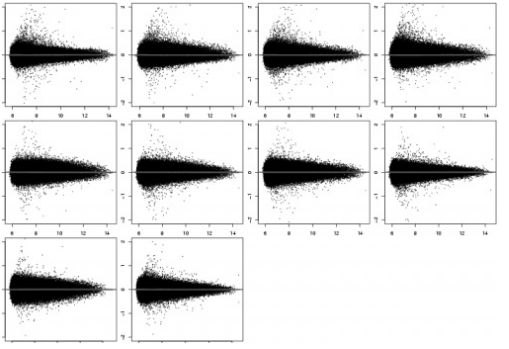

Fig. 3. 10 pairwise M versus A plots using liver (at concentration 10) dilution series data after quantile normalization.

Dilution/mixture data

The dilution/mixture data series consists of 75 HG-U95A (version 2) arrays, where two sources of RNA, liver (source A), and a central nervous system cell line (source B) are investigated. There are 30 arrays for each source, broken into 6 groups at 5 dilution levels. The remaining 15 arrays, broken into 3 groups of 5 chips, involve mixtures of the two tissue lines in the following proportions: 75 : 25, 50 : 50, and 25 : 75.

Spike-in data

The spike-in data series consists of 98 HG-U95A (version 1) arrays where 11 different cRNA fragments have been spiked in at various concentrations. There is a dilution series consisting of 27 arrays which we will examine in this paper. The remaining arrays are two sets of latin square experiments, where in most cases three replicate arrays have been used for each combination of spike-in concentrations. We make use of 6 arrays (two sets of triplicates) from one of the latin squares.

RESULTS

Probe level analysis

Figure 1 plots the densities for the $\log(PM)$ for each of the 27 arrays from the spike-in dataset, along with the distribution obtained after quantile normalization.

An M versus A plot allows us to discern intensity dependent differences between two arrays. Figure 2 shows M versus A plots for unadjusted PM for all 10 possible pairs of 5 arrays in the liver 10 group before normalization. Clear differences between the arrays can be seen by looking at the loess lines. The point clouds are not centered around $M = 0$ and we see non-linear relationships between arrays. The same 10 pairwise comparisons can be seen after quantile normalization in Figure 3. The point clouds are all centered around $M = 0$. Plots produced using the contrast and cyclic loess normalizations are similar.

Expression measures

Comparing normalization methods at the probeset level requires that one must decide on an expression measure. Although in this paper we focus only on one expression measure, the results obtained are similar when using other measures.

The expression summary used in this paper is a robust combination of background adjusted PM intensities and is outlined in Irizarry *et al.* (2003). We call this method the Robust Multichip Average (RMA). RMA estimates

are based upon a robust average of $\log_2(B(PM))$, where $B(PM)$ are background corrected PM intensities. The expression measure may be used on either the natural or log scales.

Irizarry *et al.* (2003) contains a more complete discussion of the RMA measure, and further papers exploring its properties are under preparation.

Probeset measure comparisons

Variance comparisons In the context of the dilution study, consider the five arrays from a single RNA source within a particular dilution level. We calculate expression measures for every probeset on each array and then compute the variance and mean of the probeset expression summary across the five arrays. This is repeated for each group of 5 arrays for the entire dilution/mixture study. We do this after normalization by each of our three complete data methods.

Plotting the log of the ratio of variances versus the average of the log of the mean (expression measure across arrays) allows us to see differences in the between array variations and intensity dependent trends when comparing normalization methods. In this case, the expression measures have all been calculated on the natural scale. Figure 4 shows such plots for the liver at the dilution level 10. Specifically, the four plots compare the variance ratios for quantile : unnormalized, loess : quantile, contrast : quantile and contrast : loess. The horizontal line indicates the x-axis. The other line is a loess smoother. Where the loess smoother is below the x-axis, the first method in the ratio has the smaller ratio and vice versa when the loess smoother is above the line. All three methods reduce the variance at all intensity levels in comparison to data that has not been normalized. The three normalization methods perform in a relatively comparable manner, but the quantile method performs slightly better for this dataset, as can be seen in the loess : quantile and contrast : quantile plots. Similar results are seen in comparable plots (not shown) for the other dilution/mixture groups.

We repeat this analysis with the 27 spike-in arrays, but this time we include the two baseline methods in our comparison. The complete data methods generally leave the mean level of a particular probeset at a level similar to that achieved when using unnormalized data. However, when one of the two baseline methods is used, the mean of a particular probeset is more reminiscent of the value of that probeset in the baseline array. In the natural scale, it is easy to see a mean-variance relationship, where a higher mean implies high variability. Thus, when a comparison is made between the baseline methods and the complete data methods, we find that if a baseline array which shifts the intensities higher (or lower) than the level of those of the unnormalized means is selected, then the corresponding variance of the probeset measures across arrays is higher

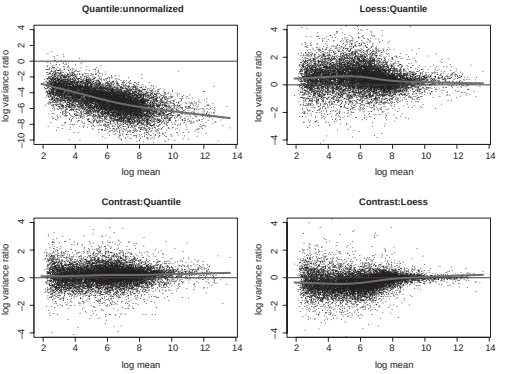

Fig. 4. $\log_2$ variance ratio versus average $\log_2$ mean for liver dilution data at concentration 10.

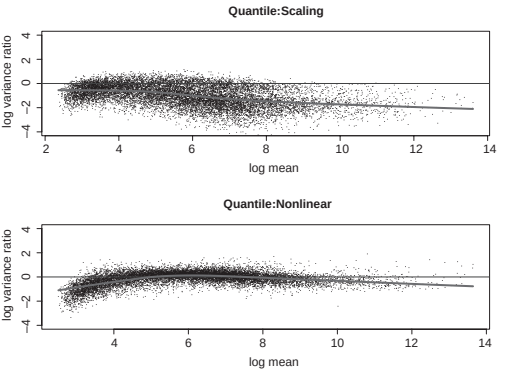

Fig. 5. $\log_2$ variance ratio versus average $\log_2$ mean using the spike-in data. Comparing the baseline methods with the quantile method.

(or lower) due to the shifting and not because of the normalization. To minimize this problem and make a fairer comparison, we work with the expression measure on the log scale when comparing the baseline methods to the complete data methods. Figure 5 compares the baseline methods to the quantile methods. We see that the quantile method reduces the between array variances more than the scaling method. The non-linear normalization performs a great deal closer to the quantile method. Similar plots (not shown) comparing the complete data methods with each other for the spike-in data demonstrate that quantile normalization has a slight edge over all the other methods.

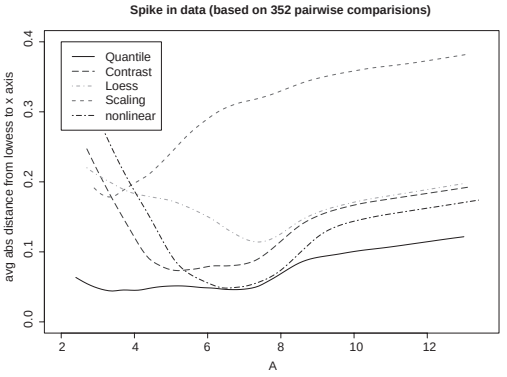

Fig. 6. Comparing the ability of methods to reduce pairwise differences between arrays by using average absolute distance from loess smoother to x-axis in pairwise M versus A plots using spike-in dataset. Smaller distances are favorable.

A similar plot (not shown) comparing the two baseline methods shows as expected that the non-linear method reduces variance when compared to the scaling method.

Pairwise comparison The ability to minimize differences in pairwise comparisons between arrays is a desirable feature of a normalization procedure. An M versus A plot comparing expression measures on two arrays should be centered around $M = 0$ if there is no clear trend towards one of the arrays. Looking at the absolute distance between a loess for the M versus A plot and the x-axis allows us to assess the difference in array to array comparisons. We can compare methods by looking at this distance across a range of intensities and averaging the distance across all pairwise comparisions.

Figure 6 shows such a plot for the spike-in data, we see that the scaling method performs quite poorly when compared to the three complete data methods. The non-linear method performs at a similar level to the complete data methods. For this dataset the quantile method is slightly better. An important property of the quantile method is that these differences remain relatively constant across intensities.

Bias comparisons One way to look at bias is in the context of the spike-in dilution series. We use data for the 27 arrays from the spike-in experiment with 11 control fragments spiked in at 13 different concentrations (0.00, 0.50, 0.75, 1.00, 1.50, 2.00, 3.00, 5.00, 12.50, 25.00, 50.00, 75.00, 100.00, 150.00 pM). We normalize each of the 27 arrays as a group using each of the quantile, contrast, cyclic loess and scaling normalizations. To the spike-in probesets, we fit the following linear model

$$\log_2 E = \beta_0 + \beta_1 \log_2 c + \epsilon$$

where E is the value of the expression measure and c are the concentrations. Note that the array with spike-in concentration 0 is excluded from the model fit, although it is used in the normalization. The ideal results would be to have slopes that are near 1. Table 1 shows the slope estimates for each of the spike-in probesets after normalization by each of the three complete data methods, the two methods using a baseline and when no normalization has taken place. For the three complete data methods for 10 out of the 11 spike-in probesets, the quantile method gives a slope closer to 1 and the non-linear method has slopes lower than the complete data methods. However, both the scaling and not normalizing have slopes closer to 1. For the non-spike-in probesets on these arrays we should see no linear relation if we fit the same model, since there should be no relation between the spike-in concentrations and the probeset measures. Fitting the linear model above, we find that there is a median slope of 0.042 for these probesets using the unnormalized data. For the quantile method the value of the median slope is -0.005. All the other normalization methods have median slope near 0. This is about the same difference in slopes as we observed for the spike-ins when comparing the unnormalized data and the best of the normalization methods. In other words, there is a systematic trend due to the manner in which the arrays were produced that has resulted in the intensities of all the probesets being related to the concentration of the spike-ins. We should adjust the spike-in slopes by these amounts. For example, we could adjust the slope of BioB-5 for the quantile method to $0.845 + 0.005 = 0.850$ and the unnormalized slope to $0.893 - 0.042 = 0.851$.

The average R^2 for the spike-in probesets, excluding CreX-3, are 0.87 for the quantile method, and 0.855, 0.849, 0.857 and 0.859 for the contrast, loess, non-linear and scaling methods, respectively. It was 0.831 for the unnormalized data. The median standard error for the slopes was 0.063 for the quantile method. For the other methods these standard errors were 0.065 (contrast), 0.068 (cyclic loess), 0.063 (non-linear), 0.065 (scaling) and 0.076 (unnormalized). Thus of all the algorithms, the quantile method has high slopes, a better fitting model and more precise slope estimates.

The slopes may not reach 1 for several reasons. It is possible that there is a 'pipette' effect. In other words we can not be completely sure of the concentrations. It is more likely that we observe concentration plus an error which leads to a downward bias in the slope estimates. Other possible reasons include the saturation of signal at the high end (this is not a concern with this data) and having a higher background effect at the lower end.

Table 1. Regression slope estimates for spike-in probesets. A slope closer to one is better

Name	Quantile	Contrast	Loess	Non-linear	Scaling	None
AFFX-BioB-5_at	0.845	0.837	0.834	0.803	0.850	0.893
AFFX-DapX-M_at	0.778	0.771	0.770	0.746	0.783	0.826
AFFX-DapX-5_at	0.754	0.747	0.728	0.731	0.764	0.807
AFFX-CreX-5_at	0.903	0.897	0.889	0.875	0.912	0.955
AFFX-BioB-3_at	0.836	0.834	0.825	0.807	0.848	0.890
AFFX-BioB-M_at	0.789	0.782	0.781	0.762	0.797	0.838
AFFX-BioDn-3_at	0.547	0.543	0.550	0.514	0.553	0.595
AFFX-BioC-5_at	0.801	0.794	0.793	0.763	0.808	0.851
AFFX-BioC-3_at	0.796	0.790	0.785	0.769	0.805	0.847
AFFX-DapX-3_at	0.812	0.804	0.793	0.776	0.815	0.859
AFFX-CreX-3_at	−0.007	−0.006	0.002	−0.007	0.005	0.046
Non-spike-in (median)	−0.005	−0.005	−0.005	−0.007	−0.001	0.042

The problems with choosing a baseline The non-linear (and scaling) method requires the choice of a baseline. In this paper we have chosen the array having the median median, but other options are certainly possible. To address these concerns we examine a set of six arrays chosen from the spike-in datasets. In particular, we choose two sets of triplicates, where the fold change of each of the spike-in probesets between the two triplicates is large. The two triplicates are chosen so that about half of the probesets are high in one triplicate and low in the other and vice versa.

We normalize this dataset using both the quantile and non-linear methods. However, for the non-linear method we experiment with the use of each of the six arrays as the baseline array. We also try using two synthetic baseline arrays: one constructed by taking probewise means and one taking probewise medians. Figure 7 shows the distribution of the mean of the probeset measure across arrays. We see that the quantile normalization produces a set of means that is very similar in distribution to the means of the unnormalized data. The means from the non-linear normalizations using each of the six different baseline arrays are quite different from each other and the unnormalized data. This is somewhat of a drawback to the baseline methods. It seems more representative of the complete data to consider all arrays in the normalization rather than to use only a single baseline and give the normalized data characteristics closer to those of one particular array. Only the mean based synthetic baseline array comes close to the unnormalized and quantile methods.

Table 2 summarizes some results from this analysis. We see that all the methods reduce the variability of the probeset measure between arrays compared to that of the unnormalized data. In each case, around 95% of the probesets have reduced variance. When compared to the quantile method, it comes out more even with a little over 50% of probesets having reduced variance for four of

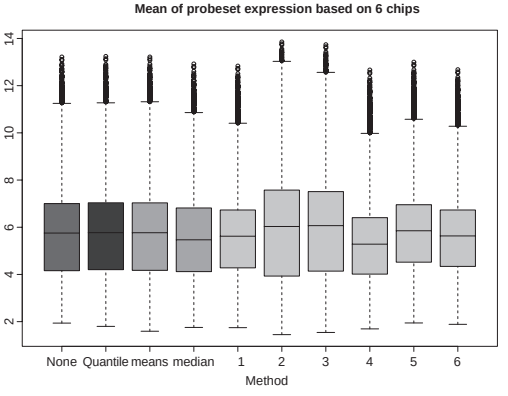

Fig. 7. Distribution of average (over 6 chips) of a probeset expression measure using different baseline normalizations.

the baselines. However, two baseline arrays perform quite poorly. As noted before, this is a reflection of the baseline methods shifting the intensities higher or lower depending on the baseline. The mean based synthetic baseline does not reduce the variabilty of the probeset measure to the same degree as the quantile method.

