THESIS FOR THE DEGREE OF DOCTOR OF ENGINEERING

Weighted Analysis of Microarray Experiments

Anders Sjögren

CHALMERS | GÖTEBORG UNIVERSITY

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University
Göteborg, Sweden, 2007

Weighted Analysis of Microarray Experiments
Anders Sjögren
ISBN 978-91-7291-939-6

Doktorsavhandlingar vid Chalmers Tekniska Högskola
Ny serie Nr 2620
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

Cover: The measured ratio of before and after treatment in the Polyp study, for 5 patients on $\log_2$ scale. Estimated correlations and variances in the proposed WAME method in blue. Weights for the different arrays in the estimate for differential expression on the diagonal. Current methods generally assume that the variation of each array is equal and that no arrays share sources of variation, i.e. that they are uncorrelated. This implies that the clouds should essentially be circular with equal size and with no or equal diagonal trends, which is clearly not the case. See Paper I and the legend of Figure 2 for further details.

Printed at the Department of Mathematical Sciences
Göteborg, Sweden, 2007

Abstract

DNA microarrays are strikingly efficient tools for analysing gene expression for large sets of genes simultaneously. The aim is often to identify genes which are differentially expressed between some studied conditions, thereby gaining insight into which cellular mechanisms are differently active between the conditions. In the measurement process, several steps exist that risk going partly or entirely wrong and quality control is therefore crucial.

In Paper I-III, a novel method is developed which integrates quality control quantitatively into the analysis of microarray experiments. The noise structure for each gene is modelled by (i) a global covariance structure matrix catching decreased quality by array-wise variances and catching shared sources of variation by correlations, and (ii) gene-wise variance scales having a prior distribution with parameters estimated from the data of all genes in an empirical Bayes manner. The variances and correlations are entirely estimated from the data. In the estimates and tests for differential expression, arrays with lower precision or arrays sharing sources of variation are downweighted. Thus, the sharp decision of entirely excluding arrays is avoided. The method is called Weighted Analysis of Microarray Experiments (WAME).

Current methods for microarray analysis generally disregard the quality variations. Simulations based on real data show that this often results in severely invalid p-values. Trusting such p-values therefore risks resulting in false biological conclusions. WAME gives increased power and valid p-values when few genes are differentially expressed and conservative p-values otherwise. Similar results are seen on simulations according to the model.

In Paper IV, WAME is used to identify genes which are differentially expressed between small and large human fat cells. WAME here successfully downweights one array that was suspected of decreased quality on biological grounds.

The WAME method is freely available as a add-on package for the R language.

List of publications

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

Paper I E. Kristiansson*, A. Sjögren*, M. Rudemo and O. Nerman. Weighted Analysis of Paired Microarray Experiments, *Statistical Applications in Genetics and Molecular Biology*: 4(1), Article 30, 2005.

Paper II E. Kristiansson, A. Sjögren, M. Rudemo and O. Nerman. Quality Optimised Analysis of General Paired Microarray Experiments, *Statistical Applications in Genetics and Molecular Biology*: 5(1), Article 10, 2006.

Paper III A. Sjögren, E. Kristiansson, M. Rudemo, and O. Nerman. Weighted analysis of microarray experiments, *submitted to BMC Bioinformatics*, 2007.

Paper IV M. Jernås, J. Palming, K. Sjöholm, E. Jennische, P.A. Svensson, B.G. Gabrielsson, M. Levin, A. Sjögren, M. Rudemo, T.C. Lystig, B. Carlsson, L.M.S. Carlsson, M. Lönn. Separation of human adipocytes by size - hypertrophic fat cells display distinct gene expression, *The Federation of American Societies for Experimental Biology Journal*: 20(9), 1540-1542, 2006.

* authors contributed equally, order was randomised

Table of contents

Acknowledgements

I would like to thank my advisors Mats Rudemo, Lena Carlsson and Olle Nerman for sharing their experience, knowledge and intuition. Furthermore, I would like to thank my PhD-student colleagues Erik Kristiansson and Louise Olofsson for the close collaborations in the statistical and biological directions, respectively.

Thanks to my colleagues and collaborators at the departments of Mathematical Statistics and Internal Medicine for inspiring suggestions and discussions, especially Gilles Guillot, John Gustafsson, Alexandra Jauhiainen, Margareta Jernås, Anna Larsson, Bob Olsson, Janeli Sarv and Magnus Åstrand.

The National Graduate School in Genomics and Bioinformatics funded this work, for which I am grateful. Thanks to Anders Blomberg for the time and efforts put into making the school active.

Finally, I would like to thank my wife Camilla and the rest of my family for their unfailing love and support during the intense last few years.

Background

On DNA microarray experiments

Inside each of the cells of organisms ranging from humans to trees and yeast, the DNA contains blueprints of a vast array of proteins. To produce a particular protein, a machinery referred to as the *central dogma* of molecular biology is invoked (Alberts et al., 1998). Here, a relevant part of the DNA (called a gene) is first essentially copied into messenger RNA (mRNA) molecules. The mRNA molecules then carry the information to the ribosomes, which produce the protein in question. The abundance of the mRNA of a gene is called the expression of that gene.

The proteins are the main working horses for a wide array of functions of the cell, including structural elements and hormones performing between-cell signalling. Therefore, monitoring the production of proteins is a way to gain insight into the inner mechanisms of the cells. Such insight could ideally help understanding diseases on a detailed level, providing the foundation of subsequent development of efficient drugs. A common problem is therefore to try and identify genes that are differentially expressed between some studied conditions, that is genes which tend to have higher abundance of mRNA in cells from one condition than from the other.

DNA microarrays is a technique for measuring the abundance of mRNA for a large set of predefined genes (typically 1000-40000), which is performed for relatively few biological samples (typically six to one hundred) in each experiment. The microarrays thereby provide snapshots of how active different mechanisms are in the measured cells at the time of the measurements.

The actual microarrays are stamp-sized plates where different spots are prepared to specifically match different genes. There are different types of microarrays, the most common being two-colour spotted cDNA arrays and one channel *in-situ* hybridised oligonucleotide arrays. From the mRNA molecules, cDNA or cRNA which is labelled with fluorescent molecules is prepared. When the cDNA or cRNA is hybridised to a microarray, the labelled molecules stick to their predefined spots. When a laser then sweeps over the surface, the amount of light emitted by the fluorescent molecules at different positions of the array can be used to measure the number of attached labelled molecules, signifying the expression of the corresponding genes.

The process of performing a microarray measurement involves a long

series of step before the data can be analysed. For an array in a patient study (i) a patient is selected, (ii) a biopsy (a piece of the tissue of interest) is taken, (iii) the RNA is extracted from the biopsy, (iv) cRNA or cDNA marked with fluorescent markers is prepared from the RNA, (v) the cRNA or cDNA is hybridised to the array and non-attached molecules are washed off, (vi) the array is scanned using a laser and photodetector and, (vii) the scanned data is preprocessed in the computer, e.g. background noise is removed and other artefacts are removed by suitable normalisation. Each of these steps in turn involves substeps. Frequently, steps are performed in batches where a number of samples are prepared in parallel.

Unfortunately, each of the involved steps risks going partly or entirely wrong, affecting individual samples or batches of samples. For example, the biopsy might contain a non-representative cell-type distribution, the RNA can become degraded during the handling of the biopsy, or some lab conditions might make some lab step work non-optimally for some samples. Thus there is a strong need for quality assessment and quality control to handle occurrences of poor quality, as is clearly pointed out in Johnson and Lin (2003) and Shi et al. (2004)

Currently, the main quality assurance solution is to try and identify quality deviations in each step and exclude or remake samples with suspected decreased quality. A short review of such methods are available in Paper I. However, even utilising an optimal quality control procedure aiming at removing low quality samples and/or individual gene measurements (e.g. spots), there will always be a marginal region with some measurements being of decreased quality without being worthless. Therefore quality control becomes a balancing act of not including samples that would introduce too much noise into the subsequent analysis, but at the same time not excluding samples that would otherwise have provided additional strength to the analysis.

Adding to the problem is the fact that some sources of the quality deviations might be shared among subsets of the arrays, e.g. when they are performed in parallel in the lab or when they share some inherent characteristics, such as when some biopsies have a similar but non-representative cell-type distribution. As will be pointed out in the next section, most current methods for analysing the data do not take quality variations in the data into account.