Looking at the 11 spike-in probesets, we calculate bias by taking the difference between the log of the ratio between spike-in concentrations and the log of the ratio of intensities in the two groups. One spike-in, Crex-3, did not seem to perform quite as well as the other spike-ins and was excluded from the analysis. Looking at the total absolute bias across the 10 spike-in probesets, we see that the non-linear method has lower total bias (compared to the quantile method) for four of the methods, but two

Table 2. Comparing variance and bias with the non-linear normalization when using different baselines

Method	% with lower var reduced cf. U	% lower var reduced cf. Q	Abs Bias	# abs Bias cf U	# abs Bias cf Q
Probewise mean	83	40	9.2	5	5
Probewise median	96	58	7.9	6	6
Non-linear 1	96	53	7.5	7	5
Non-linear 2	93	31	11.8	2	4
Non-linear 3	94	37	10.5	4	4
Non-linear 4	95	47	7.4	6	5
Non-linear 5	96	55	7.4	7	5
Non-linear 6	96	55	7.5	7	5
Quantile (Q)	95	NA	8.5	6	NA
Unnormalized (U)	NA	NA	9.7	NA	NA

are even bigger than for unnormalized data. Again, this is related to the mean–variance relationship. The four baselines shifted slightly lower in the intensity scale give the most precise estimates. Using this logic, one could argue that choosing the array with the smallest spread and centered at the lowest level would be the best, but this does not seem to treat the data on all arrays fairly. Compared to the unnormalized data, 6 of the spike-in probesets from the quantile normalized data have a smaller bias. For the non-linear normalization using array 1 as the baseline (this is the array chosen using our heuristic), 7 had smaller bias. However, looking at the other baselines, anywhere from 2 to 7 probesets had lower bias. When compared to the quantile method, the results are more even, with about an equal number of the spike-in probesets having a lower bias when using the non-linear method as when using the quantile normalization. An *M* versus *A* plot between the two groups shows all the spike-in points clearly outside the point cloud, no matter which normalization is used. This plot for quantile normalized data is shown in Figure 8.

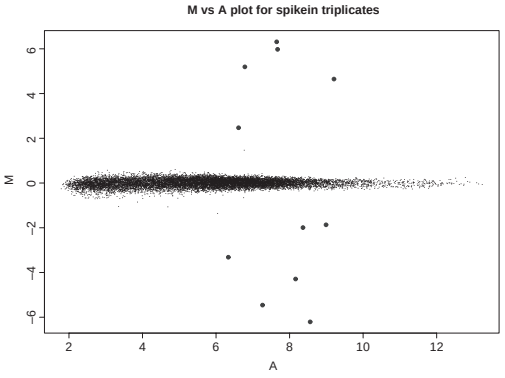

Fig. 8. *M* versus *A* plot for spike-in triplicate data normalized using quantile normalization. Spike-ins are clearly identified.

CONCLUSIONS

We have presented three complete data methods of normalization and compared these to two different methods that make use of a baseline array. Using two different datasets, we established that all three of the complete data methods reduced the variation of a probeset measure across a set of arrays to a greater degree than the scaling method and unnormalized data. The non-linear method seemed to perform at a level similar to the complete data methods. Our three complete data methods, while different, performed comparably at reducing variability across arrays.

When making pairwise comparisons the quantile method gave the smallest distance between arrays. These distances also remained fairly constant across intensities.

In relation to bias, all three complete data methods performed comparably, with perhaps a slight advantage to the quantile normalization. The non-linear method did poorer for the spike-in regressions. The scaling method had slightly higher slopes. Even so, they were more variable.

We saw that the choice of a baseline does have ramifications on down-stream analysis. Choosing a poor baseline would conceivably give poorer results. We also saw that the complete data methods perform well at both variance reduction and on the matter of bias, and in addition more fully reflect the complete set of data. For this reason we favor a complete data method.

In terms of speed, for the three complete data methods, the quantile method is the fastest. The contrast method is slower and the cyclic loess method is the most time consuming. The contrast and cyclic loess algorithms are modifications of an accepted method of normalization.

The quantile method has performed favorably, both in terms of speed and when using our variance and bias criteria, and therefore should be used in preference to the other methods.

While there might be some advantages to using a common, non-data driven, distribution with the quantile method, it seems unlikely an agreed standard could be reached. Different choices of a standard distribution might be reflected in different estimated fold changes. For this reason we prefer the minimalist approach of a data based normalization.

REFERENCES

Affymetrix (2001) Statistical algorithms reference guide, Technical report, Affymetrix.

Åstrand,M. (2001) Normalizing oligonucleotide arrays. *Unpublished Manuscript*. http://www.math.chalmers.se/~magnusaa/maffy.pdf

Bolstad,B. (2001) Probe level quantile normalization of high density oligonucleotide array data. *Unpublished Manuscript*. http://www.stat.berkeley.edu/~bolstad/

Cleveland,W.S. and Devlin,S.J. (1998) Locally-weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.*, **83**, 596–610.

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.*, **12(1)**, 111–139.

GeneLogic (2002) Datasets http://www.genelogic.com.

Hartemink,A., Gifford,D, Jaakkola,T. and Young,R. (2001) Maximum likelihood estimation of optimal scaling factors for expression array normalization. In *SPIE BIOS 2001*.

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5(3)**, 299–314.

Irizarry,R.A., Hobbs,B., Colin, F., Beazer-Barclay,Y.D., Antonellis,K., Scherf,U. and Speed,T.P. (2003) Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*. in press.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error applications. *Genome Biol.*, **2(8)**, 1–11.

Lipshutz,R., Fodor,S., Gingeras,T. and Lockart,D. (1999) High density synthetic olignonucleotide arrays. *Nature Genet.*, **21**(Suppl), 20–24.

Schadt,E., Li,C., Su,C. and Wong,W.H. (2001) Analyzing high-density oligonucleotide gene expression array data. *J. Cell. Biochem.*, **80**, 192–202.

Schadt,E., Li,C., Eliss,B. and Wong,W.H. (2002) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem.*, **84(S37)**, 120–125.

Sidorov,I.A., Hosack,D.A., Gee,D., Yang,J., Cam,M.C., Lempicki,R.A. and Dimitrov,D.S. (2002) Oligonucleotide microarray data distribution and normalization. *Information Sciences*, **146**, 65–71.

Venables,W. and Ripley,B.D. (1997) *Modern Applied Statistics with S-PLUS*, Second edn, Springer, New York.

Warrington,J.A., Dee,S. and Trulson,M. (2000) Large-scale genomic analysis using Affymetrix GeneChip®. In Schena,M. (ed.), *Microarray Biochip Technology*. BioTechniques Books, New York, **Chapter 6**, pp. 119–148.

Paper III

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 14, Number 10, 2007

Pp. 1353–1367
DOI: 10.1089/cmb.2007.0078

Improved Covariance Matrix Estimators for Weighted Analysis of Microarray Data

MAGNUS ÅSTRAND, PETTER MOSTAD, and MATS RUDEMO

ABSTRACT

Empirical Bayes models have been shown to be powerful tools for identifying differentially expressed genes from gene expression microarray data. An example is the WAME model, where a global covariance matrix accounts for array-to-array correlations as well as differing variances between arrays. However, the existing method for estimating the covariance matrix is very computationally intensive and the estimator is biased when data contains many regulated genes. In this paper, two new methods for estimating the covariance matrix are proposed. The first method is a direct application of the EM algorithm for fitting the multivariate t-distribution of the WAME model. In the second method, a prior distribution for the log fold-change is added to the WAME model, and a discrete approximation is used for this prior. Both methods are evaluated using simulated and real data. The first method shows equal performance compared to the existing method in terms of bias and variability, but is superior in terms of computer time. For large data sets (>15 arrays), the second method also shows superior computer run time. Moreover, for simulated data with regulated genes the second method greatly reduces the bias. With the proposed methods it is possible to apply the WAME model to large data sets with reasonable computer run times. The second method shows a small bias for simulated data, but appears to have a larger bias for real data with many regulated genes.

Key words: microarray data, gene expression, moderated analysis, weighted analysis.

1. INTRODUCTION

GENE EXPRESSION MICROARRAY TECHNOLOGIES measure the abundance of mRNA sequences in biological samples. Simultaneously, measurements for thousands of genes are obtained for each sample. Before acquiring the actual measurements, the samples are subject to several laboratory steps such as isolation of total RNA, transcription to complementary RNA, and fragmentation. After the wet laboratory procedures, the raw image intensity files are pre-processed, which typically involves background correction and normalization.

Department of Mathematical Sciences, Chalmers University of Technology and Göteborg University, Göteborg, Sweden.

When obtaining the biological samples (e.g., taking biopsies) and during the laboratory work, many sources of variations are introduced that may affect the final measurements. The pre-processing can reduce variability, but it can also introduce new sources of variation when the assumptions made do not hold. For instance, practically all normalization methods rely on an assumption of a rough balance between up- and down-regulation. The quality of the manufactured arrays is also known to vary. Altogether this means that data quality within an experiment often varies across the arrays.

Moreover, when sources of variations are shared between arrays, we can expect correlations between the measurements. For example, as mentioned by Kristiansson et al. (2006), mRNA may be more or less degraded, and when biopsies contain several cell types, the mixture proportions can differ. When this is the case, we can expect arrays with equal mRNA degradation or with similar mixture proportions to show more similar gene profiles compared to arrays with different mRNA degradation and cell mixture proportions.

To accommodate these differences in data quality (i.e., differences in variances and the possibility of correlations between arrays), the empirical Bayes WAME model was introduced in Kristiansson et al. (2005), and further developed in Kristiansson et al. (2006) and Sjögren et al. (2007). WAME makes use of a global covariance matrix common to all genes. The covariance matrix is scaled between genes by a gene-specific parameter assumed to be inverse-gamma distributed. Another empirical Bayes approach taking the quality issue into account is weighted LIMMA (Ritchie et al., 2006). However, here only differences in variances across arrays are considered; correlations between arrays are assumed to be zero.

Several other empirical Bayes and fully Bayesian procedures are found in the literature. Baldi and Long (2001) derive a posterior probability of regulation for each gene based on two posterior models, whereas Lönnstedt and Speed (2002) compute the posterior odds for differential expression. In Smyth (2004), a moderated t-statistic is used and shown to be a monotonic function of the posterior odds. Closely related to the moderated t-statistic are penalized t-tests such as the SAM method in Tusher et al. (2001).

In Broët et al. (2002), a fully Bayesian model is applied directly to the observed mean differences between two conditions. For each gene, a posterior distribution across an unknown number of regulation levels is derived. Lönnstedt and Britton (2005) explore several fully Bayesian models, of which one shows good fit to the data sets examined. However, in terms of accurate ranking of genes, the empirical Bayes models perform at least as good the fully Bayesian model.

A Bayesian model incorporating array effects (i.e., normalization) is suggested by Lewin et al. (2006). Based on the joint posterior distribution a gene-selection procedure using multiple criteria is proposed. A fully Bayesian model, specifically designed for Affymetrix types of arrays, which does not require replicates is used by Hein and Richardson (2006).

The current version of WAME estimates the global covariance matrix under the temporary assumption of no regulated genes. Once the covariance matrix estimator is obtained the assumption is relaxed and the remaining model parameters are estimated. However, due to the temporary assumption, the covariance matrix estimator is biased when regulated genes exists. Also, the computational procedure used is very computational intensive resulting in long computer run times. The aim of this paper is to eliminate or reduce these two drawbacks. We recapitulate the results of Kristiansson et al. (2006) and Sjögren et al. (2007) describing the WAME model for analyzing designed microarray experiments. Then an alternative computational procedure (method I) is introduced which considerably reduces the computational time required for estimating the global covariance matrix. Secondly, the WAME model is expanded with a prior distribution for the log fold-change and an estimation procedure (method II) relaxing the assumption of no regulation is described. These two methods are compared with the current WAME procedure on simulated and real data.

2. DESIGNED MICROARRAY EXPERIMENTS

Let y_{ig} be the log-scale gene expression measurement for gene g on array i, for a total of G genes and p arrays, and put $y_g = (y_{1g}, \ldots, y_{pg})^T$. For a designed microarray experiment the vector $\mu_g = \mathrm{E}[y_g]$, the log-intensity profile for gene g across the arrays, is determined by a full rank $p \times k$ design matrix D through $\mu_g = D\gamma_g$ where γ_g is a gene-specific parameter vector of length k. The aim is to estimate $C\gamma_g$ and thus the matrix C specifies the linear combinations of γ_g that are of interest.

As an example, consider a one-channel experiment comparing two conditions with two replicates for each condition. With condition one on array 1 and 3, and condition two on array 2 and 4, the design matrix D_1 below can be used (the design matrix is generally not uniquely determined). Put $\gamma_g = (\gamma_{g1}, \gamma_{g2})^T$, where γ_{g1} denotes the mean intensity for gene g, and γ_{g2} is half the intensity difference between the two conditions. For a two-color dye-swapped spotted array experiment with four arrays and condition one colored green on array 1 and 3 the design matrix D_2 below would be a natural choice when analyzing the logged R/G ratio. The parameter vector consists of one element with the same interpretation as γ_{g2} in the first example.

$$D_1 = \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix} \quad D_2 = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}$$

For estimating the logged fold-change, we use $C = [0, 2]$ and $C = [2]$ in the first and second example, respectively.

3. WEIGHTED ANALYSIS

The model suggested by Kristiansson et al. (2005) uses a global covariance structure and a gene-specific scaling factor to model dependency between arrays. Formally, for $g = 1, \ldots, G$ let

$$\begin{aligned} y_g | c_g &\sim \mathrm{N}_p(\mu_g, c_g \Sigma) \\ c_g &\sim \Gamma^{-1}(m/2, m\nu/2). \end{aligned} \tag{1}$$

Here $\mathrm{N}_p(\mu, \Sigma)$ denotes a p-dimensional normal distribution with mean μ and covariance matrix Σ and $\Gamma^{-1}(\alpha, \beta)$ is the inverse-gamma distribution with density function $f(x) = \beta^\alpha \exp\{-\frac{\beta}{x}\} x^{-(\alpha+1)} / \Gamma(\alpha)$ for $x > 0$. As described the model is not identifiable. The likelihood evaluated at (Σ, ν, m) equals the likelihood evaluated at $(\delta^{-1}\Sigma, \delta\nu, m)$ for any $\delta > 0$. To get an identifiable model, we can use a restriction on Σ, e.g., $\mathrm{trace}(\Sigma) = p$, or a restriction on ν, say $\nu = 1$. Note that Kristiansson et al. (2005) uses another parameterization ($m = 2\alpha$, $\nu = 1/\alpha$) with no restriction on Σ.

The method for estimating the covariance matrix Σ of model (1) suggested by Kristiansson et al. (2005) only works for designs where no regulation implies $\mu_g = 0$. This typically holds for the dye-swap example above. But not for the first example with design matrix D_1 since the first parameter γ_{g1} is the mean log-intensity which is strictly positive for most genes. However, as shown by Sjögren et al. (2007), such data can be transformed by subtracting the vector $D\hat{\gamma}_g^0$ from y_g, where $\hat{\gamma}_g^0$ is a suitable estimator for γ_g under the assumption of no regulation. For unpaired designs comparing two conditions, such as our first example, this means subtracting the mean intensity across the arrays for each gene, that is

$$y_{ig} \to y_{ig} - \bar{y}_{\cdot g},\ i = 1, \ldots, p, \quad \text{where} \quad \bar{y}_{\cdot g} = \frac{1}{p} \sum_{j=1}^{p} y_{jg}.$$

The design matrix is transformed accordingly by subtracting the column mean from each column. This means that the first column is a vector of zeros and can be omitted, while the second column is unchanged. Hence, for the transformed data we can use the design matrix D_2 of the second example.

Given an estimator of Σ together with estimates of m and ν these are treated as known, and weighted moderated F- and t-tests are derived. When C consists of one row only, the weighted moderated t-tests for *the* linear combination in C is based on the unbiased minimum variance estimator of $C\gamma_g$:

$$\widehat{C\gamma_g} = C(D^T \Sigma^{-1} D)^{-1} D^T \Sigma^{-1} y_g. \tag{2}$$

Under H_0: $C\gamma_g = 0$, the weighted moderated t-statistic

$$\tilde{t} = \sqrt{\frac{p-k+m}{C(D^T\Sigma^{-1}D)^{-1}C^T}}\frac{\widehat{C\gamma_g}}{\sqrt{m\nu + \mathrm{RSS}_g}} \tag{3}$$

is t-distributed with $p-k+m$ degrees of freedom. Here

$$\mathrm{RSS}_g = y_g^T\left(\Sigma^{-1} - \Sigma^{-1}D(D^T\Sigma^{-1}D)^{-1}D^T\Sigma^{-1}\right)y_g \tag{4}$$

is the weighted residual sum of squares. See Kristiansson et al. (2006) for details.

4. ESTIMATING Σ, m, AND ν: METHOD I

In this section, an alternative method, referred to as method I, for estimation of Σ, m and ν is described. As for the method used by Kristiansson et al. (2005), it is based on the temporary assumption of no regulated genes. When μ_g is set to zero for all genes model (1) describes a multivariate t-distribution with zero mean and m degrees of freedom. For the purpose of fitting a multivariate t-distribution, the EM algorithm (Dempster et al., 1977) and extensions of it have been shown to be powerful tools. Many of the applications deal with missing data, and the situation with unknown degrees of freedom was considered by Lange et al. (1989). In comparison, our situation is very simple. The degrees of freedom is still unknown, but the mean vector is zero and there are no missing data.

4.1. Estimation using the EM-algorithm

With the EM algorithm a current approximation, the Q function, of the real log-likelihood function, is found in the E-step and maximized in the M-step. The observed data is augmented with missing or unobserved data into a complete data set. The Q function is then defined as the expectation of the log-likelihood function of the complete data set. The expectation is with respect to the conditional distribution of the missing or unobserved data, given the observed data and the current parameter estimate. The M step results in an updated parameter which is again feed into the E and M step, resulting in another estimate, and so on. When the change in the parameter estimate is small enough the iterations are stopped.

In our situation the gene-specific scaling factors, the c_gs, are the missing data. We use the restriction $\nu = 1$, and we assume independence between genes. Under model (1), with μ_g equal to zero and ν equal to 1, the joint density function of (y_g, c_g) is

$$p(y_g, c_g|\Sigma, m) = p_{N_p}(y_g|0, c_g\Sigma)p_{\Gamma^{-1}}(c_g|m/2, m/2)$$

where p_{N_p} and $p_{\Gamma^{-1}}$ are the density function of the multivariate normal and inverse gamma distributions respectively. Hence, with $T_g = (y_g, c_g)$ and $\theta = (\Sigma, m)$, the contribution from T_g to the complete log-likelihood of $T_1, \ldots, T_G$ is (up to a constant)

$$\begin{aligned}\mathcal{L}_g(\theta, T_g) = &-\frac{1}{2}\log(|\Sigma|) - \frac{y_g^T\Sigma^{-1}y_g}{2c_g} - \frac{m+p+2}{2}\log(c_g)\\ &- \frac{m}{2c_g} + \frac{m}{2}\log\left(\frac{m}{2}\right) - \log(\Gamma(m/2)).\end{aligned} \tag{5}$$

The E-step consists of evaluating the conditional expectation of the complete log-likelihood. Under the assumption of independence between genes, we can treat each gene individually. Hence, we need to evaluate the conditional expectation of $\mathcal{L}_g(\theta, T_g)$ with respect to the distribution of $T_g|y_g$ governed by the parameter θ_0. That is, the distribution of $c_g|y_g, \theta_0$ which is $\Gamma^{-1}(m^*/2, m^*\nu^*/2)$ where

$$m^* = m_0 + p \quad \text{and} \quad \nu^* = \frac{y_g^T\Sigma_0^{-1}y_g + m_0}{m_0 + p}.$$

For a random variable $x \sim \Gamma^{-1}(\alpha, \beta)$ we have $\mathrm{E}[\log(x)] = \log(\beta) - \psi(\alpha)$ and $\mathrm{E}[x^{-1}] = \alpha/\beta$ from the properties of the log-gamma and gamma distribution, respectively (Johnson et al., 1995), where ψ is the digamma function. So the E-step (up to a constant) results in a sum across all genes where the contribution from gene g is

$$\begin{aligned} Q_g(\theta, \theta_0) &= \mathrm{E}[\mathcal{L}_g(\theta, T_g)|y_g, \theta_0] \\ &= -\frac{1}{2}\log(|\Sigma|) - \frac{y_g^T \Sigma^{-1} y_g}{2} \cdot \frac{m_0 + p}{y_g^T \Sigma_0^{-1} y_g + m_0} \\ &\quad + \frac{m}{2}\log\left(\frac{m}{2}\right) - \log(\Gamma(m/2)) - \frac{m}{2} \cdot \frac{m_0 + p}{y_g^T \Sigma_0^{-1} y_g + m_0} \\ &\quad - \frac{m+2}{2}\left(\log\left(\frac{y_g^T \Sigma_0^{-1} y_g + m_0}{2}\right) - \psi\left(\frac{m_0 + p}{2}\right)\right). \end{aligned} \tag{6}$$

In the M step, the Q function, $Q(\theta, \theta_0)$, equal to the sum of $Q_g(\theta, \theta_0)$ across all genes g, is to be maximized with respect to $\theta = (\Sigma, m)$. The part of (6) depending on Σ can be written as

$$-\frac{1}{2}\log(|\Sigma|) - \frac{z_g^T \Sigma^{-1} z_g}{2} \quad \text{where} \quad z_g = y_g \cdot \sqrt{\frac{m_0 + p}{y_g^T \Sigma_0^{-1} y_g + m_0}}.$$

Hence, the value of Σ maximizing the Q-function is equal to the ordinary maximum likelihood estimator for the multivariate normal distribution with known zero mean, given data $z_1, \ldots, z_G$:

$$\hat{\Sigma} = \frac{1}{G}\sum_{g=1}^{G} z_g z_g^T.$$

The value of m maximizing Q is found using numerical optimization of the function f:

$$f(m) = \frac{m}{2}\Big(\log(m) - \log(2) - S_1\Big) - \log\left(\Gamma(m/2)\right)$$

where

$$S_1 = \frac{1}{G}\sum_{g=1}^{G}\left[\log\left(\frac{y_g^T \Sigma_0^{-1} y_g + m_0}{2}\right) - \psi\left(\frac{m_0 + p}{2}\right) + \frac{m_0 + p}{y_g^T \Sigma_0^{-1} y_g + m_0}\right].$$

This finalizes the updating of m and Σ. It remains to set start values. We do this assuming independence and equal variances: $\Sigma = \alpha I_p$ for $\alpha > 0$ and I_p equals the identity matrix. With $\mathrm{MSS}_g = y_g^T y_g / p$ we get

$$\begin{aligned} \mathrm{E}[\log(\mathrm{MSS}_g)] &= \log(m\alpha/p) + \psi(p/2) - \psi(m/2) \quad \text{and} \\ \mathrm{var}(\log(\mathrm{MSS}_g)) &= \psi'(m/2) + \psi'(p/2). \end{aligned}$$

Here ψ and ψ' are the digamma and trigamma functions, respectively. Start values m and $\Sigma = \alpha I_p$ are found by setting the sample mean and squared sample standard deviation of MSS_g across all genes equal to the theoretical ones above.

The method just described uses the restriction $\nu = 1$. However, later on we will use the restriction $\mathrm{trace}(\Sigma) = p$. That is, the obtained estimates (under $\nu = 1$) are transformed according to

$$\hat{\nu} \to \mathrm{trace}(\hat{\Sigma})/p \quad \text{and} \quad \hat{\Sigma} \to p\hat{\Sigma}/\mathrm{trace}(\hat{\Sigma}).$$

5. ESIMATING Σ, m, AND ν: METHOD II

In this section, model (1) is extended with a prior distribution for μ_g. Before specifying the model we define a linear transformation of the vector y_g. Assume D, C, and γ_g are as described in Section 2 and suppose the matrix C has one row only. Thus, with $\mu_g = D\gamma_g$ we wish to estimate $\delta_g = C\gamma_g$ or test the hypothesis H_0: $\delta_g = 0$. For doing this we find full rank matrices A and B whose columns define linear combinations of y_g having expectation zero and δ_g, respectively. From D and C form the $p \times p$ and $p \times 1$ matrices

$$A_0 = I - D(D^T D)^{-1} D^T \quad \text{and} \quad B = D(D^T D)^{-1} C^T.$$

Since A_0 is of rank $p-k$ only, we let A be a $p \times (p-k)$ matrix whose columns span the same subset of $\mathbb{R}^p$ as the columns of A_0, for example derived by iteratively removing linearly dependent columns from A_0. With $\mathrm{E}[y_g] = \mu_g = D\gamma_g$ some algebra gives

$$\mathrm{E}[A^T y_g] = 0 \quad \text{and} \quad \mathrm{E}[B^T y_g] = C\gamma_g = \delta_g.$$

All other linear combinations of y_g, that are linearly independent of the ones in A and B, measure other aspects of γ_g. Hence, for the purpose of estimating δ_g, or testing the hypothesis H_0 it is natural to only use the linear combinations of y_g determined by the columns of A and B. The legitimacy of this statement can also be shown using the principle of invariance under a group of linear transformations. Given the group of transformations it follows that only test-statistics depending on a maximal invariant needs to be considered. In our case, the linear combinations determined by A and B form such a maximal invariant. For details, see chapter 6 of Lehmann (1986).

Now, let x_g be the $q = p - k + 1$ sized vector of *transformed* log-intensities defined in terms of the $p \times q$ transformation matrix P:

$$x_g = P^T y_g \quad \text{where} \quad P = [A; B]. \tag{7}$$

Then $\mathrm{E}[x_g] = (0, \ldots, 0, \delta_g)^T$ and for vectors with such a mean structure, we specify the following model. For $g = 1, \ldots, G$ let

$$\begin{aligned} x_g | c_g &\sim \mathrm{N}_q(\mu_g, c_g \Sigma) \\ c_g &\sim \Gamma^{-1}(m/2, m\nu/2) \\ \mu_g &= (0, \ldots, 0, \delta_g)^T \\ \delta_g &\sim \begin{cases} \equiv 0 & \text{with prob. } \psi_0 \\ F(\beta) & \text{with prob. } 1 - \psi_0. \end{cases} \end{aligned} \tag{8}$$

The difference from model (1) is the structure of μ_g and the distributional assumption on δ_g. The continuous distribution $F(\beta)$ depends on a parameter vector β with density function $f(x|\beta)$. As for model (1) a restriction on ν or Σ is needed for identifiability. The covariance matrix Σ can be divided according to

$$\Sigma = \begin{bmatrix} \Sigma_A & \Sigma_{AB} \\ \Sigma_{AB}^T & \Sigma_B \end{bmatrix}$$

where Σ_A is the covariance matrix for all dimensions but the last, Σ_B is just the variance of the last dimension and Σ_{AB} consists of the covariances between the last dimension and all other dimensions of x_g.

Model (8) is fitted in two steps. First, the last dimension of x is dropped. That is, we only use the linear combinations of y_g for which the design implies zero mean. Hence, the assumption of zero mean used in method I is fulfilled and the method can be applied without risk of introducing bias into the estimates of m, ν and Σ_A. The hyperparameters m and ν are then treated as known and equal to the estimated values. In step 2 estimates of Σ and β are computed under the restriction $\mathrm{trace}(\Sigma_A) = q - 1$ which was also used in step 1. Thus the only information carried over from step 1 about Σ_A is the scale. It is also possible to

derive estimates of the submatrices Σ_{AB} and Σ_B by holding Σ_A fixed but we find it easier to re-estimate the complete covariance matrix Σ. The model is fitted by replacing the continuous distribution $F(\beta)$ by a discrete version $\tilde{F}(\beta)$ with equally spaced support points $\alpha_1, \ldots, \alpha_r$. Thus, the prior distribution of δ is discrete on $\{0, \alpha_1, \alpha_2, \ldots, \alpha_r\}$. Throughout this article we have used $r = 100$, which generally appears to be large enough. The support points are set so that the range covers 99.9% of the observed data. As for method I, we will use the EM algorithm, treating the c_g's and δ_g's as missing data. The procedure is described below.