Statistical analysis of microarray experiments

An important class of problems problem related to microarray experiments is to try and identify genes that are differentially expressed between some studied conditions, that is to try and find genes which have higher typical abundance of mRNA in cells from one condition than from the other. Here, the conditions might for example refer to different treatments of patients or strains of yeast. To formalise this problem, a differentially expressed gene is defined as a gene with higher expected abundance in samples from one condition compared to the other. An interpretation of the expected abundance in a condition is the average abundance in all possible biological samples from the condition. Since all possible samples cannot be measured, a few samples are randomly selected from each condition and they are measured with DNA microarrays. Statistics is then used to draw conclusions on the difference in expression between the entire populations of possible biological samples.

A statistical model is often first assumed for how the measurements are distributed. For each gene, a function of the data (a statistic) is formed, which is designed to have an extreme value when the data suggests a differential expression. The statistic can then be used to test for differential expression for the different genes. The distributional assumptions are then used to determine the critical value for the statistic, that is a value of the statistic which is improbable to be exceeded if the gene is in fact non differentially expressed. This means that if the statistic for a gene exceeds the critical value, the test rejects that the gene is not differentially expressed and the gene is called significantly differentially expressed.

In the models below, a transformed version of the abundance is used, where the base two logarithm ($\log_2$) is taken for each value. This transformation has the property that a two-fold increase or decrease in the original measurement will result in an addition or subtraction, respectively, of one to the $\log_2$-value. Thus increases and decreases are treated symmetrically. In addition, the $\log_2$ values better fit standard statistical models.

The most common statistical model is the Ordinary Linear Model (Arnold, 1980) which give rise to the ordinary t- and F-tests. In this case the model implies that for each fixed gene, the measurements are normally distributed with expected values according to the respective conditions they come from. Furthermore, for each gene the noise (i.e. the random deviations from the expected values) of the different arrays are assumed to be independently distributed and to have equal variances. The likelihood ratio statistic then becomes the ordinary t-statistic which is computed as

the ratio between the estimated differential expression and the estimated standard deviation of that difference. A large absolute value of the t-statistic thus means that the difference is larger than can be explained by chance if the gene is in fact non-differentially expressed.

In microarray experiments there are often relatively few arrays, resulting in highly variable estimates of the standard deviation for each gene. To use the information in the large number of measured genes to handle this problem, an empirical Bayes approach (Robbins, 1956; Maritz, 1970) can be taken. Here, the variances of the genes are assumed to be inverse gamma distributed *a-priori*, i.e. before data from the gene is observed. The two parameters of the inverse gamma distribution are then estimated from the data of all genes. This approach has been used by Baldi and Long (2001), Lönnstedt and Speed (2002) and in the R (R Development Core Team, 2006) package LIMMA (Smyth, 2004). Essentially, the parameters are (i) a global typical variance for all genes and (ii) the variability of the variances for different genes. Now, the estimate of the variance for each gene becomes a compromise between the ordinary gene-wise estimate and the estimate of the global variance. Here, more emphasis in placed on the global variance estimate when there is a low estimated variability of the variances for different genes, i.e. when the variances tend to be similar for different genes. The corresponding statistics are called moderated t- or F-statistics, since the ordinary variance estimates are moderated by the global estimate in the statistics.

Full hierarchical Bayes models have been presented, e.g by Lönnstedt and Britton (2005). However, Lönnstedt and Britton (2005) note that the existing empirical Bayes methods have at least as good performance as their novel models, and that the long computation times for the full hierarchical Bayes methods not seem worth spending.

A statistic similar to the moderated t-statistic was proposed by Efron et al. (2001) and Tusher et al. (2001), but without a underlying distributional model. Instead a permutation based procedure is used to estimate the distribution of the statistic for non-differentially expressed genes. The method is thus non-parametric and the underlying assumption is that the arrays are exchangeable for non-differentially expressed genes. This procedure is available in the SAM plugin for Microsoft Excel.

The empirical Bayes approach above was extended by Ritchie et al. (2006), introducing array-specific variances to model array-wide quality deviations. It can be noted that this paper appeared after Paper I, which introduced a more general model for paired designs.

In some microarray experiments, additional information is available. For example, shared sources of variation may be known or quantitative quality measures may be available, e.g. spot shape features or residuals from the fitting of probe-level models (Bolstad, 2004). It is possible to explicitly model some such sources of variation, for example using random or fixed effects (cf. Bakewell and Wit (2005)) and to include quality measures as covariates. However, such models would likely focus on some of the clearer sources of variation but leave out more involved and hard modelled sources.

In the parametric models above the noise of the different arrays is assumed to be independent and/or to have equal variances. This essentially implies that the quality of each array is identical and/or that the random deviations of each array has nothing to do with the deviations of any other arrays. The exchangeability assumption in the non-parametric methods similarly implies that the quality of each array is identical and no two arrays are more similar than any other two. However, it was previously argued that the qualities of the different arrays are expected to vary. Also, some sources of the quality deviations are shared among subsets of the arrays and their random deviations from the expected value will be partly connected. Thus the assumptions of the models seem questionable.

If the assumptions of the model used do not hold, two problems occur. First, the statistic is designed from the model assumptions to give as good inference results as possible. If the assumptions fail, one can expect to lose power in the analysis. Second, when deriving p-values and the critical value for the statistic, the model assumptions are used. When the model assumptions fail, the test results and p-values might therefore be misleading, potentially leading to erroneous biological conclusions.

Contributions of the thesis

The suspicion that data from different arrays coming from different biological samples have unequal quality is verified by examination of microarray experiments coming from several different experimental designs and microarray techniques. The model assumptions of the current methods are thus found to be invalid in many cases.

An alternative analysis method is therefore proposed, where the noise of different arrays is not assumed independent or to have equal variances. Instead the model is a generalised linear model where each array has a variance of its own and each pair of arrays have a correlation. Each

gene then has a factor which scales the array-wise variances. This factor models the variation in variance for different genes and is assigned an inverse gamma prior distribution. The model is thus a generalisation of the one from the empirical bayes method in LIMMA.

The method is called Weighted Analysis of Microarray Experiments (WAME) since an array is downweighted when it has high variance and/or when it is positively correlated with other arrays, making them partially contain the same information. Since the arrays are weighted according to their quality, the sharp and sometimes subjective choice is avoided of entirely excluding arrays with suspected decreased quality.

An advantage of the method is that the quality is entirely determined from the actual data. Thus, array-wide quality deviations from all steps are considered, which is in contrast to other quality assurance techniques which only regard one or a few steps in experiments with a certain type of microarrays. An addition, it makes it applicable to different DNA microarray techniques, e.g. both *in-situ* hybridised one-channel arrays and two-colour spotted cDNA arrays.

For several examined real experiments, the difference in test results and p-values are significant between WAME and the traditional models, where the traditional models suggest a much larger number of differentially expressed genes. Resample-based simulations based on the real data are performed, where the noise is real and the signal added. Here, the p-values of WAME are valid when few genes are highly differentially expressed and conservative otherwise. The commonly used methods instead give p-values which can be highly optimistic or conservative, depending on how the quality deviations of the arrays happen to occur. Consequently, p-values and derived entities, such as estimates of the number of differentially expressed genes, from the standard methods are questionable to use without careful model verifications.

It has recently been argued that the expression of different genes are highly dependent, making the law of large number normally inapplicable (Klebanov and Yakovlev, 2006) and e.g. the standard false discovery rate estimator imprecise (Pawitan et al., 2006). However, since array-specific quality deviations are not modelled in the underlying statistics being examined, the statistics of the different genes will share a common bias. The observed strong dependencies might thus be due to failed model assumptions rather than due to the nature of microarray data *per se*, e.g. through substantial long-range gene-gene interactions. This is further discussed in Paper III.

Using similar simulations where the noise is taken from resamples of real data and a synthetic differential expression is added to randomly selected genes. WAME is then found to have higher power of identifying the differentially expressed genes than the commonly used methods. Similar results are noted on fully parametric simulations.

The model is basically constructed in a stepwise manner in Paper I-III, where Paper I treats two-condition experimental designs with pairwise measurements, Paper II treats general experimental designs with pairwise measurements and Paper III treats unpaired experimental designs.

In Paper IV, WAME is used to identify genes that are differentially expressed between small and large human adipocytes (fat cells). The biological importance is that the risk of metabolic complication, including type 2 diabetes and cardiovascular disease, is increased not only by the amount and location of adipose (fat) tissue, but also by the size of the fat cells. The active mechanisms in cells of different size are therefore of interest. WAME here successfully downweights one array that was suspected of decreased quality on biological grounds and would otherwise have been entirely excluded.