5.1. *Estimating Σ using a discrete prior for δ*

With $\alpha_0 = 0$ we have $P(\delta = \alpha_j) = \psi_j$ for $j = 0, \ldots, r$. Thus the joint density function of (x_g, c_g, δ_g) is

$$p(x_g, c_g, \delta_g = \alpha_j | \Sigma, m, \nu, \psi_0, \ldots, \psi_r) = p_{N_q}(x_g | \lambda\alpha_j, c_g\Sigma) \times p_{\Gamma^{-1}}(c_g | m/2, m\nu/2) \times \psi_j$$

where λ is a vector of length q equal to $(0, \ldots, 0, 1)^T$. Following the recipe of method I, with $T_g = (x_g, c_g, \delta_g)$ and $\theta = (\Sigma, \psi_0, \ldots, \psi_r)$, the contribution to the complete log-likelihood from T_g is (up to a constant)

$$\mathcal{L}_g(\theta, T_g) = \sum_{j=0}^{r} I_{(\delta_g = \alpha_j)} \left(\log(\psi_j) - \frac{1}{2}\log(|\Sigma|) - \frac{(x_g - \lambda\alpha_j)^T \Sigma^{-1} (x_g - \lambda\alpha_j)}{2c_g} \right)$$

where I_A is the indicator function for the event A. For the E-step, we need to find the conditional expectation of

$$I_{(\delta_g = \alpha_j)} \quad \text{and} \quad \frac{I_{(\delta_g = \alpha_j)}}{c_g}.$$

The distribution of $c_g | \delta_g = \alpha_j, x_g, \theta_0$ is $\Gamma^{-1}(m^*/2, m^* \nu^*_{gj}/2)$ with

$$m^* = m + q \quad \text{and} \quad \nu^*_{gj} = \frac{(x_g - \lambda\alpha_j)^T \Sigma_0^{-1} (x_g - \lambda\alpha_j) + m\nu}{m + q}.$$

Further, integrating out c_g yields the density of $x_g | \delta_g = \alpha_j, \theta_0$

$$p(x_g | \delta_g = \alpha_j, \theta_0) \propto \left[(x_g - \lambda\alpha_j)^T \Sigma_0^{-1} (x_g - \lambda\alpha_j) + m\nu \right]^{-(m+q)/2} = \left[m^* \nu^*_{gj} \right]^{-(m+q)/2}$$

and we get

$$\begin{aligned} p(\delta_g = \alpha_j | x_g, \theta_0) &\propto p(x_g, \delta_g = \alpha_j | \theta_0) \;=\; p(x_g | \delta_g = \alpha_j, \theta_0)\; p(\delta_g = \alpha_j | \theta_0) \\ &\propto \left[m^* \nu^*_{gj} \right]^{-(m+q)/2} \cdot \psi_{0j}. \end{aligned}$$

Hence, with $C_g^{-1} = \sum_{j=0}^{r} \left[m^* \nu^*_{gj} \right]^{-(m+q)/2} \cdot \psi_{0j}$ we have

$$p^*_{gj} = \mathrm{E}\left[I_{(\delta_g = \alpha_j)} | x_g, \theta_0 \right] \;=\; p(\delta_g = \alpha_j | x_g, \theta_0) \;=\; C_g \left[m^* \nu^*_{gj} \right]^{-(m+q)/2} \cdot \psi_{0j}$$

and

$$\mathrm{E}\left[\frac{I_{(\delta_g = \alpha_j)}}{c_g} \middle| x_g, \theta_0 \right] = p(\delta_g = \alpha_j | x_g, \theta_0) \cdot \mathrm{E}\left[1/c_g | \delta_g = \alpha_j, x_g, \theta_0 \right] \;=\; p^*_{gj} / \nu^*_{gj}.$$

Thus, $\mathrm{E}\left[\mathcal{L}_g(\theta, T_g) | x_g, \theta_0 \right]$ is equal to

$$\sum_{j=0}^{r} p^*_{gj} \left(\log(\psi_j) - \frac{1}{2}\log(|\Sigma|) - \frac{(x_g - \lambda\alpha_j)^T \Sigma^{-1} (x_g - \lambda\alpha_j)}{2\nu^*_{gj}} \right). \tag{9}$$

Turning to the M-step we start with the part of (9) that depends on the ψ_j's. Summing up the contributions from $T_1, \ldots, T_G$ we need to maximize

$$\sum_{g=1}^{G}\sum_{j=0}^{r}\log(\psi_j)\ p^*_{gj} = \sum_{j=0}^{r}\log(\psi_j)\ \hat{n}_j \quad \text{where} \quad \hat{n}_j = \sum_{g=1}^{G} p^*_{gj}. \tag{10}$$

This means that the maximization with respect to the ψ_j's can be done using the sufficient statistics $\hat{n}_0, \ldots, \hat{n}_r$. The parameters $\psi_0, \ldots, \psi_r$ must form a proper density and for $j > 0$, $\psi_j = Cf(\alpha_j|\beta)$ for a constant $C > 0$ where f is the density function of the continuous distribution F in (8). With these restrictions, (10) reduces to

$$\hat{n}_0 \log\Big(1 - C\sum_{j=1}^{r} f(\alpha_j|\beta)\Big) + \sum_{j=1}^{r}\hat{n}_j \log\Big(Cf(\alpha_j|\beta)\Big)$$

which should be maximized with respect to β and the constant C. The optimal C is

$$\hat{C} = \Big(1 - \hat{n}_0/\hat{n}.\Big)\Big/\sum_{j=1}^{r} f(\alpha_j|\beta) \quad \text{where} \quad \hat{n}. = \hat{n}_0 + \cdots + \hat{n}_r$$

which means that $\hat{\psi}_0 = \hat{n}_0/\hat{n}.$. The optimal β is found by maximizing

$$\sum_{j=1}^{r}\hat{n}_j \log\Big(f(\alpha_j|\beta)\Big) - (\hat{n}. - \hat{n}_0)\log\Big(\sum_{j=1}^{r} f(\alpha_j|\beta)\Big).$$

To find the optimal Σ, we introduce scaled and shifted versions of each observed vector x_g:

$$z_{gj} = (x_g - \lambda\alpha_j)\cdot\sqrt{p^*_{gj}/v^*_{gj}}.$$

Then the part of (9), which should be maximized with respect to Σ, is

$$-\sum_{g=1}^{G}\sum_{j=0}^{r}\left(p^*_{gj}\ \frac{1}{2}\log(|\Sigma|) + \frac{z_{gj}^T\Sigma^{-1}z_{gj}}{2}\right) = -\frac{G}{2}\log(|\Sigma|) - \sum_{g=1}^{G}\sum_{j=0}^{r}\frac{z_{gj}^T\Sigma^{-1}z_{gj}}{2}$$

since $p^*_{g0} + \cdots + p^*_{gr} = 1$ for all g. This means that the optimal Σ is

$$\hat{\Sigma} = \frac{1}{G}\sum_{g=1}^{G}\sum_{j=0}^{r} z_{gj}z_{gj}^T.$$

The updating of Σ and β is completed by scaling $\hat{\Sigma}$ to comply with the restriction $\text{trace}(\Sigma_A) = q - 1$.

6. SIMULATION WITH SPECIFIED COVARIANCE MATRIX

Using the same setup as Kristiansson et al. (2005), five different cases where studied as listed in Table 1. Each simulated dataset consists of 10,000 genes and four arrays. Data for each gene was generated as follows: $y_1, \ldots, y_4$ were generated as iid from either a standard normal distribution or a scaled (to unit variance) t-distribution ($df = 5$). For a fixed covariance matrix Σ and a scaling factor c generated from the $\Gamma^{-1}(2,2)$–distribution, the vector $y = (y1, \ldots, y_4)^T$ was multiplied by the square rot matrix of $c\Sigma$: $y \to (c\Sigma)^{1/2}y$. The covariance matrix was either the diagonal matrix $\Sigma = \text{diag}(0.4, 0.8, 1.2, 1.6)$ or the Σ described in Table 2. If a gene was selected to be up (or down) regulated, a scalar Δ was added to (or subtracted from) the vector y: $y \to y \pm \mathbf{1}\Delta$. The Δ's were generated from a uniform distribution between 0 and 2 (or 0 and 4).

TABLE 1. DESCRIPTION SUMMARY OF THE FIVE CASES SIMULATED WITH SPECIFIED COVARIANCE MATRIX

Case	Correlation	Heavy tails	Regulated genes	Method I $\hat{m}$	Method I $\hat{v}$	Method II $\hat{m}$	Method II $\hat{v}$	$1-\hat{\psi}_0$
1	No	No	No	4.00	1.00	4.02	1.00	0.31%
2	Yes	No	No	4.00	1.00	4.02	1.00	0.28%
3	Yes	Yes	No	3.05	0.77	3.11	0.78	0.02%
4	Yes	No	Medium	4.03	1.09	4.02	1.01	5.67%
5	Yes	No	Large	3.54	1.16	4.00	1.02	6.24%

"Heavy tails" denotes replacing normal distribution of the y's with the t-distribution with 5 degrees of freedom. When regulation is present, 5% of the genes were selected to be regulated in each direction. For correlation "No," the true covariance matrix Σ is diagonal with diagonal elements 0.4, 0.8, 1.2, and 1.6. For correlation "Yes," Σ is as listed in Table 2. The size of the regulation is uniform between 0 and 2 (Medium), or between 0 and 4 (Large). The means of $\hat{m}$, $\hat{v}$, and $\hat{\psi}_0$ are calculated across 1000 simulated data sets. The corresponding STDs for $\hat{m}$ and $\hat{v}$ ranged between 1.2% and 2.6% of the listed means. True values of m and v are 4 and 1, respectively. For $1-\psi_0$, the true value is 0% in cases 1, 2, and 3, and 10% in cases 4 and 5.

For method II, the distribution F is the symmetric mixture of two Γ-distributions in Table 3. Further, since method II estimates the covariance of transformed log-intensities ($x = P^T y$) the obtained estimates were transformed into estimates of the covariance matrix of y using

$$\hat{\Sigma} \to P\hat{\Sigma}P^T \quad \text{where} \quad P = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix}.$$

The result is found in Tables 1 and 2. In cases 1 and 2, when all assumptions of the models are true, including that of no regulated genes of method I, the estimates of m, v and Σ are all unbiased. When the assumption on conditional normal distribution of the y's is violated in case 3, the estimate of Σ appears to be as good as in cases 1 and 2. However, the shape and scale parameters are systematically too small.

In cases 4 and 5, where the assumption of method I on all genes being unregulated is violated, the method I estimate of Σ becomes biased. A bias is also observed for method II but compared to method I the bias is small. The WAME method for estimating Σ showed identical results with method I in terms of bias and variability for all five cases (results not shown).

In Figure 1, a subset of 100 of the fitted Γ-densities are shown for cases 4 and 5. Compared with the true scaled density, the scaled fitted densities have generally too little mass especially for δ close to zero. This is in line with the underestimated proportion of regulated genes in Table 1.

TABLE 2. PERFORMANCE SUMMARY OF METHODS I AND II WITH RESPECT TO BIAS OF THE ESTIMATOR $\hat{\Sigma}$ OF Σ

	Σ	Bias: Mean $(\hat{\Sigma})-\Sigma$ Method I	Method II
4	$\begin{bmatrix} 0.4 & 0.2 & 0.1 & 0.0 \\ 0.2 & 0.8 & 0.4 & 0.2 \\ 0.1 & 0.4 & 1.2 & 0.6 \\ 0.0 & 0.2 & 0.6 & 1.6 \end{bmatrix}$	$\begin{bmatrix} 0.04 & 0.05 & 0.06 & 0.07 \\ 0.05 & 0.01 & 0.04 & 0.05 \\ 0.06 & 0.04 & -0.01 & 0.03 \\ 0.07 & 0.05 & 0.03 & -0.04 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & -0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & -0.00 \end{bmatrix}$
5	$\begin{bmatrix} 0.4 & 0.2 & 0.1 & 0.0 \\ 0.2 & 0.8 & 0.4 & 0.2 \\ 0.1 & 0.4 & 1.2 & 0.6 \\ 0.0 & 0.2 & 0.6 & 1.6 \end{bmatrix}$	$\begin{bmatrix} 0.09 & 0.12 & 0.13 & 0.15 \\ 0.12 & 0.03 & 0.09 & 0.12 \\ 0.13 & 0.09 & -0.03 & 0.07 \\ 0.15 & 0.12 & 0.07 & -0.09 \end{bmatrix}$	$\begin{bmatrix} 0.01 & 0.01 & 0.02 & 0.02 \\ 0.01 & 0.00 & 0.01 & 0.01 \\ 0.02 & 0.01 & -0.00 & 0.01 \\ 0.02 & 0.01 & 0.01 & -0.01 \end{bmatrix}$

The means are based on 1000 simulated data sets. The different cases are further described in Table 1. The table shows results for cases 4 and 5 only; for cases 1, 2, and 3, neither method I or II showed any bias.

TABLE 3. DESCRIPTION OF PRIORS

$$\Gamma\text{-mixture:}\quad \delta \sim \begin{cases} \equiv 0 & \text{with probability } \psi_0 \\ \pm\Gamma(\alpha,\beta) & \text{each side with probability } (1-\psi_0)/2 \end{cases}$$

$$\text{Asymmetrical } \Gamma\text{-mixture:}\quad \delta \sim \begin{cases} \equiv 0 & \text{with probability } \psi_0 \\ +\Gamma(\alpha_1,\beta_1) & \text{with probability } \psi_1 \\ -\Gamma(\alpha_2,\beta_2) & \text{with probability } 1-\psi_0-\psi_1 \end{cases}$$

$$\text{Gaussian kernel:}\quad \delta \sim \begin{cases} \equiv 0 & \text{with probability } \psi_0 \\ F_{N,0.2} & \text{with probability } 1-\psi_0 \end{cases}$$

$F_{N,\alpha}$ denotes a kernel density using a Gaussian kernel with bandwidth adjusted by the factor α—that is, the default bandwidth (as defined in Silverman [1986], page 48, equation 3.11) multiplied with α.

A second simulation with m set to 1 was also performed (method I only, data not shown). The performance of the estimators follows the same pattern as with m equal to 4. Further, using a subset of 100 data sets from cases 4 and 5, a third analysis was done (data not shown). In this analysis, the distribution F of δ was specified by a Gaussian kernel density (Table 3). With this choice of prior the mean estimated proportion of regulated genes was 9.6 and 9.1 for cases 4 and 5, respectively, and the estimate of Σ showed no bias.

7. SIMULATION BASED ON REAL DATA

A second simulation was done based on the 18 arrays of the severe group of the chronic obstructive pulmonary disorder (COPD) data set (Spira et al., 2004), publicly available at the Gene Expression Omnibus repository (*www.ncbi.nlm.nih.gov/geo/*, series reference number GSE1650) consisting of data from Affymetrix arrays of type HG U133A. Using RMA expression indexes (Irizarry et al., 2003), 1000 data sets were generated and analyzed as follows: A random subset sized 8 was drawn from the 18 arrays and

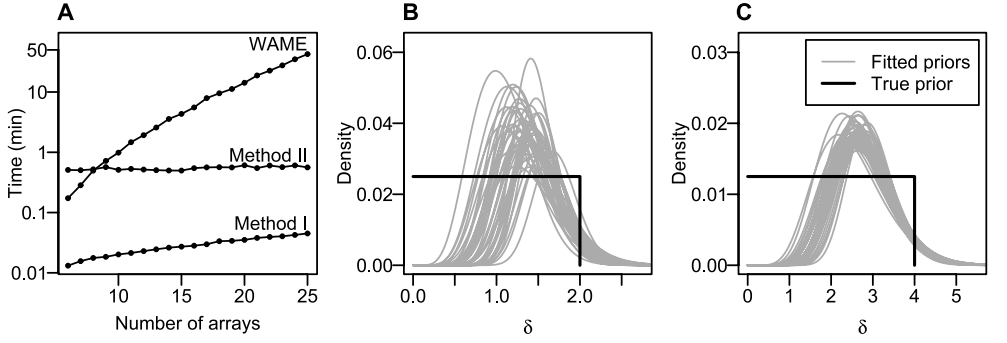

FIG. 1. **(A)** Computational time for estimating Σ as a function of the number of arrays. Point-wise mean with each mean based on 20 different data sets. **(B)** Fitted priors for 100 of the simulated datasets for case 4 of the simulation with specified covariance matrix. Gray curves are fitted densities, times the probability for a gene being regulated: $\hat{f} \times (1-\hat{\psi}_0)$. Black heavy line shows the true prior. Only the positive parts of the priors are shown. **(C)** As B, but for case 5.

divided into two equally sized groups. At this point, method I was used on the transformed log-intensities to estimate Σ, ν, and m. A random subset of 10% from the 22,283 probe sets was selected. For each selected probe set, a true regulation generated from the Γ-distribution (shape $= 2$, scale $= 4$) was added (or subtracted) to one of the two groups of four arrays. The scaled density is shown in Figure 2. After adding regulated genes method I and II were used on the transformed log-intensities. This is the same setup as used by Sjögren et al. (2007). Note that method I as well as method II was applied on transformed log-intensities. If y is the vector of eight log-intensities, analysis was done on the vector x of seven transformed log-intensities obtained as in (7) with

$$D^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} 0 & 2 \end{bmatrix}$$

specifying the transformation, as described in Section 5. For this simulation we do not know the true values of Σ, ν, and m. But we can compare the estimators obtained using method I on data *without* any regulated genes, with the ones obtained from the corresponding data *with* regulated genes. Ideally, the estimators of Σ, ν, and m should not be affected by the added regulation. The symmetric mixture of two Γ-densities in Table 3 was used as prior for the regulated genes for method II. Hence, the same density as the one used to generate the regulation. The result is summarized in Table 4 and Figure 2.

As seen in Table 4, all covariances as well as variances of the Σ_A part of Σ are fairly unaffected by the regulation for both methods. However, for method I the estimators of ν, m, and the variance Σ_B are clearly affected when regulation is added. The most obvious bias is seen when looking at the variance Σ_B, which is overestimated by a factor of 1.6. The result for method II is closer to the ideal with a slight bias for Σ_B of 3%. Moreover, the mean estimated proportion of regulated genes is close to the true proportion. In graph A of Figure 2, a subset of the fitted Γ-densities are plotted. The fitted densities are close to the true scaled density for δ above 0.75, whereas the fitted densities differ more from the true density for δ closer to zero. Graph B show the empirical distribution of the p-values obtained using method I and II. Due to the overestimated variance Σ_B, the p-values of method I are conservative, while method II shows no sign of producing conservative p-values.

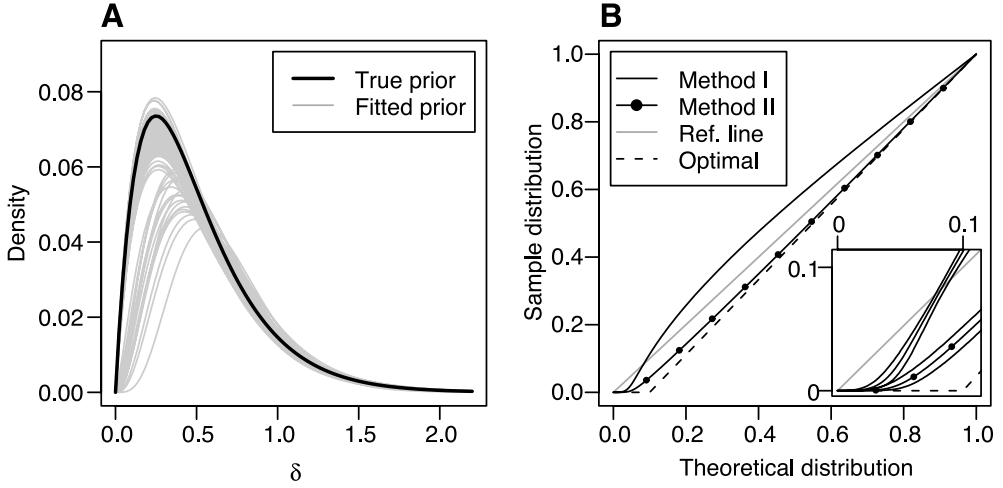

FIG. 2. Summary for simulation based on real data. **(A)** Fitted prior for δ for 100 of the simulated data sets. Gray curves are densities, times the probability for a gene being regulated: $\hat{f} \times (1 - \hat{\psi}_0)$. Black line is the true scaled density. Only the positive part is shown. **(B)** Pointwise median of the empirical distribution of observed p-values. Dashed line indicates optimal result under maximal power for detecting all regulated genes. The close-up also displays pointwise 5% and 95% quantiles, together with the pointwise median.

TABLE 4. SUMMARY FOR SIMULATION BASED ON REAL DATA

	Method I		*Method II*	
	Mean	*STD*	*Mean*	*STD*
Differences of covariances	0.000	0.003	0.000	0.002
Ratios of variances Σ_A	1.000	0.002	1.000	0.001
Ratios of variance Σ_B	1.660	0.098	1.026	0.017
Ratios of v	1.078	0.014	1.001	0.007
Ratios of m	0.881	0.028	0.991	0.024
$1 - \psi_0$ (%)	—	—	9.1	0.8

For each simulated dataset, three different estimators of Σ, v, and m were computed: (1) using method I *without* added regulation, (2) using method I *with* added regulation, and (3) method II *with* added regulation. Differences and ratios of 2 and 1 (method I) and 3 and 1 (method II) are summarized by mean and STD across all simulated data sets. Proportion of regulated genes was obtained using method II only. Differences and ratios should ideally be zero and 1, respectively. The true proportion of regulated genes was 10%.

8. COMPUTER RUN TIME

Using increased sized subsets of the complete COPD data set of 30 arrays, methods I and II together with the WAME method was used and the computer run time was recorded. A true regulation was added as done in Section 7, and 20 subsets were selected for each subset size. The mean computer run time is shown in Figure 1. Note the log-scale on the y-axis. Both methods I and II have computer run times less than one minute, whereas the WAME method requires nearly 50 minutes for 25 arrays.

9. REAL DATA

Using a dataset with very clear differences between some of the designed groups, the methods were compared in the situation where regulated genes exist. RNA from two strains of mice, each strain exposed to three treatments was hybridized to Affymetrix arrays. The obtained CEL-file data was pre-processed using the RMA (Irizarry et al., 2003) method. A PCA plot of the 34 arrays is shown in graph A of Figure 3. The plot shows very clear separation between the two strains for all three treatments. Moreover, for strain A the three treatments are also well separated. Hence, using within treatment comparisons of the two strains and within strain A comparisons of the three treatments 6 comparisons can be studied which undoubtedly are under H_1: Regulated genes exists. The result is summarized in Table 5 and graph B of Figure 3.

TABLE 5. MOUSE DATA SET

	Prior	*A0/B0*	*A1/B1*	*A2/B2*	*A0/A1*	*A0/A2*	*A1/A2*
$1 - \hat{\psi}_0$	Γ-mixture	6.67%	3.63%	4.75%	3.44%	1.43%	1.09%
	Asymmetrical Γ-mixture	6.50%	3.45%	7.19%	7.10%	7.21%	1.77%
	Gaussian kernel	5.22%	3.18%	4.36%	3.72%	1.80%	1.69%
$\hat{\Sigma}_B$	0	3.46	4.99	4.03	12.53	4.73	5.08
	Γ-mixture	2.58	4.38	3.23	10.37	4.36	4.89
	Asymmetrical Γ-mixture	2.58	4.37	3.07	9.97	4.07	4.80
	Gaussian kernel	2.62	4.37	3.20	10.30	4.27	4.85

Estimated proportion of regulated genes ($1 - \hat{\psi}_0$) in upper part of the table, and variance of last the dimension ($\hat{\Sigma}_B$) in lower part of the table for the six pair-wise group comparisons. Prior 0 denotes results obtained when using method I for estimating the covariance matrix; A0/B0 denotes the comparison between strain A and B, both under treatment 0.

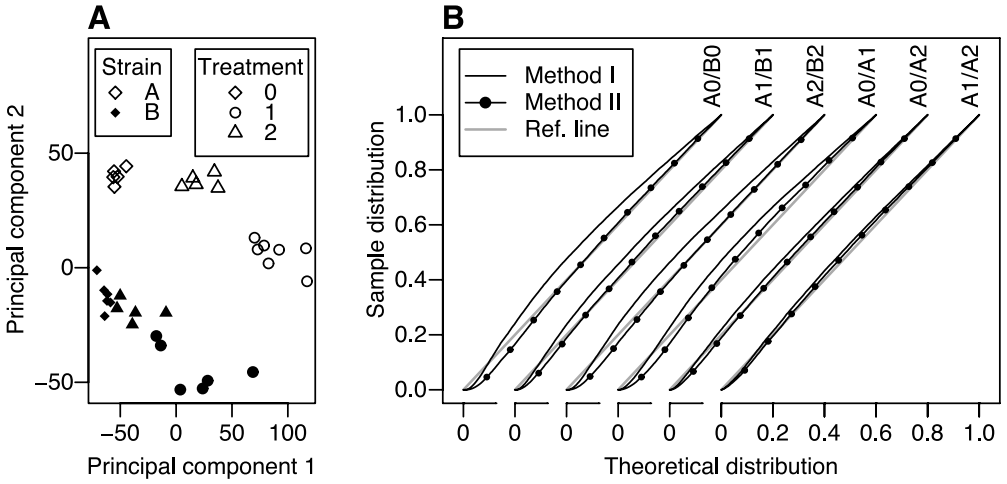

FIG. 3. Mouse data set. **(A)** PCA plot for the 34 arrays. **(B)** Empirical distribution of observed p-values for the six pairwise comparisons. The p-values of method II are the ones obtained using the asymmetric Γ-mixture as prior for δ_g. A0/B0 denotes the comparison between strain A and B, both under treatment 0.

Methods I and II were applied to transformed log-intensities for each of the six comparisons. For method II, all three priors in Table 3 were used as the prior distribution F for δ. The Σ-estimators were all scaled so that the mean of the variances of the Σ_A part of Σ equals 1. Also the columns of the P-matrix determining the transformation were scaled to unit length. Thus we would expect the variance Σ_B to be fairly close to one. But as seen in Table 5 all estimators of Σ_B are above 2.5 and for the A0 versus A1 comparison estimators as high as 10 were obtained. Moreover, in graph B of Figure 3, the empirical p-value distributions show the same S-shape as seen for method I in Figure 2. This indicates that method I as well as II produces biased estimators of Σ_B.

10. CONCLUSION

We have presented two new methods, I and II, for estimating the global covariance matrix Σ of the WAME model suggested by Kristiansson et al. (2005). Method I is a direct application of the EM algorithm for fitting the multivariate t-distribution. Compared with the method used by Kristiansson et al. (2005) method I shows identical bias and variability on simulated data with specified covariance matrix, but has superior computer run times. From the simulation based on real data we found that for the vector of transformed data it was only the estimator of Σ_B that was affected when regulation was added. Analytically, it can be shown that neither the minimum variance estimator (2) nor the weighted residual sum of squares (4) is affected when Σ is altered by changing Σ_B only. It is only the global constant $C(D^T\Sigma^{-1}D)^{-1}C^T$ in the denominator of the weighted moderated t-statistic (3) that is affected by such a change of Σ. Hence, just as long as Σ is positive definite, the ranking of the genes using (3) is invariant with respect to the estimated value of Σ_B. This means that the bias of $\hat{\Sigma}$ is merely a problem when setting the scale for the t-statistic (3) so that reliable p-values or an appropriate cut-off can be obtained. Indeed, this is a problem shared with quite a few methods and general techniques for setting the threshold given a z-score exists, e.g., local FDR (Efron, 2005) and mixture models such as Dean and Raftery (2005) and Broët et al. (2002).

With method II, we try to reduce the bias by modeling the unknown log fold-change using the prior distribution F. A discrete approximation, $\tilde{F}$, is used giving a finite set of regulation levels. Hence, method II is also a mixture model where restrictions are imposed onto the mixture proportions through the distribution F. For simulated data, method II greatly reduces the bias but for real data the estimated Σ still appears biased, most likely due to false model assumptions. The question is in what way.

According to the model used by method II, the prior distribution is the same for all genes. The model used by Lönnstedt and Britton (2005) shares this property, but for other models the unknown gene-specific variance also determines the variance for the distribution of the log fold-change—for example, the models used by Lönnstedt and Speed (2002) and Smyth (2004), where the conditional distribution of the log fold-change is normal with zero mean and variance proportional to c_g. Thus, alternatively the distribution of δ_g could be specified conditionally on c_g.

Method II worked well for simulated data with different variances across genes. But for data where all genes have the same variance, method II can produce strange estimators. For example, with F equal to the normal distribution, the same likelihood would be obtained with the proportion of regulated genes set to one as when set to zero. Similar results would also be obtained with F equal to the mixture of two gamma distributions. This means that method II can produce strange estimators when the variances do not differ enough and/or the arrays are too few so that gene variances can not be sufficiently separated.

In summary, method I is a substantial improvement for weighted analysis of microarray data, making it possible to analyze large datasets with acceptable computer run times. Also, the case with missing data is within range using existing applications of the EM algorithm (Lange et al., 1989). With the discrete approximation $\tilde{F}$, method II is a useful extension of the WAME model managing with any choice of distribution F. Although being flexible for the choice of F, method II still suffers from lack of fit to data resulting in biased estimators of Σ. Both methods are implemented using a mixture of R and C code available at *www.math.chalmers.se/∼astrandm*.