Summary of papers

Summary of Paper I : *Weighted Analysis of Paired Microarray Experiments*

Since several consecutive steps are involved before analysable data is obtained in microarray experiments, the data of the different arrays are suspected to have different quality. Three real datasets with paired two-sample experimental designs are examined, the *Swirl* data by Dudoit and Yang (2003), the *Polyp* data by Benson et al. (2004), and the *Cardiac* data by Hall et al. (2004). All datasets reveal strong deviations in quality between the different arrays, or pairs of one-channel arrays in case of one-channel microarrays (see Figure 1-3). As expected, the different arrays have different variability and, in addition, correlations between the noise of certain arrays are evident.

A model called WAME is therefore formulated for paired two-sample experiments, where the quality deviations are modelled by different variances for different arrays as well as by correlations between them in a covariance matrix, catching both unequal precision and shared sources of variation. Genes have different variability (biological and technical), which is modelled by a gene-specific variance scaling factor with an inverse gamma prior distribution. Given this structure, the pair-wise measured log_2-ratios for each gene are assumed to be normally distributed.

The model can be summarised as follows: For each gene g one $\log_2$ ratio is observed for each of the n pairs,

$$\mathbf{X}_g = (X_{g1}, \ldots, X_{gn}) \, .$$

Further, μ_g denotes the differential expression, Σ is the $n \times n$ covariance matrix catching the quality deviations, c_g is the gene-specific variance scaling factor, and α is a hyperparameter determining the spread of the prior distribution for c_g. Then for fixed μ_g, Σ and α,

$$\begin{aligned} c_g &\sim \Gamma^{-1}(\alpha, 1) \text{ and} \\ \mathbf{X}_g \mid c_g &\sim \mathrm{N}_n \left(\mu_g \mathbf{1}, c_g \Sigma\right) . \end{aligned} \tag{1}$$

For two-sample paired designs, this model is a generalisation of the widely used empirical Bayes model proposed by Lönnstedt and Speed (2002), which is implemented in the LIMMA package (Smyth, 2004). The models are equivalent when no correlations or unequal variances exist in Σ.

The contribution of the WAME model is thus the modelling of the array-specific qualities.

The estimation of the covariance matrix Σ is complicated for a number of reasons. Primarily, Σ and μ_g cannot be maximum likelihood estimated simultaneously, since there are trivial solutions that give infinite likelihood (e.g. when the expected value is set to the measurement of one array and the corresponding variance is set to zero). To circumvent this problem, it is temporarily assumed than no differential expressed genes exist, and thus that only noise is measured. A likelihood based estimator of Σ and α is thus proposed. Since the estimates will be based on thousands of genes, they are expected to be precise and are treated as known in the gene-wise inference below.

The main goal with the inference is to examine the differential expression μ_g. The gene-wise maximum likelihood estimate of μ_g is derived to be a weighted mean

$$\hat{\mu}_g = \mathbf{w}^{\mathrm{T}} \mathbf{X}_g$$

where

$$\mathbf{w}^{\mathrm{T}} = \frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}} \, .$$

Thus, arrays having increased variance or being positively correlated with other arrays are downweighted. The likelihood ratio test for identifying differentially expressed genes can be realised by a weighted moderated t-statistic

$$T_g = \sqrt{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}\,(n-1+2\alpha)}\;\frac{\hat{\mu}_g}{\sqrt{S_g+2}} \, ,$$

where S_g is a residual sum of squares like function of $\mathbf{X}_g$ taking Σ into account. When the gene is not differentially expressed, T_g becomes t-distributed with $n-1+2\alpha$ degrees of freedom.

Since the covariance structure is entirely general and since it is estimated entirely from the data, array-wide quality deviations in all steps (biological and technical) are objectively incorporated into the analyses.

A simulation study is performed where data is simulated according to the WAME model. Some widely used alternative methods are included for comparison: the ordinary t-statistic, the moderated t-statistic in LIMMA, Efron's penalized t-statistic (Efron et al., 2001) and ranking by fold-change, i.e. the average over all arrays. When correlations and/or different variances are included, receiver operating characteristic (ROC) curves reveal that WAME has the highest power. When no correlations

exist and the variances are equal, LIMMA and WAME performs approximately equally well. In all cases, the moderated statistics outperform the traditional methods, i.e. the t-test or the fold-change.

The point estimator of the covariance structure matrix Σ is evaluated on data simulated according to the WAME model with and without differentially expressed genes, as well as on data with heavier tails. The estimator turns out to be precise in all cases. When differentially expressed genes exist, the differential expression is interpreted as a correlated noise by the estimator and the correlations and variances are biased upwards. Otherwise, the estimator is accurate.

On the real datasets, distinctly non-equal variances and non-zero correlations are estimated, which corresponds well to visual inspection of plots of the data. The corresponding weights for the different arrays in the estimate of the differential expression become highly unequal. In the *Polyp* data, one of the arrays was suspected in advance to be of decreased quality on biological grounds and was excluded in the original publication. That array was significantly downweighted by WAME. Evaluation of the model on the real datasets by comparison of the empirical and theoretical distributions suggests that the inverse gamma family of prior distribution for c_g is flexible enough and that the t-statistic is t-distributed as suggested by the model. In the *SWIRL* data, the conditions are well controlled and it is believed that very few differentially expressed genes exist. Here, it is noted that the observed distribution of the t-statistics from LIMMA and WAME are somewhat unequal, where WAME follows the theoretical null-distribution well while LIMMA provides t-statistics which seem too optimistic.

Summary of Paper II : *Quality Optimised Analysis of General Paired Microarray Experiments*

In Paper I, experiments with paired samples coming from two conditions were examined. Now, WAME is extended to handle experiments with paired samples and arbitrary numbers of conditions. This is performed by generalising the model (1) to a generalised linear model (Arnold, 1980), being a generalisation of the LIMMA model (Lönnstedt and Speed, 2002; Smyth, 2004) by adding the covariance structure matrix Σ between the arrays. In Paper I, the differential expression was directly defined by the expected value of the measurements. Now, for each gene g a set of

parameters $\boldsymbol{\gamma}_g$ are introduced, typically being the expected value for each experimental condition. The expected value vector $\boldsymbol{\mu}_g$ of the arrays and the differential expression $\boldsymbol{\delta}_g$ are then defined as linear combinations of $\boldsymbol{\gamma}_g$ by a matrix C and a design matrix D. The model becomes

$$\begin{aligned} c_g &\sim \Gamma^{-1}(\alpha, 1) \text{ and} \\ \mathbf{X}_g \mid c_g &\sim \mathrm{N}_n\left(\boldsymbol{\mu}_g, c_g\Sigma\right), \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_g = D\,\boldsymbol{\gamma}_g$ and the differential expression is defined by $\boldsymbol{\delta}_g = C\,\boldsymbol{\gamma}$.

The covariance matrix Σ and the hyperparameter α are essentially estimated as in Paper I, again temporarily assuming $\boldsymbol{\mu}_g = 0$ for all genes. This assumption is valid in the paired setting if no genes are differentially expressed between any pairwise measured conditions. The estimator can be expected to give results even if a few genes are differentially expressed between some conditions.

The estimator of $\boldsymbol{\delta}_g$ and the test for differential expression are derived by first transforming the data by multiplication by the inverse of the square root matrix of Σ, $\Sigma^{-1/2}$,

$$\tilde{\mathbf{X}}_g = \Sigma^{-1/2}\,\mathbf{X}_g \ .$$

Letting $\tilde{D} = \Sigma^{-1/2}D$ and $\tilde{\boldsymbol{\mu}}_g = \tilde{D}\,\boldsymbol{\gamma}_g$, properties of the normal distribution gives

$$\tilde{\mathbf{X}}_g \mid c_g \sim \mathrm{N}_n\left(\tilde{\boldsymbol{\mu}}_g, c_g\,\mathrm{I}\right) \tag{3}$$

where the identity matrix is denoted by I. Thus standard data is obtained which is independent and identically distributed between the arrays.

The resulting likelihood ratio test statistic becomes a weighted moderated F-statistic (or t-statistic when C has one row) and the maximum likelihood estimator of the differential expression vector becomes a vector of weighted averages,

$$\hat{\boldsymbol{\delta}}_g = C(D^{\mathrm{T}}\Sigma^{-1}D)^{-}D^{\mathrm{T}}\Sigma^{-1}\,\mathbf{X}_g\,,$$

where the generalised inverse is any matrix satisfying $AA^{-}A = A$.

A time-course experiment is simulated according to the WAME model with moderate correlations between arrays within time points and low to moderate correlations between arrays from different time points. The differential expression vector is here the vector of all changes in expected value between nearby timepoints. The weighted moderated F-statistic thus aims at identifying genes which are differentially expressed between

any timepoints. The results are compared to the ordinary F-statistic, the moderated F-statistic in LIMMA and a weighted moderated F-statistic in LIMMA, which is based on the assumption of unequal variances but no correlations. A substantial advantage in power is noted for WAME. This is expected since the data are simulated according to the WAME model and relevant correlations and differences in variance exist. Still it is hoped to give an indication of the gains on real data.

Two real datasets are examined, the ApoAI dataset (Callow et al., 2000) and the Cardiac dataset (Hall et al., 2004). In both cases, the statistic show a good fit to the null-distribution. *ApoAI* is a well controlled mouse knockout study where few genes are expected to be differentially expressed. Here, the variation of the estimates of differential expressions are lower for WAME, which is beneficial. Similar to in the *SWIRL* study examined in Paper I, the moderated t-statistics from LIMMA are optimistic.

Summary of Paper III : *Weighted analysis of microarray experiments*

In Paper I and Paper II the covariance-structure matrix Σ is estimated using a temporary assumption that $\boldsymbol{\mu}_g = \mathbf{0}$ for most genes, i.e. that the measurements of most genes consist solely of biological and technical noise. For unpaired designs where one-channel microarrays are used to measure gene expression in different conditions, the signal can be decomposed into two parts: (i) an expected expression in a reference condition and (ii) the differential expression, i.e. a difference between the expected expression in each condition compared to in the reference condition. For many such experiments, the differential expression can be assumed zero for most genes. However, the expected expression in the reference condition can in general not be assumed zero for any genes. Therefore, the previous estimator of Σ is not applicable.

In Paper III, the assumptions are relaxed to only assume that most genes are non-differentially expressed, i.e. $\boldsymbol{\delta}_g = \mathbf{0}$. The trick used is to transform the data and consider

$$\mathbf{Y}_g = \mathbf{X}_g - \tilde{\boldsymbol{\mu}}_g^0 \tag{4}$$

where $\tilde{\boldsymbol{\mu}}_g^0$ is a suitable linear estimator of $\boldsymbol{\mu}_g$ which is unbiased under the assumption of no differential expression, $\boldsymbol{\delta}_g = 0$, and which preserves

the estimability of the differential expression $\boldsymbol{\delta}_g$, based only on the transformed data.

In the unpaired designs above, the transformation can simply be a common subtraction for all arrays of an estimate of the expected expression in the reference condition. Since the difference in mean value between the arrays from the different conditions is unchanged by the transformation, the differential expression will still be estimable.

By design, the transformed data will have expected value zero for non-differentially expressed genes. Since the transformation is linear, the transformed data will furthermore be normally distributed conditional on c_g, although with a different design matrix. Therefore, the method presented in Paper II can essentially be applied to the transformed data, including tests and estimates for the differential expression. It is shown that the tests and estimators are in fact unchanged by the transformation, if the covariance-structure matrices for the transformed and untransformed data are known. The difference is that the covariance-structure matrix for the transformed data can be estimated.

To investigate the properties of the new version of WAME, two real datasets are examined, the COPD data by Spira et al. (2004) and the Atrium data by Barth et al. (2005). Relevant unequal variances and correlations are visually observed (see Figure 4-6) and estimated using WAME, which leads to unequal weights for the different arrays.

When p-values are computed using (i) WAME, (ii) the ordinary t-statistic, (iii) the moderated t-statistic and (iv) the weighted moderated t-statistic, the observed distributions of the p-values differ substantially. Here WAME is considerably more conservative. Since the model assumptions seem to fail for the methods not modelling correlations and/or different variances, it is suspected that their p-values are optimistic and invalid. However, it cannot be ruled out that the p-values of WAME are conservative due to the assumption of no differentially expressed genes, which might partly explain the difference in distributions.

To make a closer examination, data is first simulated with real noise and without differentially expressed genes. This is performed by repeatedly selecting two random subgroups of four arrays from within one group in the original data. Since the same condition is samples twice, no differentially expressed genes exist. The different methods are then performed on those groups. Figure 7 shows the empirical p-value distributions for the resampled Atrium data analysed with the four methods. For WAME, the resulting p-value distributions are very close to the expected uniform.

For OLM (the ordinary t-statistic), LIMMA and weighted LIMMA there is a high variability between the p-value distributions and they are in many cases substantially different from the expected uniform. Thus it seems questionable to trust p-values, or derived entities such as estimates of false discovery rates and the number of differentially expressed genes, from the standard methods without careful prior model verification.

To evaluate the power of the methods in the studied datasets, a known differential expression is added to randomly selected genes in the resampled data above. Here, WAME performs best, followed by LIMMA and weighted LIMMA, which in turn perform significantly better than the traditional methods of performing t-tests or ranking by fold-change. Using the same kind of simulated data, it is further suggested that when the proportion of differentially expressed genes increases, the p-values of WAME become conservative, but does not lose power.

In some of the studied datasets, some arrays in certain groups get weights with different signs compared to the other arrays of that group. A random effects model is shown to elucidate why this happens.

It is furthermore noted that previously observed strong correlations between gene-wise statistics might be due to failure of taking array-wide quality deviations into account in the model and not due to the nature of microarray data *per se*.

Summary of Paper IV :
Separation of human adipocytes by size - hypertrophic fat cells display distinct gene expression

Is is known that the risk of metabolic complication, including type 2 diabetes and cardiovascular disease, is increased not only by the amount and location of adipose tissue (fat), but also by the size of the adipocytes (fat cells). Human fat cells can change 20-fold in diameter and several thousand-fold in volume. Enlargement of subcutaneous abdominal adipocytes is associated with insulin resistance and is an independent predictor of type 2 diabetes.

In Paper IV, a simple and accurate procedure is first developed for separating adipocytes from adipose tissue into two populations according to size. A paired two-sample microarray experiment is then performed where the gene expression is measured separately for large and small adipocytes from three patients. Here WAME is used and biologically interesting genes are identified, of which three are further studied: SAA, TM4SF1

and leptin. Using a database containing expression profiles from a large number of human tissues and cell types, the genes are indicated to be highly expressed in adipocytes compared to other tissues. A separate set of biological samples are used to verify the differential expression of the three genes using realtime RT-PCR. Finally, the differential expression on protein level is indicated using immunohistochemistry.

In the microarray analysis, three patients coded as 13, 15 and 16 were included. The biological sample for Patient 13 was strongly suspected to be of decreased quality. For each Patient a ratio between the measured expression of large and small adipocytes was $\log_2$-transformed and used. Figure 8 shows the $\log_2$-ratios for all genes, for each pair of patients. Estimated variances and correlations as well as weights are also shown. Here the data from the array for Patient 13 is found to have roughly 14 times higher variance than the data from the arrays for the other two patients. Nevertheless, the figure shows that genes having a high $\log_2$ ratio in Patient 15 or Patient 16 tend to have a high $\log_2$ ratio also in Patient 13, so the data from Patient 13 appears to contain some signal and thus not to be worthless.

If the WAME model is suitable for the data and the estimates of the variances and correlations are correct, the ordinary unweighted estimate of differential expression (excluding Patient 13) has 9% higher variance than the weighted estimate from WAME. In addition, one extra degree of freedom is obtained in the t-statistic, further increasing the power. If instead Patient 13 is included in the ordinary unweighted estimate, the variance will become 229% higher and the assumptions of the standard methods will clearly fail.

Additional papers

During the graduate studies, the author has in addition coauthored the following papers:

- M. Benson, L.M.S. Carlsson, M. Adner, M. Jernås, M. Rudemo, A. Sjögren, P.A. Svensson, R. Uddman and L.O. Cardell. Gene profiling reveals increased expression of uteroglobin and other anti-inflammatory genes in glucocorticoid-treated nasal polyps, *Journal of Allergy and Clinical Immunology*: 113(6), 1137-1143, 2004.

- B.G. Gabrielsson, L.E. Olofsson, A. Sjögren, M. Jernås, A. Elander, M. Lönn, M. Rudemo and L.M.S. Carlsson. Evaluation of reference genes for studies of gene expression in human adipose tissue. *Obesity Research.* 13(4), 649-652, 2005.

References

B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Robert, and P. Walter. *Essential Cell Biology*. Garland, 1998.

S.F. Arnold. *The Theory of Linear Models and Multivariate Analysis*. John Wiley & Sons, 1980.