ACKNOWLEDGMENTS

We thank Anders Sjögren and Olle Nerman for valuable discussions. The research was supported by the Gothenburg Mathematical Modeling Center and the Gothenburg Stochastic Center.

REFERENCES

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.

Broët, P., Richardson, S., and Radvanyi, F. 2002. Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *J. Comput. Biol.* 9, 671–681.

Dean, N., and Raftery, A. 2005. Normal uniform mixture differential gene expression detection for cDNA microarrays. *BMC Bioinform.* 6, 173.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* 39, 1–38.

Efron, B. 2005. Local false discovery rates [Technical report]. Stanford University, Stanford, CA. Available at: *www-stat.stanford.edu/∼brad/papers/False.pdf*. Accessed October 15, 2007.

Hein, A.M., and Richardson, S. 2006. A powerful method for detecting differentially expressed genes from GeneChip arrays that does not require replicates. *BMC Bioinform.* 7, 753.

Irizarry, R.A., Hobbs, B., Collin, F., et al. 2003. Exploration, normalization, and summaries of high-density oligonucleotide array probe level data. *Biostatistics* 4, 249–264.

Johnson, N., Kotz, S., and Balakrishnan N. 1995. *Continuous Univariate Distributions, Volume 2*, 2nd ed. Wiley, New York.

Kristiansson, E., Sjögren, A., Rudemo, M., et al. 2005. Weighted analysis of paired microarray experiments. *Statist. Appl. Genet. Mol. Biol.* 4, 30.

Kristiansson, E., Sjögren, A., Rudemo, M., et al. 2006. Quality optimised analysis of general paired microarray experiments. *Statist. Appl. Genet. Mol. Biol.* 5, 10.

Lange, K.L., Little, R.J.A., and Taylor, J.M.G. 1989. Robust statistical modelling using the *t* distribution. *J. Am. Statist. Assoc.* 84, 881–896.

Lehmann, E.L. 1986. *Testing Statistical Hypotheses*, 2nd ed. Chapman & Hall, New York.

Lewin, A., Richardson, S., Marshall, C., et al. 2006. Bayesian modelling of differential gene expression. *Biometrics* 62, 1–9.

Lönnstedt, I., and Britton, T. 2005. Hierarchical Bayes models for cDNA microarray gene expression. *Biostatistics* 6, 279–291.

Lönnstedt, I., and Speed, T.P. 2002. Replicated microarray data. *Statist. Sinica* 12, 31–46.
Ritchie, M., Diyagama, D., Neilson, J., et al. 2006. Empirical array quality weights in the analysis of microarray data. *BMC Bioinform.* 7, 261.
Silverman, B.W. 1986. *Density Estimation*. Chapman & Hall, New York.
Sjögren, A., Kristiansson, E., Rudemo, M., et al. 2007. Weighted analysis of general microarray experiments. *BMC Bioinformatics* 8, 387.
Smyth, G.K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Mol. Biol.* 3, 3.
Spira, A., Beane, J., Pinto-Plata, V., et al. 2004. Gene expression profiling of human lung tissue from smokers with severe emphysema. *Am. J. Respir. Cell Mol. Biol.* 31, 601–610.
Tusher, V.G., Tibshirani, R., and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* 98, 5116–5121.

Address reprint requests to:
Dr. Magnus Åstrand
Mathematical Statistics
Chalmers University
Göteborg, Sweden

E-mail: astrandm@chalmers.se

Paper IV

Empirical Bayes models for multiple probe type microarrays at the probe level

Magnus Åstrand*[1], Petter Mostad[1] and Mats Rudemo[1]

[1]Mathematical Sciences, Chalmers University of Technology, and Mathematical Sciences, Göteborg University, S-41296, Göteborg, Sweden

Email: Magnus Åstrand*- astrandm@chalmers.se;magnus.astrand@astrazeneca.com; Petter Mostad - mostad@chalmers.se; Mats Rudemo - rudemo@chalmers.se;

*Corresponding author

Abstract

Background: When analyzing microarray data a primary objective is often to find differentially expressed genes. With empirical Bayes and penalized t-tests the sample variances are adjusted towards a global estimate, producing more stable results compared to ordinary t-tests. However, for Affymetrix type data a clear dependency between variability and intensity-level generally exists, even for logged intensities, most clearly for data at the probe level but also for probe-set summarizes such as the MAS5 expression index. As a consequence, adjustment towards a global estimate results in an intensity-level dependent false discovery rate.

Results: We propose two new methods for finding differentially expressed genes, Probe level Locally moderated Weighted median-t (PLW) and Locally Moderated Weighted-t (LMW). Both methods use an empirical Bayes model taking the dependency between variability and intensity-level into account. A global covariance matrix is also used allowing for differing variances between arrays as well as array-to-array correlations. PLW is specially designed for Affymetrix type arrays (or other multiple-probe arrays). Instead of making inference on probe-set summaries, comparisons are made separately for each perfect-match probe and are then summarized into one score for the probe-set.

Conclusions: The proposed methods are compared to 12 existing methods using five spike-in data sets. PLW has the most accurate ranking of regulated genes in four out of the five data sets, and LMW consistently performs better than all examined moderated t-tests when used on RMA expression indexes.

Background

Microarrays are widely used for measuring gene expression in biomedical research. For the purpose of finding differentially expressed genes there exist numerous methods. In early studies genes where often ranked with respect to fold-change. Genes showing fold-change above 2 (or 3) were regarded as potentially regulated and were selected for further investigation. The obvious drawback with such an approach, as pointed out by many authors, is that genes with high fold-change may also be highly variable and thus with low significance of the regulation. On the other hand, since the number of replicates in many studies is small, variance estimators computed solely within genes are not reliable in that very small values can occur just by chance. As a consequence the ordinary t-test suffers from low power and is not a better option for filtering out regulated genes.

Many methods have been proposed to improve on the variance estimator in order to find more powerful statistical tests for differential expression. In empirical Bayes methods [1–8] and the penalized t-test suggested in [9], the gene-specific variance estimator is modified in order to produce more stable results. With proportions determined by the accuracy of the gene-specific variance estimators, a mixture of the gene-specific variance estimator and a global variance estimate is used in place of the gene-specific variance estimator in the denominator of the t-test. Similarly, in the Significance Analysis of Microarrays (SAM) method [10] and the method suggested in [11], a constant is added to the gene-specific sample standard deviation.

Another approach is to pool variance estimators for genes having similar expression level, thus modeling the variance as a function of intensity-level. For example Eaves et al. [12] use a weighted average of the gene-specific variance estimator and a pooled estimate based on the 500 genes with most similar mean expression level, and Jain et al. [13] suggest the local-pooled-error method (LPE) where a variance function fitted to estimated variances and mean intensities is used. Comander et al. [14] pool genes with respect to minimum intensity rather than mean intensity, and Hu et al. [15] use a hierarchical model with a linear relationship between variance and intensity-level. Of these four methods, only the one suggested in [15] takes the accuracy of the gene-specific variance estimators into account when setting the weights for the gene-specific estimator and the pooled estimator, respectively. On the other hand Hu et al. [15] only deal with a linear relationship between variance and intensity-level. A variance to intensity-level dependency is also utilized in the moderated t-test suggested in [6]. The method proposed builds on the moderated t-test suggested in [2,3] with the addition of fitting a loess curve in the scatter plot of logged variance estimators against mean intensity when estimating the model parameters.

The type of arrays considered in this paper is the Affymetrix GeneChip arrays. These arrays are one color arrays and each gene is represented by a set of probes, the probe-set, consisting of 10-16 probe-pairs. Each probe-pair consists of one perfect match (PM) probe and one mismatch (MM) probe. The probes are 25 bases long and the PM and MM probes have identical sequences of bases except for the middle probe which in the MM probe is set to the complementary base of that in the PM probe. The MM probes are thus designed to measure the background intensity for the corresponding PM probe. The standard way of dealing with the multiple-probes is to derive a summary measurement, an expression index, for each probe-set (gene) and array (sample), for example using the RMA method [16] or the Affymetrix

MAS5 algorithm. The expression indexes are then used in downstream analysis by only considering the expression index itself, the precision of the expression index is ignored. However, in the fully Bayesian probe-level BGX model [17] information about the accuracy of the expression index is obtained as a complete distribution which is subsequently used when computing the posterior distribution of differential expression. Also, the probe-level measurement error from the probabilistic probe-level model multi-mgMOS [18] is used when computing the probability of positive log-ratio in the PPLR method [19].

For Affymetrix type arrays a dependency between variability and intensity-level generally exists, even for log-transformed data. Figure 1 shows scatter plots of sample variance versus sample mean calculated on logged PM intensities (background corrected and normalized) and three different expression indexes: RMA, GCRMA and MAS5. Except for the RMA expression index a clear dependency between variability and intensity-level exists, with a unique signature for each type of pre-processing of the raw CEL-file data. The GCRMA expression index shows increasing variability with intensity-level while MAS5 shows the opposite relationship. As a consequence, methods assuming constant variance as well as methods adjusting the gene-specific variance (or standard deviation) estimators towards a global estimate suffer from intensity-level dependent false discovery rates. Figure 2 shows an example where the moderated t-test in the R-package LIMMA [3] was used on MAS5 expression indexes computed on a set of replicated arrays. The false discovery rate obtained with LIMMA follows the same pattern as in the right lower panel in Figure 1 where the same data set is used.

The aim of variance stabilizing transformations is to reduce or eliminate the problem of dependency between variability and intensity-level. A family of transformations, the generalized-log family (glog), was introduced in [20–22] and further used in [23, 24]. A comparison of the glog family with the started logarithm transformation [25] and the log-linear hybrid transformation [26] is presented in [27]. It is concluded that the glog family is "probably the best choice when it is convenient to use it", but it is also noted that the direct interpretation of differences as logged ratios for microarray data when using the ordinary log-transformation, does not hold when using such variance stabilizing transformations. Generally, the glog family effectively stabilizes the variance when applied to raw Affymetrix probe-level data, for example using the parameter estimation procedure described in [21]. However, the transformations implicitly defines a background correction, and when applied to data already having been subject to another background correction (or further processed data), the glog transformations may not be able to capture the structure of the dependency between variability and intensity-level. This applies, for example, to probe-level data background corrected using the RMA default background method, and MAS5 expression indexes, see Figure 2. Thus, there is a need for more flexible solutions, and in short, Figures 1 and 2 may be seen as a motivation for the methods proposed in this paper.

The hierarchical Bayesian model WAME proposed and developed in [4, 5, 7, 8] is in the present paper extended to incorporate the variability to intensity-level dependency. The Probe level Locally moderated Weighted median-t method (PLW) applies the extended model to logged PM intensities resulting in moderated and weighted t-statistics for all PM probes. In the final step of PLW the median t-statistic of all

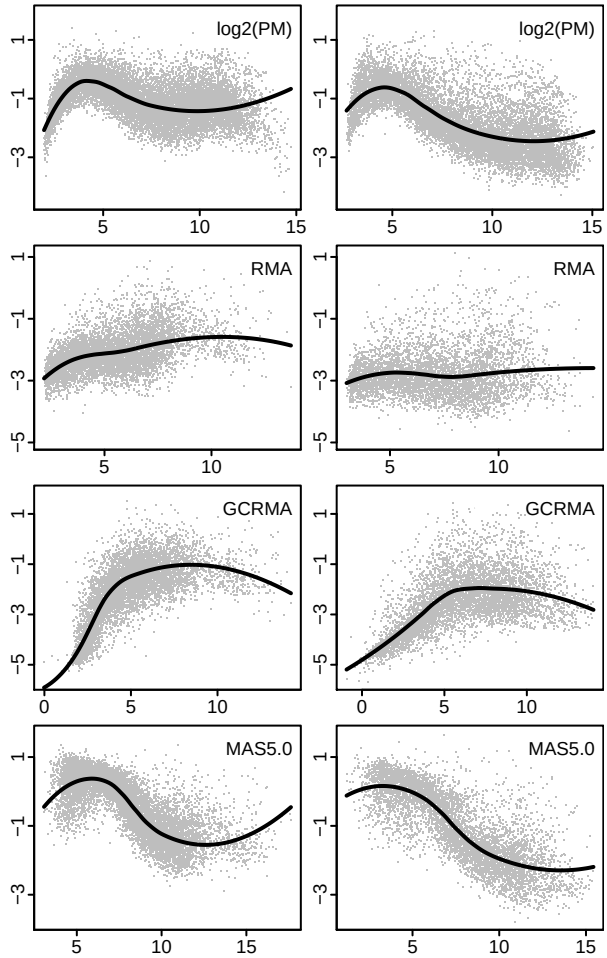

Figure 1: Scatter plots of sample variance (logged with base 2) against mean intensity for logged PM intensities and three expression indexes. Left and right panels show data set A and B, respectively (see Section *Data sets* on page 8).

PM probes building up each probe-set is computed, and this median is the value used for ranking the probe-sets with respect to differential expression.

The Locally Moderated Weighted-t method (LMW) is a more general method intended for single probe type of arrays or summary measures of multiple probe type arrays, such as RMA and MAS5. LMW use the same model as PLW but since only one t-statistic is obtained for each probe-set no median is calculated. The proposed methods are compared with existing methods on five publicly available spike-in data sets.

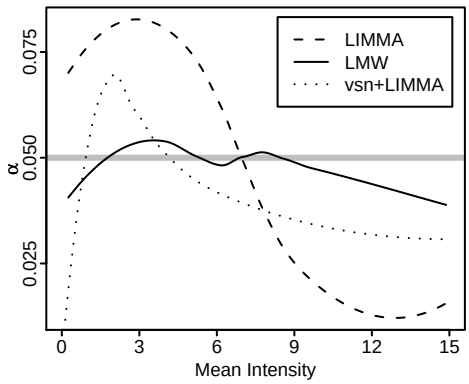

Figure 2: False discovery rate (α) calculated on re-sampled data and plotted against mean intensity. Data sets of size 6 were sampled from the complete data set B (see Section *Data sets*) of 18 replicated arrays and then analyzed using the Affymetrix MAS5 algorithm followed by a two group analysis of 3+3 arrays using the moderated t-test in the R-package LIMMA [3], on logged MAS5 indexes and indexes transformed using the variance stabilizing transformation in the R-package vsn [21], and the proposed method LMW using logged MAS5 indexes. False discovery rate were obtained by averaging over the sampled data sets using loess-curves fitted to mean intensity and indicator of significance (1 if the probe-set is among the 5% probe-sets with highest absolute statistic, 0 otherwise). The mean intensities of each data set are shifted to the range [0,15].