D.J. Bakewell and E. Wit. Weighted analysis of microarray gene expression using maximum-likelihood. *Bioinformatics*, 21(6):723–729, 2005.

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.

A.S. Barth, S. Merk, E. Arnoldi, L. Zwermann, P. Kloos, M. Gebauer, K. Steinmeyer, M. Bleich, S. Kaab, M. Hinterseer, H. Kartmann, E. Kreuzer, M. Dugas, G. Steinbeck, and M. Nabauer. Reprogramming of the human atrial transcriptome in permanent atrial fibrillation. *Circulation Research*, 96(9):1022–1029, 2005.

M. Benson, L. Carlsson, M. Adner, M. Jernås, M. Rudemo, A. Sjögren, P.A. Svensson, R. Uddman, and Cardell L.O. Gene profiling reveals increased expression of uteroglobin and other anti-inflammatory genes in glucocorticoid-treated nasal polyps. *Journal of Allergy and Clinical Immunology*, 113(6):1137–1143, 2004.

B.M. Bolstad. *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. PhD thesis, University of California, Berkeley, California, 2004.

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 Research*, 10(12):2022–2029, 2000.

S. Dudoit and J.Y.H. Yang. Bioconductor R packages for exploratory data analysis and normalization of cDNA microarray data. In G. Parmigiani, E.S. Garett, R.A. Irizarry, and S.L. Zeger, editors, *The Analysis of Gene Expression Data*. Springer, 2003.

B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

J.L. Hall, S. Grindle, X. Han, D. Fermin, S. Park, Y. Chen, R.J. Bache, A. Mariash, Z. Guan, S. Ormaza, J. Thompson, J. Graziano, S.E. de Sam Lazaro, S. Pan, R.D. Simari, and L.W. Miller. Genomic profiling of the human heart before and after mechanical support with a ventricular assist device reveals alterations in vascular signaling networks. *Journal of Physiological Genomics*, 17(3):283–291, 2004. URL `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GDS558`.

W. Huber, A. von Heydebreck, and M. Vingron. Analysis of microarray gene expression data. In M. et al. Bishop, editor, *Handbook of Statistical Genetics, 2nd Edition.* John Wiley & Sons, 2003.

K. Johnson and S. Lin. QA/QC as a pressing need for microarray analysis: meeting report from CAMDA'02. *BioTechniques*, 34(suppl):S62–S63, 3 2003.

L. Klebanov and A. Yakovlev. Treating expression levels of different genes as a sample in microarray data analysis: Is it worth a risk? *Statistical Applications in Genetics and Molecular Biology*, 5(1):Article 9, 2006.

I. Lönnstedt and T. Britton. Hierachical bayes models for cdna microarray gene expression. *Biostatistics*, 6(2):279–291, 2005.

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

J.S. Maritz. *Empirical Bayes Methods.* Methuen & Co. Ltd., 1970.

Y. Pawitan, S. Calza, and A. Ploner. Estimation of the false discovery proportion under general dependence. *Bioinformatics*, 22(24):3025–3031, 2006.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2006. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

M. Ritchie, D. Diyagama, J. Neilson, R. van Laar, A. Dobrovic, A. Holloway, and G. Smyth. Empirical array quality weights in the analysis of microarray data. *BMC Bioinformatics*, 7(1):261, 2006.

H. Robbins. An empirical Bayes approach to statistics. In J. Neyman, editor, *Third Berkeley Symposium on Mathematics and Probability*, pages 157–163, 1956.

L. Shi, W. Tong, F. Goodsaid, F.W. Frueh, H. Fang, T. Han, J.C. Fuscoe, and D.A. Casciano. QA/QC: Challenges and pitfalls facing the microarray community and regulatory agencies. *Expert Review of Molecular Diagnostics*, 4(6):761–777, 2004.

G.K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

A. Spira, J. Beane, V. Pinto-Plata, A. Kadar, G. Liu, V. Shah, B. Celli, and J.S. Brody. Gene expression profiling of human lung tissue from smokers with severe emphysema. *American Journal of Respiratory Cell and Molecular Biology*, 31(6):601–610, 2004.

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

Paper I

Statistical Applications in Genetics and Molecular Biology

Volume 4, *Issue* 1 2005 *Article* 30

Weighted Analysis of Paired Microarray Experiments

Erik Kristiansson* Anders Sjögren†

Mats Rudemo‡ Olle Nerman**

*Chalmers University of Technology, first two authors contributed equally, erikkr@math.chalmers.se

†Chalmers University of Technology, first two authors contributed equally, anders.sjogren@math.chalmers.se

‡Chalmers University of Technology, rudemo@math.chalmers.se

**Chalmers University of Technology, nerman@math.chalmers.se

Weighted Analysis of Paired Microarray Experiments*

Erik Kristiansson, Anders Sjögren, Mats Rudemo, and Olle Nerman

Abstract

In microarray experiments quality often varies, for example between samples and between arrays. The need for quality control is therefore strong. A statistical model and a corresponding analysis method is suggested for experiments with pairing, including designs with individuals observed before and after treatment and many experiments with two-colour spotted arrays. The model is of mixed type with some parameters estimated by an empirical Bayes method. Differences in quality are modelled by individual variances and correlations between repetitions. The method is applied to three real and several simulated datasets. Two of the real datasets are of Affymetrix type with patients profiled before and after treatment, and the third dataset is of two-colour spotted cDNA type. In all cases, the patients or arrays had different estimated variances, leading to distinctly unequal weights in the analysis. We suggest also plots which illustrate the variances and correlations that affect the weights computed by our analysis method. For simulated data the improvement relative to previously published methods without weighting is shown to be substantial.

KEYWORDS: Quality control, QC, Quality Assurance, QA, Quality Assessment, Empirical Bayes, DNA Microarray

---

*We would like to thank Mikael Benson, Lars Olaf Cardell, Lena Carlsson and Margareta Jernås for valuable discussions and access to the data from (Benson et al., 2004). We are also indebted to two referees and the editor for a series of valuable comments. Erik Kristiansson and Anders Sjögren wish to thank the National Research School in Genomics and Bioinformatics for support. Olle Nerman wishes to thank University of Canterbury and John Angus Erskine Bequest, New Zeeland for support by an Erskine Visiting Fellowship in spring 2005.

1 Introduction

DNA microarrays are strikingly efficient tools for analysing gene expression for large sets of genes simultaneously. They are often used to identify genes that are differentially expressed between two conditions, e.g. before and after some treatment. A drawback is that the technology involves several consecutive steps, each exhibiting large quality variation. Thus there is a strong need for quality assessment and quality control to handle occurrences of poor quality, as is clearly pointed out in Johnson and Lin (2003) and Shi et al. (2004).

Despite the observed need for effective quality control, standard operating procedures for quality assurance of the entire chain of processing steps have only recently been proposed (Ryan et al., 2004, for one-channel experiments). However, even utilising an optimal quality control procedure aiming at removing low quality arrays and/or individual gene measurements (e.g. spots), there will always be a marginal region with some measurements being of decreased quality without being worthless, as noted in Ryan et al. (2004). Consequently, it should be possible to make progress by integrating quality control quantitatively into the analysis following the lab steps and low-level analysis, taking quality variations into account.

When integrating the quality concept into the analysis, the quality of different parts of the dataset should ideally be estimated from data and used in the subsequent selection of differentially expressed genes. Here we introduce a method, called *Weighted Analysis of paired Microarray Experiments* (referred to as WAME), for the analysis of paired microarray experiments, e.g. comparison of pairs of treatment conditions and many two-colour experiments. WAME aims at estimating array- or repetition-wide quality deviations and integrates the quality estimates in the statistical analysis. Only the observed gene expression ratios are used in the quality assessment, making the method applicable to most paired microarray experiments, independent of which DNA microarray technology is used.

In short WAME identifies and downweights repetitions (biological or technical) of pairs (corresponding to individuals or to arrays) with decreased quality for many genes. Repetitions with positively correlated variations, e.g. caused by shared sources of variation, are similarly down-weighted. Thus, estimates of differential expression with improved precision and tests with increased power are provided.

As a useful complement to the WAME analyses we suggest pair-wise plots of log-ratios of gene expression measurements. Such plots are supplied for all three real datasets analysed, and are particularly useful for understanding

which patients or arrays are up- or down-weighted.

In the adopted model, log ratios of measured RNA-levels are assumed normally distributed. The covariance structure is specified by parameters of two types: (i) a global covariance matrix signifying different quality for different repetitions and (ii) gene specific multiplicative factors. The latter have inverse gamma prior distribution with one gene-specific parameter, which is estimated by an empirical Bayes method.

The paper is organised as follows. In Section 2, some background on microarray quality and a selected literature review are presented. This is followed by a detailed description of our model. Methods for estimating the parameters and a likelihood ratio test for identifying differentially expressed genes are derived. We give a summary of the computational procedure including a reference to R code available from the Internet. In the following section simulations are used to compare WAME to four currently used methods: (i) average fold change, (ii) ordinary t-test, (iii) the penalized t-statistic of Efron et al. (2001), and (iv) the moderated t-statistic of Smyth (2004). Next, WAME is applied to three real datasets, the *Cardiac* dataset of Hall et al. (2004), the *Polyp* dataset of Benson et al. (2004) and the *Swirl* dataset (Dudoit and Yang, 2003). The results obtained are discussed in a subsequent section and some derivations and mathematical details are given in an appendix.

2 Background

To put the quality control aspect of our model into context, the different steps and sources of variation in typical paired microarray experiments are outlined below. In addition, a selection of publications dealing with quality control for microarray experiments are briefly reviewed.

2.1 Sources of variation in typical microarray experiments

The first step, after decision on experimental design, of a microarray experiment aiming at identifying differentially expressed genes would typically be to determine how biological samples should be acquired. In experiments dealing with homogeneous groups of single cell organisms, such as yeast, in highly controlled environments, this task is typically less complex than when dealing with heterogeneous groups of multicellular organisms, such as humans. Here selection of subjects and cells from the relevant organ, e.g. by

biopsy or laser dissection, are complicated tasks.

From the biological sample the following lab-steps are performed: RNA extraction, reverse transcription (and *in vitro* transcription), labelling, hybridisation to arrays and scanning. The parts of the scanned images corresponding to the different genes (i.e. spots or probe-pairs) are identified and quantified. In addition, background correction may be performed. Subsequently, normalisation of the quantified measurements is performed to handle global differences. In the case of Affymetrix type arrays, 11-20 pairs of quantitative measurements are combined into one expression level estimate for each gene. For an experiment of paired type, one $\log_2$-ratio of the expression level estimates is calculated for each pair and gene. These $\log_2$-ratios are then used to examine which genes are differentially expressed.

In several of the steps mentioned above there are substantial quality variations. For example, the quantity and quality of the RNA in biopsies may vary considerably. There are sometimes evidence of poor quality making it possible to remove obviously worthless samples. Nevertheless, there will always be a marginal region with measurements of reduced quality without being worthless. In addition, some variations are hard to detect before the actual normalised $\log_2$-ratios are computed, e.g. non-representative tissue distribution in human biopsies. An additional aspect of quality control is systematic errors, where the variations of different repetitions are correlated. This could be due to shared sources of variation, such as simultaneous processing in lab steps or non-representative tissue composition in the biopsies.

Another potentially important factor is the quality of the arrays used for the measurements. Flaws in the manufacturing process might make measurements for individual genes inferior. This is more of a problem in the case of spotted arrays, for which there are only one or a few spots per gene. However, such bad spots can often be detected. The quality control in the actual manufacturing of microarrays is certainly very important but will not be further discussed here.

2.2 A brief review of some relevant publications

In Johnson and Lin (2003) and Shi et al. (2004) the general need for improved quality assurance in the context of DNA microarray analysis is emphasised. Tong et al. (2004) implement a public microarray data and analysis software and note that "Although the importance of quality control (QC) is generally understood, there is little QC practise in the existing microarray databases". They include some available measures of quality for different steps in the analysis in their database.

Dumur et al. (2004) survey quality control criteria for the wet lab steps of Affymetrix arrays, going from RNA to cDNA. Additionally, three sources of technical variation (hybridisation day, fluidic scan station, fresh or frozen cDNA) are evaluated using an ANOVA model.

Ryan et al. (2004) present guidelines for quality assurance of Affymetrix based microarray studies, utilising a variety of techniques for the different steps, some of which are shown to agree. A sample quality control flow diagram is suggested, including steps from extracted RNA to the quantified arrays.

Chen (2004) aims screens out ineligible arrays (Affymetrix type) using a graphical approach to display grouped data. Park et al. (2005) similarly aim at identifying outlying slides in two-channel experiments by using scatterplots of transformed versions of the signals from the two channels.

Tomita et al. (2004) use correlation between arrays (Affymetrix type) to evaluate the RNA integrity of the individual arrays, by forming an average correlation index (ACI). The ACI is shown to correlate with several existing quality factors, such as the 3'-5' ratio of GAPDH.

Li and Wong (2001) and Irizarry et al. (2003) introduce estimates of expression levels for probe-sets in Affymetrix type arrays, based on linear models of probe-level data. Bolstad (2004) extend the use of such probe-level models (PLM), e.g. by plotting residuals from the robust regression. It is thereby possible to visually assess the quality of the actual scanned and hybridised arrays, potentially detecting errors in certain steps of microarray experiments based on Affymetrix type arrays.

Several papers have been written on the quality control of individual measurements of genes (spots or probes). Wang et al. (2001, 2003) define a spot-wise composite score from various quantitative measures of quality of individual spots in spotted microarrays. They further perform evaluations on several in-house datasets, showing that when bad spots are removed, the variance of all gene-specific ratios in one chip is decreased. In Hautaniemi et al. (2003) Bayesian networks are used to discriminate between good and bad spots with training data provided by letting experienced microarray users examine the arrays by hand.

In the papers discussed above the countermeasure against low-quality spots or arrays is to treat them as outliers and to remove them. Again, there will always be a marginal region with some measurements being of decreased quality without being worthless. An interesting approach using weighted analysis of the microarray gene expression data is due to Bakewell and Wit (2005). The starting point is a variance component model for the log expression mean for a spot i with variance $\sigma_b^2 + \sigma_i^2/m_i$, where σ_b^2 is

the variance between spots while σ_i^2 is the variance between pixels in spot i and m_i is the effective number of pixels. For each gene the spots are weighted inversely proportional to estimated variances, and different genes are essentially treated independent of each other. Only quality deviations of the actual hybridised spots are included in the model.

In Yang et al. (2002) the variance of different print tip groups or arrays in cDNA experiments are estimated by a robust method. The need for scale normalisation between slides is determined empirically, e.g. by displaying box plots for the log ratios of the slides.

The model we propose (WAME) assesses the quality of different arrays quantitatively by examining the computed $\log_2$-ratios. Thus, quality deviations in all steps leading to the gene expression estimates are included, as long as the quality deviations occur for a wide variety of measured genes. Furthermore, shared systematic errors are taken care of via estimated covariances between repetitions. The assessed qualities are incorporated into the analysis based on the statistical model presented in the next sections.

In microarray experiments there are often relatively few replicates, resulting in highly variable gene-specific variance estimates. To use the information in the large number of measured genes to handle this problem, an empirical Bayes approach (Robbins, 1956; Maritz, 1970) can be taken, determining a prior distribution from the data, thus moderating extreme estimates. This approach has been used in Baldi and Long (2001), Lönnstedt and Speed (2002) and Smyth (2004).

3 The model

The experimental layouts studied in the present paper are restricted to comparisons of paired observations from two conditions. For each gene $g = 1, \ldots, N_G$ and each pair of measurements $i = 1, \ldots, N_I$, let X_{gi} with expected value μ_g be the normalised $\log_2$-ratio of the observed gene expressions from the two conditions. Thus, μ_g measures the expected $\log_2$ ratio of the RNA concentrations of the two conditions.

In Section 2.1 it was noted that there may exist dependencies between repetitions, e.g. due to systematic errors. Furthermore, different arrays may have different precision in their measurements of the gene expressions. To describe this, we use a covariance structure matrix Σ which models precision as individual variances for the different repetitions and dependencies between repetitions as covariances.

Due to both technical and biological reasons the observations for the dif-

ferent genes have different variability, and a gene-specific multiplicative factor c_g for the covariance matrix is introduced. The c_g-variables for different genes are assumed to be independent. Given c_g the vector $\mathbf{X_g}$ consisting of all repetitions for gene g is assumed to have a N_I-dimensional normal distribution with mean vector $\mu_g\mathbf{1}$ and covariance matrix $c_g\Sigma$. The vectors $\mathbf{X_g}$ for different genes are also assumed independent. This independence assumption is optimistic but we believe that it is not critical in the Σ-estimation step owing to the large number of genes.