Results and Discussion

Model and methodology

Given a set of n arrays let y_{ip} be the background corrected and normalized log-intensity on array i for PM probe p and put $y_p = (y_{1p}, \ldots, y_{np})^T$. The PM probes are divided into G (disjoint) probe-sets $\mathcal{G}_1, \ldots, \mathcal{G}_G$ and thus there are a total of $P = |\mathcal{G}_1| + \cdots + |\mathcal{G}_G|$ probes. For $p = 1, \ldots, P$ assume

$$\begin{aligned} y_p | c_p &\sim \mathrm{N}_n(\mu_p, c_p \Sigma) \\ c_p &\sim \Gamma^{-1}(\tfrac{1}{2}m, \tfrac{1}{2}m \cdot \nu(\bar{\mu}_p)) \end{aligned} \tag{1}$$

where μ_p is the log-intensity profile for probe p across the n arrays with mean log-intensity level $\bar{\mu}_p$, Σ is an $n \times n$ covariance matrix, m is a real-valued parameter, and $\nu(\cdot)$ is a smooth real-valued function. N_n denotes an n-dimensional normal distribution, and $\Gamma^{-1}(a, b)$ denotes the inverse-gamma distribution with shape parameter a and scale parameter b. A cubic spline is used to parameterize the function $\nu(\cdot)$. Given set of K interior spline-knots

$$\nu(x) = \exp\{H(x)^T \beta\}$$

where β is a parameter vector of length $2K - 1$ and $H : \mathbb{R} \to \mathbb{R}^{2K-1}$ is a set of B-spline basis functions, see chapter 5 of [28].

As in the model suggested in [4] the model in Equ. 1 makes use of a global covariance matrix, thus allowing differing variances as well as correlations between arrays. To account for the dependency between variability and intensity-level the scale-parameter of the Γ^{-1}-distribution depends on the mean log-intensity level $\bar{\mu}_p$ for the probe through the smooth function ν.

We assume that the vector μ_p is determined by a full rank $n \times k$ design matrix D and a parameter vector γ_p of length k. The aim is to estimate and test hypothesis for δ_p, a linear combination of γ_p specified by a $1 \times k$ matrix C. In summary,

$$\mu_p = D\gamma_p \quad \text{and} \quad \delta_p = C\gamma_p \ .$$

For the special case of comparing two conditions, with n_1 and n_2 arrays from conditions 1 and 2, respectively, the design matrix D is an $(n_1 + n_2) \times 2$ matrix. For example, with $n_1 = 3$ and $n_2 = 4$ we can use

$$D^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } \gamma_p = \begin{bmatrix} \gamma_{p1} \\ \gamma_{p2} \end{bmatrix}$$

and thus $\mu_p = (\gamma_{p1}, \gamma_{p1}, \gamma_{p1}, \gamma_{p2}, \gamma_{p2}, \gamma_{p2}, \gamma_{p2})^T$. With $C = [-1 \quad 1]$ we have $\delta_p = \gamma_{p2} - \gamma_{p1}$, thus δ_p is the logged fold change between conditions 2 and 1.

However, instead of estimating the parameters of the model in Equ. 1 we use a reduced model derived from Equ. 1 through a linear transformation of the vector y_p. Define the $n \times n$ and $n \times 1$ matrices

$$A_0 = I - D(D^T D)^{-1} D^T \quad \text{and} \quad B = D(D^T D)^{-1} C^T \ .$$

Since A_0 is of rank $n - k$ only we let A be an $n \times (n - k)$ matrix whose column space equals that of A_0. With $q = n - k + 1$ form the $n \times q$ transformation matrix M and the vector z_p of length q

$$M = [A; B] \quad \text{and} \quad z_p = M^T y_p \tag{2}$$

giving the reduced model

$$\begin{aligned} z_p | c_p &\sim \mathrm{N}_q\Big((0, \ldots, 0, \delta_p)^T, c_p \Sigma_z\Big) \\ c_p &\sim \Gamma^{-1}(\tfrac{1}{2}m, \tfrac{1}{2}m \cdot \nu(\bar{\mu}_p)) \end{aligned} \tag{3}$$

where $\Sigma_z = M^T \Sigma M$.

The reduced model is fitted using the EM algorithm [29] as described in Section *Parameter estimation*. The c_p's are treated as missing data and we replace the unknown intensity-level for probe p, $\bar{\mu}_p$, with the observed mean intensity across arrays, $\bar{y}_p$. Given estimators of the parameters Σ_z, m, and β we proceed as if these parameters are known, and weighted moderated t-tests are computed for each probe p. The unbiased minimum variance estimator of δ_p is

$$\hat{\delta_p} = (\lambda^T \Sigma_z^{-1} \lambda)^{-1} \lambda^T \Sigma_z^{-1} z_p \tag{4}$$

where λ is the vector $(0, \ldots, 0, 1)^T$ of length q. The weighted moderated t-statistic is defined as

$$\tilde{t}_p = \sqrt{\frac{q+m-1}{(\lambda^T \Sigma_z^{-1} \lambda)^{-1}}} \frac{\hat{\delta}_p}{\sqrt{m \exp\{H(\bar{y}_p)^T \beta\} + \mathrm{RSS}_p}} \tag{5}$$

and under H_0: $\delta_p = 0$ it can be shown that $\tilde{t}_p$ is t-distributed with $q+m-1$ degrees of freedom. Here

$$\mathrm{RSS}_p = z_p^T \Big(\Sigma^{-1} - \Sigma^{-1} \lambda (\lambda^T \Sigma^{-1} \lambda)^{-1} \lambda^T \Sigma^{-1} \Big) z_p \tag{6}$$

is the weighted residual sum of squares. See [5] for details. The PLW statistic for the probe-set $\mathcal{G}$ is then defined as

$$\mathrm{PLW}_{\mathcal{G}} = \mathrm{median} \Big\{ \tilde{t}_p : p \in \mathcal{G} \Big\} \ . \tag{7}$$

The LMW and PLW methods are implemented in the R package plw [30].

Parameter estimation

The $q \times q$ covariance matrix Σ_z of the reduced model in Equ. 3 is divided according to

$$\Sigma_z = \begin{bmatrix} \Sigma_A & \Sigma_{AB} \\ \Sigma_{AB}^T & \sigma_B^2 \end{bmatrix}$$

where Σ_A is the covariance matrix for all but the last dimension of z_p and σ_B^2 is the variance of the last dimension (indexes A and B refer to the corresponding sub-matrices of the transformation matrix M in Equ. 2). The reduced model is fitted in two steps. First the parameters m, β and the sub-matrix Σ_A are estimated by dropping the last dimension of the vectors z_p. Since the reduced model is not identifiable without a restriction on the function ν or the covariance matrices Σ_z we use the restriction $\mathrm{trace}(\Sigma_A) = q - 1$. Secondly, the parameters m and β are held fixed and Σ_z is estimated using the complete z_p vectors. Temporarily the assumption of no regulated genes is used ($\delta_p = 0$ for all probes) and Σ_z is estimated under the restriction that the trace of the Σ_A part should be equal to $q - 1$.

In step 1, we let x_p denote the sub-vector of z_p obtained by dropping the last element. Under the reduced model x_p is distributed according to the model in Equ. 1 with $\Sigma = \Sigma_A$, $\mu_p = 0$, $n = q - 1$, and using the EM-algorithm an iterative procedure for estimating m, β and Σ_A is obtained. Given estimates of the previous iteration, m_0, β_0 and Σ_{A0}, updated estimates are found as follows. Let

$$w_p = \frac{m_0 + q - 1}{x_p^T \Sigma_{A0}^{-1} x_p + m_0 \exp\{H(\bar{y}_p)^T \beta_0\}} \ .$$

The updated estimate of Σ_A is

$$\hat{\Sigma}_A = \frac{1}{P} \sum_{p=1}^{P} w_p x_p x_p^T \tag{8}$$

and the updated estimate of β is found by numerical maximization of the function

$$h(\beta) = \frac{1}{P}\sum_{p=1}^{P}\Big(H(\bar{y}_p)^T\beta - w_p \exp\{H(\bar{y}_p)^T\beta\}\Big) \ . \tag{9}$$

With $\hat{\beta}$ equal to the updated estimate of β let

$$\begin{aligned} S &= h(\hat{\beta}) - \log(m_0 + q - 1) \\ &\quad + \psi\left(\frac{m_0 + q - 1}{2}\right) + \frac{1}{P}\sum_{p=1}^{P}\log(w_p) \end{aligned}$$

where $\psi(x) = \frac{d}{dx}\log\Gamma(x)$ is the digamma function. The updated estimate of m is then found using numerical maximization of the function

$$f(m) = m\Big(\log(m) + S\Big) - 2\log\Big(\Gamma(m/2)\Big) \ . \tag{10}$$

In step 2 a similar iterative procedure is used to estimate Σ_z. With Σ_{z0} denoting the estimate of Σ_z from the previous iteration and with w_p re-defined as

$$w_p = \frac{\hat{m} + q}{z_p^T\Sigma_{z0}^{-1}z_p + \hat{m}\exp\{H(\bar{y}_p)^T\hat{\beta}\}}$$

where $\hat{m}$ and $\hat{\beta}$ are the estimates obtained in step 1, an updated estimate of Σ_z is computed according to Equ. 8 with z_p replacing x_p. In order for the estimators of Σ_A and Σ_z, in step 1 and 2, respectively, to comply with the trace restriction the updated estimates are scaled at the end of each iteration. See additional file 1: Supplement1.pdf for more details.

Data sets

The two data sets used in Figures 1 and 2 are publicly available at the Gene Expression Omnibus repository [31] with series or sample reference number indicated below. Data set A consists of the 18 arrays from the severe group of the COPD data set [32] (series reference number GSE1650), where Affymetrix arrays of type HG U133A were used. In data set B the 18 arrays with normal tissue where selected from a lung tumor data set [33] (sample reference numbers GSM47958-GSM47976, excluding GSM47967). Here the HG-U95A arrays were used.

Five spike-in data sets were used to evaluate the proposed methods. In the Affymetrix U95 and 133A Latin Square data sets [34] arrays of type HG-U95A and HG-U133A, respectively, were used. The Affymetrix U95 data set consists of data from 59 arrays divided into 19 groups of size 3, and one group of size 2. From the 20 groups there are 178 possible pair-wise group comparisons each with 16 [35] known differentially expressed genes among the 12626 genes present on the arrays. The Affymetrix 133A data set comprise data from 42 arrays with a total of 22300 probesets of which 42 were spiked in at known concentration. The 42 arrays are divided into 14 groups of size 3 and thus there are 91 possible pair-wise group comparisons. As

done in the Affycomp II assessments [35] we exclude 271 probe-sets which are likely to cross-hybridize to spike-in probe-sets. The sequence of each spike-in clone was blasted against all HG-U133A target sequences (∼600bp regions from which probes are selected). A threshold of 100bp identified 271 probe-sets which are available in the affycomp R-package.

From the Gene Logic Tonsil and AML data sets [36] all groups with 3 replicated arrays were used, giving a total of 12 and 10 groups, respectively. For these data there are 11 genes spiked in at known concentration, which can be studied in 66 and 45 pair-wise group comparisons, respectively. Both data sets were obtained using the Affymetrix HG-U95A arrays having 12626 genes.

The Golden Spike data set [37] consists of 6 arrays of type Drosgenome1 divided into 2 groups of equal size. The samples used in this experiment consist of mRNA from 3866 genes, of which 1331 are differentially expressed between the groups. The Drosgenome1 array has a total of 14010 genes, thus 10144 of these should not be expressed, 2535 should be expressed but not regulated, and 1331 should be expressed and regulated. It has been observed that the 2535 genes make these data atypical [38]. We have chosen to exclude the 2535 genes from the analysis, thus only using 11475 genes of which 1331 are known to be regulated. Also, since all 1331 genes are up-regulated it is necessary to take special care in the normalization. In place of the quantile normalization method, the default for the RMA method, we used the contrast normalization [39] and fitted the normalization curve using PM probes from the 11475 unregulated genes only. To present comparable result only, methods relying on a normalization procedure at the probe-set level (PPLR and BGX) and methods for which the default normalization method is not the quantile method (logit-t) were excluded when analyzing this data set.

Comparison with existing methods

Using the spike-in data sets listed above the proposed methods, PLW and LMW were compared with 12 existing methods for ranking genes. The 12 methods include ranking with respect to: observed fold change (FC), ordinary t-test, the moderated t-test in the R-package LIMMA [3], the weighted moderated t-test in the R-package WAME.EM [8], Efron's penalized t-test [11] and the Shrink-t method [9] in the R-package st, the SAM method [10] in the R-package samr, the Local-pooled-error test [13] in the R-package LPE, and the Intensity-Based Moderated T-statistic (IMBT) [6] using the R-code available at `http://eh3.uc.edu/r/ibmtR.R`. All these methods (including LMW) were applied to RMA expression indexes obtained using the R-package affy, while PLW was applied to logged PM intensities, background corrected and normalized using the default methods of RMA. With LMW 4-6 spline-knots (depending on the number of probe-sets) were used for the function ν, whereas 12 knots were used in PLW (the spline-knots are set using an internal function in the R-package plw). Note that the RMA method was applied only to the arrays involved in each group comparison, as opposed to running the RMA method using all arrays of each data set.

We also compared with the PPLR method [19] applied to the expression index and probe-level measurement error of the multi-mgMOS model [18] available in the R-package puma, the logit-t procedure implemented in the R-package plw according

Table 1: Area under ROC curves up to 100 false positives rounded to nearest integer value with an optimum of 100. Numbers within parenthesis are within data set ranks for the methods compared. Methods are ordered with respect to mean rank across the five data sets.

	Affymetrix		Golden	Gene Logic	
Method	U95	133A	Spike	Tonsil	AML
PLW	96(1)	93(6)	40(1)	87(1)	86(1)
LMW	96(2)	94(1)	32(3)	84(3)	80(4)
LPE	94(5)	93(10)	38(2)	84(2)	85(2)
WAME	95(3)	94(2)	32(7)	81(5)	78(7)
Efron-t	94(6)	93(4)	32(5)	79(7)	79(5)
IBMT	95(4)	94(3)	32(8)	78(8)	76(8)
FC	92(11)	93(5)	31(10)	83(4)	85(3)
logit-T	94(9)	92(11)	-(-)	80(6)	79(6)
LIMMA	94(7)	93(7)	32(9)	76(9)	75(9)
SAM	94(8)	93(8)	32(5)	74(11)	74(10)
Shrink-t	94(10)	93(9)	32(4)	75(10)	73(11)
PPLR	88(12)	90(12)	-(-)	71(12)	69(12)
t-test	85(13)	86(13)	25(11)	57(13)	52(13)
# of genes	12626	22029	11475	12626	12626
# of spikes	16	42	1331	11	11
# of groups	20	14	2	12	10

to the description in [40], and the BGX method [17] as implemented in the R-package bgx.