In microarray experiments, the number of experimental units is typically fairly small and estimates of c_g utilising only information from the measurements with gene g may be highly variable. Therefore prior information is introduced as a prior distribution for c_g, which serves to moderate the estimates of c_g. The prior for c_g is assumed to be an inverse gamma distribution with a parameter α determining the spread of the distribution, in effect determining the information content in the prior. The inverse gamma distribution is a conjugate prior distribution for the variance of a normal distribution and has as such been used in Bayesian and empirical Bayesian analysis of microarray data before (Baldi and Long, 2001; Lönnstedt and Speed, 2002; Smyth, 2004).

The model can be summarised as follows: We observe $\mathbf{X}_g = (X_{g1}, \ldots, X_{gN_I})$ where $g = 1, \ldots, N_G$. Let Σ be a covariance matrix with N_I rows and columns, c_g a set of gene-specific variance scaling factors and α a hyperparameter determining the spread of the prior distribution for c_g. Then for fixed μ_g, Σ and α,

$$\begin{aligned} c_g &\sim \Gamma^{-1}(\alpha, 1) \text{ and} \\ \mathbf{X}_g \mid c_g &\sim \mathbf{N}_{N_I}\left(\mu_g\mathbf{1}, c_g\Sigma\right), \end{aligned} \tag{1}$$

and all variables corresponding to different genes are assumed independent.

4 Inference

4.1 Estimation of a scaled version of the matrix Σ

Estimating Σ may appear easy but it turns out to be rather intricate and there are several issues involved.

Firstly, there are trivial solutions that give infinite likelihood of the model. For instance, if the gene-specific mean value μ_g is equal to the observation of one of the repetitions the likelihood goes to infinity as the corresponding variance goes to zero. To avoid this complication the assumption that the

differential expression of most genes is approximately zero is introduced temporarily. This assumption is not as consequential as it might sound, since it is made by most of the procedures that have become *de facto* standard in the (preceding) normalisation step, one example being the loess normalisation method (Yang et al., 2002). Nevertheless, it does limit the set of experimental setups that can be treated and the proportion of genes that are regulated must not be too large. The impact of this assumption is further investigated by the simulation study in Section 5.2. For the rest of Section 4.1, μ_g is thus set equal to zero for all $g = 1, \ldots, N_G$.

Another issue is the scaling of Σ. For each gene, the covariance matrix is scaled with the random variable c_g which has an inverse gamma distribution with a parameter which is unknown in a first stage. To address this issue, the estimation of Σ is performed in two steps. In the first step, a transformation is applied to $\mathbf{X}_g$ such that the transformed vector has a distribution that is independent of c_g. To simplify notation the index g will be dropped from $\mathbf{X}_g$ and c_g in the rest of this section. Let $\mathbf{U} = (U_1, \ldots, U_{N_I})$ where

$$U_i = \begin{cases} X_1 & \text{if } i = 1 \\ X_i/X_1 & \text{if } 2 \leq i \leq N_I \end{cases} . \tag{2}$$

The distribution of the vector $\mathbf{U}$ has the density

$$f_{\mathbf{U} \mid c,\Sigma}(\mathbf{u}) = f_{\mathbf{X} \mid c,\Sigma}(\mathbf{x}(\mathbf{u}))|J(\mathbf{u})|$$

where J is the corresponding Jacobian. Some algebra shows that the scaling factor c cancels for $U_2, \ldots, U_{N_I}$ and by integrating over U_1, we get the density

$$\begin{aligned} f_{U_2,\ldots,U_{N_I} \mid \Sigma}(u_2, \ldots, u_{N_I}) &= \int_{-\infty}^{\infty} f_{\mathbf{U} \mid c,\Sigma}(\mathbf{u})\, du_1 \\ &= C\, |\Sigma|^{-1/2} \left[v^{\mathrm{T}} \Sigma^{-1} v\right]^{-N_I/2}, \end{aligned} \tag{3}$$

where C is a normalisation constant and $v = (1, u_2, \ldots, u_{N_I})$. The distribution (3) is independent of c and the marginal distribution of u_i is a Cauchy distribution translated with $\rho_{1,i}\sigma_{i,i}/\sigma_{1,1}$ and scaled with $\sqrt{1-\rho_{1,i}^2}\sigma_{i,i}/\sigma_{1,1}$, where $\rho_{1,i}$ is the correlation between X_1 and X_i and $\sigma_{i,i}$ is the variance of X_i. This shows that $\rho_{1,i}$ and $\sigma_{i,i}/\sigma_{1,1}$ are identifiable. Analogously, from the one dimensional Cauchy distributions of $U_j/U_k = X_j/X_k$, , $j = 2, \ldots, N_I$, $k = 2, \ldots, N_I$ and $j \neq k$, it follows that all other correlations and variance ratios are identifiable as well.

From (3) we see that the distribution of $(U_2, \ldots, U_{N_I})$ is unchanged if we multiply Σ with a constant. Let us therefore fix one element of Σ, e.g. we

set the first element in the first row equal to one. Let Σ^* denote the matrix thus obtained. Then

$$\Sigma^* = \lambda\Sigma, \tag{4}$$

and the constant λ will be estimated together with the hyperparameter α as described below in Section 4.2. Thus estimation of the covariance matrix Σ will be carried out in two steps: first estimate Σ^* with one element fixed and then estimate λ.

Numerical maximum likelihood based on the distribution (3) is used to produce a point estimate of Σ^*. Here the number of unknown parameters are $N_I(N_I+1)/2$, growing as N_I^2. To get an efficient implementation C/C++ is combined with R (R Development Core Team, 2004). The resulting computational time for three arrays is less than a second and for 30 arrays it takes a few hours.

4.2 Estimation of the hyperparameter α and the scale λ

In this section, we develop methods for estimation of the hyperparameter α as well as the scale parameter λ in (4). From the model assumptions in Section 3 we recall that c_g has an inverse gamma distribution with hyperparameter α, e.g.

$$c_g \mid \alpha \sim \Gamma^{-1}(\alpha, 1).$$

The inference of α will be based on the statistic

$$S_g = (A\mathbf{X}_g)^{\mathrm{T}}(A\Sigma A^{\mathrm{T}})^{-1}A\mathbf{X}_g,$$

where A is an arbitrary $N_I - 1 \times N_I$ matrix with full rank and each row sum equal to 0. It follows that the distribution of S_g conditioned on c_g is a scaled chi-square distribution with $N_I - 1$ degrees of freedom,

$$S_g \mid c_g \sim c_g \cdot \chi^2_{N_I-1}\,.$$

The unconditional distribution of S_g can be calculated by use of the fact that a gamma distribution divided by another gamma distribution has an analytically known distribution, a beta prime distribution (Johnson et al., 1995, page 248). Thus,

$$S_g \mid \alpha \sim 2 \times \beta'\left((N_I-1)/2, \alpha\right), \tag{5}$$

which has the density function

$$f_{S_g \mid \alpha}(s_g) = \frac{1}{2}\frac{\Gamma(\alpha + (N_I-1)/2)}{\Gamma(\alpha)\Gamma((N_I-1)/2)}\frac{(s_g/2)^{(N_I-1)/2-1}}{[1+s_g/2]^{\alpha+(N_I-1)/2}}\,.$$

In the same fashion, denote the variance estimator based on Σ^* in (4) by S_g^*, that is,

$$S_g^* = (AX_g)^{\mathrm{T}}(A\Sigma^* A^{\mathrm{T}})^{-1}AX_g \,. \tag{6}$$

It follows that, $S_g^* = S_g/\lambda$ so

$$S_g^* \mid \alpha, \lambda \sim 2/\lambda \times \beta'\left((N_I - 1)/2, \alpha\right). \tag{7}$$

Assuming independence between the genes, α and λ can now be estimated by numerical maximum likelihood. The estimated value of the (unscaled) covariance matrix Σ can then be calculated from Equation (4). Results from simulations show that the estimation of α and λ is accurate enough for realistic values (results not shown). In the following sections, these parameters are therefore assumed to be known.

4.3 The posterior distribution of c_g

The posterior distribution of c_g is not explicitly used in the calculations above, but still of general interest. As previously mentioned, the distribution of S_g conditioned on c_g is a scaled chi-square distribution with $N_I - 1$ degrees of freedom. Since chi-square distributions and inverse gamma distributions are conjugates, the posterior distribution of c_g given S_g is an inverse gamma distribution as well. We find

$$c_g \sim \Gamma^{-1}(\alpha, 1)$$

$$c_g \mid S_g \sim \Gamma^{-1}\left(\alpha + (N_I - 1)/2, 1 + \frac{S_g}{2}\right),$$

and the prior can be interpreted as representing 2α pseudo observations, which add a common variance estimate to all genes. A discussion regarding the use of this model in microarray analysis can be found in Lönnstedt and Speed (2002) and Smyth (2004) and a general discussion in Robert (2003) Section 4.4.