Due to long computer run times the comparison with the BGX method is restricted to the Gene Logic AML data set using a subset of probe-sets only (the run time for one single analysis of 6 arrays with all 12626 probe-sets is more than 24 hours). The subset of size 1011 consists of probe-sets number 6000-7002 (excluding 6030, 6367, and 6463) together with the 11 spiked probe-sets and the same subset was used in [17]. The probe-set numbering is as obtained when loading data into R using the R-package affy.

For each spike-in data set and method ROC-curves were calculated. Also, for the analysis using a complete set of probe-sets, the area (AUC) under the ROC curve up to 25, 50, 100 and 200 false positives was computed. In the comparison with BGX using only 1011 probe-sets, AUC was computed up to 2, 4, 8 and 16 false positives in order to cover the same false positive range as for the complete probe-set comparisons.

ROC curves for a subset of the compared methods are found in Figure 3 and AUC values up to 100 false positives from the complete probe-set analysis are found in Table 1 (ROC curves for all methods and AUC up to 25, 50, 200 false positives are available as supplementary material. See additional file 2: Supplement2.pdf and file 3: Supplement2.pdf, respectively) Overall, three of the four methods taking the variability-to-intensity-level dependency into account (PLW, LMW and LPE)

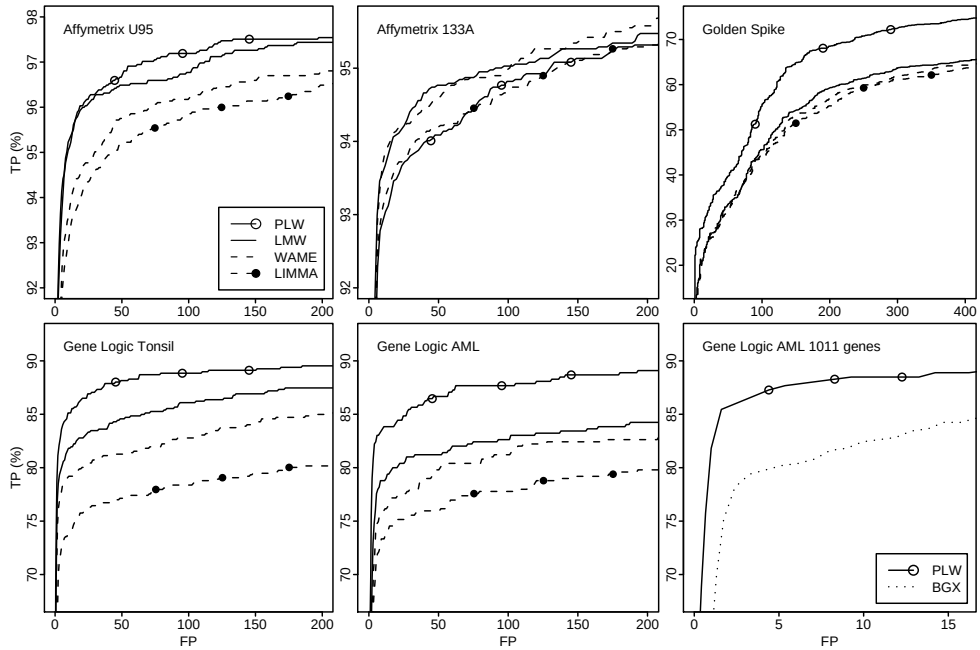

Figure 3: ROC curves for a subset of the compared methods. The horizontal axis shows the number of false positives (FP) and the vertical axis the proportion of true positives found (TP).

performed better than the other methods, with the proposed method PLW having the highest AUC on four of the five data sets. The fourth method taking the variability-to-intensity-level dependency into account (IMBT) performed comparably well on the Affymetrix and Golden Spike data sets but less so on the two Gene Logic data sets. Ranking genes with respect to FC performs quite well on the Affymetrix U133A and the two Gene Logic data sets but not on the other two data sets. Among the penalized and moderated t-test methods, WAME and Efron-t consistently perform better than the other ones. However, the difference between these methods for the two Affymetrix Latin Square and the Golden Spike data sets are small, compared to the difference in AUC obtained using the two Gene Logic data sets. Thus, the two Gene Logic data sets appear slightly different from the other three. The PPLR method based on the multi-mgMOS model [18] was ranked as number eleven with only the ordinary t-test having lower AUC values.

The ROC curves obtained using the subset of 1011 probe-sets from the Gene Logic AML data set are found in the lower right panel of Figure 3. The PLW method shows consistently higher true positive rate compared with BGX and the AUC up to 8 false positives (scaled so that optimum is 100) is 84 and 75 for PLW and BGX, respectively.

The second proposed method LMW differs from existing moderated and penalized t-test in that the global variance estimator (which gene-specific estimators are

adjusted towards) varies with intensity-level. Actually this is the only difference between LMW and the WAME method. The LPE method also uses a global variance estimator that varies with intensity-level. But opposed to using a weighted mean of the global and gene-specific estimator, only the global estimator is used in the denominator of the LPE statistic. Thus for genes with similar intensity-level, LPE is basically identical to ranking using fold change. Hence, since LMW consistently performs better than WAME, and LPE has higher AUC than fold change in four of the five data test, modeling the global variance estimator as a function of intensity is worthwhile doing. Further, having in mind that the GCRMA and MAS5 expression indexes showed a clear dependency between variability and intensity-level in Figure 1, whereas the RMA expression index only showed a weak dependency, this kind of variance modeling might be even more important in analysis based on GCRMA or MAS5 expression indexes.

Also, Figure 2 shows that the false discovery rate obtained by adjusting towards a global estimate that varies with intensity-level results in a much more stable false discovery rates compared to using a (truly) global estimate.

Both logit-t and PLW do inference on background corrected and normalized logged PM intensities resulting in multiple statistics which are then summarized into one by the median statistic for each probe-set, in contrast to first summarizing PM intensities and then doing inference. With this being the only difference between PLW and LMW, and one of the differences between logit-t and the ordinary t-test using RMA expression indexes (they also use different background correction and normalization methods), we find that computing statistics for each PM probe and then summarizing shows better performance compared to the other option.

More complicated models often come with the prize of longer computer run times. Of the methods evaluated the BGX model and the PPLR method together with the multi-mgMOS model are the most computer intense ones. The computer run time for one single two group analysis of 3+3 HG-U95A arrays with data from 12626 genes is more than 24 hours with BGX and 1.5 hours for PPLR+multi-mgMOS (using the recommended EM method of PPLR) on a 2.2 GHz AMD Opteron machine. The corresponding time (including pre-processing of PM and MM data) is 2-3 minutes for PLW and 9 seconds for the moderated t-test in LIMMA.

Conclusions

We have presented two new methods for ranking genes with respect to differential expression: Probe level Locally moderated Weighted median-t (PLW) and Locally Moderated Weighted-t (LMW). Both methods perform very well compared to existing methods with PLW having the most accurate ranking of regulated genes in four out of five examined spike-in data sets. With LMW we show that introducing an intensity-level dependent scale parameter for the prior distribution of the gene-specific variances improves the performance of the moderated t-test. Also, compared to the moderated t-statistic, LMW shows a much more stable false discovery rate across intensity-levels when used on MAS5 expression indexes. In the PLW method inference is performed directly on logged PM intensities and the median of the resulting moderated t-statistics for each probe-set is used to find differentially expressed genes. Overall the PLW method performs better than all compared methods and

thus probe-level inference appears to be preferable over the standard approach using gene expression indexes for inference.

Authors contributions

MÅ provided the initial idea, formulated the model, derived and programmed the estimation procedure, performed the analysis of real and simulated data, and drafted the manuscript. All authors finalized and approved the final version of the manuscript.

Acknowledgements

The research was supported by the Gothenburg Mathematical Modeling Center and the Gothenburg Stochastic Centre.

References

1. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes**. *Bioinformatics* 2001, **17**(6):509–519.
2. Lönnstedt I, Speed TP: **Replicated microarray data**. *Statistica Sinica* 2002, **12**:31–46.
3. Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments**. *Stat. Appl. Genet. Mol. Biol.* 2004, **3**:article 3.
4. Kristiansson E, Sjögren A, Rudemo M, Nerman O: **Weighted Analysis of Paired Microarray Experiments**. *Stat. Appl. Genet. Mol. Biol.* 2005, **4**:article 30.
5. Kristiansson E, Sjögren A, Rudemo M, Nerman O: **Quality optimised analysis of general paired microarray experiments**. *Stat. Appl. Genet. Mol. Biol.* 2006, **5**:article 10.
6. Sartor MA, Tomlinson CR, Wesselkamper SC, Sivaganesan S, Leikauf GD, Medvedovic M: **Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments**. *BMC Genomics* 2006, **7**:article 538.
7. Sjögren A, Kristiansson E, Rudemo M, Nerman O: **Weighted analysis of general microarray experiments**. *BMC Bioinformatics* 2007, **8**:article 387.
8. Åstrand M, Mostad P, Rudemo M: **Improved covariance matrix estimators for weighted analysis of microarray data**. *J. Comput. Biol.* 2007, **Accepted**.
9. Opgen-Rhein R, Strimmer K: **Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach**. *Stat. Appl. Genet. Mol. Biol.* 2007, **6**:article 9.
10. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *PNAS* 2001, **98**(9):5116–5121.
11. Efron B, Tibshirani R, Storey JD, Tusher V: **Empirical Bayes Analysis of a Microarray Experiment**. *J. Amer. Statist. Assoc.* 2001, **96**:1151–1160.

12. Eaves IA, Wicker LS, Ghandour G, Lyons PA, Peterson LB, Todd JA, Glynne RJ: **Combining Mouse Congenic Strains and Microarray Gene Expression Analyses to Study a Complex Trait: The NOD Model of Type 1 Diabetes**. *Genome Res.* 2002, **12**(2):232–243.

13. Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee JK: **Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays**. *Bioinformatics* 2003, **19**(15):1945–1951.

14. Comander J, Natarajan S, Gimbrone M, Garcia-Cardena G: **Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation**. *BMC Genomics* 2004, **5**:17.

15. Hu J, Wright FA: **Assessing Differential Gene Expression with Small Sample Sizes in Oligonucleotide Arrays Using a Mean-Variance Model**. *Biometrics* 2007, **63**:41–49.

16. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249–264.

17. Hein AM, Richardson S, Causton HC, Ambler GK, Green PJ: **BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data**. *Biostatistics* 2005, **6**(3):349–373.

18. Liu X, Milo M, Lawrence ND, Rattray M: **A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips**. *Bioinformatics* 2005, **21**(18):3637–3644.

19. Liu X, Milo M, Lawrence ND, Rattray M: **Probe-level measurement error improves accuracy in detecting differential gene expression**. *Bioinformatics* 2006, **22**(17):2107–2113.

20. Munson P: **A 'consistency' test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations.** *Gene Logic Workshop of Low Level Analysis of Affymetrix GeneChip Data* 2001.

21. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression**. *Bioinformatics* 2002, **18**:S96–104.

22. Durbin BP, Hardin JS, Hawkins DM, Rocke DM: **A variance-stabilizing transformation for gene-expression microarray data**. *Bioinformatics* 2002, **18**:S105–S110.

23. Durbin BP, Rocke DM: **Estimation of transformation parameters for microarray data**. *Bioinformatics* 2003, **19**(11):1360–1367.

24. Geller SC, Gregg JP, Hagerman P, Rocke DM: **Transformation and normalization of oligonucleotide microarray data**. *Bioinformatics* 2003, **19**(14):1817–1823.

25. Tukey JW: *Exploratory Data Analysis*. Addison-Wesley 1977.

26. Holder D, Raubertas RF, Pikounis VB, Svetnik V, Soper K: **Statistical analysis of high density oligonucleotide arrars: a SAFER approach.** *Gene Logic Workshop of Low Level Analysis of Affymetrix GeneChip Data* 2001.

27. Rocke DM, Durbin B: **Approximate variance-stabilizing transformations for gene-expression microarray data**. *Bioinformatics* 2003, **19**(8):966–972.

28. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning, Volume 1*. Springer, first edition 2001.

29. Dempster AP, Laird NM, Rubin DB: **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *J. Roy. Statist. Soc. Ser. B* 1977, **39**:1–38.

30. Åstrand M: **plw: An R implementation of Probe level Locally moderated Weighted median-t (PLW) and Locally Moderated Weighted-t (LMW)**[http://www.math.chalmers.se/~astrandm].

31. **Gene Expression Omnibus repository**[http://www.ncbi.nlm.nih.gov/geo/].

32. Spira A, Beane J, Pinto-Plata V, Kadar A, Liu G, Shah V, Celli B, Brody JS: **Gene Expression Profiling of Human Lung Tissue from Smokers with Severe Emphysema**. *Am. J. Respir. Cell Mol. Biol.* 2004, **31**(6):601–610.

33. Stearman RS, Dwyer-Nield L, Zerbe L, Blaine SA, Chan Z, Bunn PAJ, Johnson GL, Hirsch FR, Merrick DT, Franklin WA, Baron AE, Keith RL, Nemenoff RA, Malkinson AM, Geraci MW: **Analysis of Orthologous Gene Expression between Human Pulmonary Adenocarcinoma and a Carcinogen-Induced Murine Model**. *Am J Pathol* 2005, **167**(6):1763–1775.

34. **Affymetrix U95 and 133A Latin Square spike-in data sets**[http://www.affymetrix.com/support/datasets.affx].

35. Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP: **A benchmark for Affymetrix GeneChip expression measures**. *Bioinformatics* 2004, **20**(3):323–331.

36. **Gene Logic spike-in data sets**[http://www.genelogic.com/newsroom/studies/].

37. Choe S, Boutros M, Michelson A, Church G, Halfon M: **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset**. *Genome Biology* 2005, **6**(2):R16.

38. Irizarry R, Cope L, Wu Z: **Feature-level exploration of a published Affymetrix GeneChip control dataset**. *Genome Biology* 2006, **7**(8):404.

39. Åstrand M: **Contrast Normalization of Oligonucleotide Arrays**. *J. Comput. Biol.* 2003, **10**:95–102.

40. Lemon W, Liyanarachchi S, You M: **A high performance test of differential gene expression for oligonucleotide arrays**. *Genome Biology* 2003, **4**(10):R67.