4.4 Inference about μ_g

In this section we derive a statistical test for differential expression based on the WAME model. The hypotheses for gene g can be formulated as

$$H_0: \text{ gene } g \text{ is not regulated } (\mu_g = 0)$$
$$H_A: \text{ gene } g \text{ is regulated } (\mu_g \neq 0).$$

A test suitable for the hypothesis H_0 is the likelihood ratio test (LRT) based on the ratio of the maximum values of the likelihood function under the different hypotheses. With our notation we reject H if

$$\frac{\sup\limits_{H_A} L\left(\mu_g|\mathbf{x}_g\right)}{\sup\limits_{H_0} L\left(\mu_g|\mathbf{x}_g\right)} = \frac{\sup\limits_{\mu_g \neq 0} L\left(\mu_g|\mathbf{x}_g\right)}{L\left(0|\mathbf{x}_g\right)} \geq k, \tag{8}$$

where k, $1 \leq k < \infty$, sets the level of the test. To calculate the likelihood function, we need to integrate over c_g, e.g.,

$$\begin{aligned} L\left(\mu_g|\mathbf{x}\right) &= \int f_{\mathbf{x}\,|\,\mu_g,c_g,\Sigma}(\mathbf{x}) f_{c_g\,|\,\alpha}(c_g)\, dc_g \\ &= (2\pi)^{-N_I/2} \left|\Sigma\right|^{-1/2} \frac{\Gamma(N_I/2+\alpha)}{\Gamma(\alpha)} \left[\frac{\left(\mathbf{x}_g - \mu_g\mathbf{1}\right)^{\mathrm{T}} \Sigma^{-1} \left(\mathbf{x}_g - \mu_g\mathbf{1}\right)}{2} + 1\right]^{-(\alpha+N_I/2)}. \end{aligned}$$

It is now straight forward to calculate the denominator $L(0|\mathbf{x}_g)$ in (8) and some algebra shows that the numerator is maximised by $\hat{\mu}_g = \bar{x}_g^w$, where

$$\bar{x}_g^w = \frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\mathbf{x}_g\ , \tag{9}$$

is a *weighted mean value* of the observations. Analogously, we define the random variable $\bar{X}_g^w$ by replacing $\mathbf{x}_g$ with $\mathbf{X}_g$. Then,

$$\bar{X}_g^w | c_g \sim \mathbf{N}\left(\mu_g, \frac{c_g}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right)$$

and it can be shown that

$$\mathbf{w}^{\mathrm{T}} = \frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}} \tag{10}$$

is the weight vector that minimises the variance of $\mathbf{w}^{\mathrm{T}}\mathbf{X}_g$. The weights in equation (10) will depend on the covariance matrix as follows. A repetition with high variance will have a low weight while a repetition with low variance will have a high weight. Moreover, a positive high correlation between repetitions will cause decreased weights. Note that if a repetition is highly correlated with a repetition with lower variance, its weight can actually become negative. According to the theory, this is nothing strange but practically this is of course not satisfying. Fortunately, such extreme cases seem to be rare in the microarray context and if they appear, the source of the correlation should be investigated and one could consider removing the negatively weighted repetition.

Evaluation of the likelihood function at 0 and $\bar{x}_g^w$ and a few calculations show that the inequality (8) is equivalent to

$$\frac{|\bar{x}_g^w|}{\sqrt{s_g + 2}} \geq k'$$

where s_g is the observed value of S_g defined in Section 4.2 and k' is some non-negative constant. Define

$$T_g = \sqrt{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}\,(N_I - 1 + 2\alpha)}\;\frac{\bar{X}_g^w}{\sqrt{S_g + 2}} \tag{11}$$

and reject the null hypothesis if

$$|T_g| \geq k'',$$

where k'' is another non-negative constant. The statistic T_g will be referred to as the *weighted moderated t-statistic* since it is a weighted generalisation of the moderated t-statistic derived by Lönnstedt and Speed (2002) and refined by Smyth (2004). Indeed, if all repetitions have equal estimated variances and no estimated correlations, T_g becomes equivalent to the result in Section 3 in Smyth (2004). To calculate the value of k'' that corresponds to a given level of the test, the distribution of T_g needs to be derived. Under the null hypothesis, it turns out to be a t-distribution with $2\alpha + N_I - 1$ degrees of freedom,

$$T_g \;\sim\; t_{2\alpha+N_I-1}\;. \tag{12}$$

4.5 Summary of the computational procedure

The computational procedure is summarized below in eight steps including three types of model control. R code corresponding to these steps is available from `http://wame.math.chalmers.se`.

(i) To estimate Σ^*, optimize numerically the product of the right members of (3) for all genes as a function of Σ with the element in the upper left corner set equal to 1. For each gene $v = (1, u_2, \ldots, u_{N_I})$ with $u_2, \ldots, u_{N_I}$ given by (2).

(ii) Compute $S_g^*, g = 1, \ldots, N_G$, in (6) for some full rank $N_I - 1 \times N_I$ matrix A with zero row sums.

(iii) Estimate α and λ by numerical maximum likelihood with the distribution (7) for $S_g^*, g = 1, \ldots, N_G$, assumed to be independent.

(iv) Compute $\Sigma = (1/\lambda)\Sigma^*$.

(v) For each gene g compute $\bar{X}_g^w$ from (9) with $\mathbf{x}_g$ replaced by $\mathbf{X}_g$ and compute T_g from (11) with $S_g = \lambda S_g^*$. From the T_g-values p-values may be computed from the distribution (12) and a gene ranking list may be produced.

(vi) Compute the empirical distribution of T_g, $g = 1, \ldots, N_G$, and plot it together with the density of the theoretical distribution (12) as a model control. The corresponding q-q plot is expected to coincide with the theoretical distribution in the central part but typically not in the tails.

(vii) Compute the empirical distribution of S_g, $g = 1, \ldots, N_G$, and plot it together with the density of the theoretical distribution (5) as a model control.

(viii) As an additional model control, plot pairwise $\log_2$-ratios for repetitions as in Figures 3, 6 and 8 below.

5 Results from simulations

5.1 Comparison to similar gene ranking methods

A simulation study was done to compare the performance of WAME to four published methods. These methods were

- Average fold-change
- Ordinary t-statistic
- Efron's penalized t-statistic
- Smyth's moderated t-statistic

The average fold-change for a gene is simply the mean value over all the observed $\log_2$-ratios and the ordinary t-statistic is the average fold-change divided by the corresponding sample standard deviation. These two methods have traditionally been popular gene ranking methods and it is therefore interesting to see how they perform. Another method introduced in Efron et al. (2001) is the penalized t-statistic which is a modified version of the ordinary t-statistic where a constant has been added to the sample standard deviation. The motivation for this adjustment is the unreliability of the t-statistic in situations when only a few repetitions are used. The constant used here was chosen as the 90th percentile of the empirical distribution of the sample standard deviations, according to Efron et al. (2001). Finally, the moderated t-statistic is included. It was developed and implemented by Smyth (2004) and it is available in the R package LIMMA (Smyth et al., 2003). The moderated t-statistic can be seen as a refined version of the B-statistic which was first presented in Lönnstedt and Speed (2002). In the paired microarray context, WAME is a generalisation of LIMMA in the sense that the two models are identical when all repetitions have the same variance and no correlations exist.

All methods were applied to a series of simulated datasets with different settings and the number of true positives as a function of false positives was plotted, generating several so called receiver operating characteristic (ROC) curves. The average over 100 datasets was used to produce a single curve where each dataset was created as follows. The number of genes (N_G) was fixed to 10000, the number of repetitions (N_I) to 4 and the hyperparameter α to 2. These values were chosen since they are typical for real datasets. The covariance matrix Σ is fixed and for each gene g the following steps were done.

1. c_g was sampled from an inverse gamma distribution according to the model specification.

2. A vector of $N_I = 4$ independent observations was drawn from a normal distribution with mean value zero and variance one. This vector was then multiplied by the square-root matrix of Σ.

3. If this particular gene was selected to be regulated, then the absolute mean value for each of the N_I elements was drawn from a uniform distribution between 0 and 2.

5% of the genes were randomly selected and set to be upregulated. Analogously, 5% were downregulated resulting in totally 10% regulated genes. It should be noted that it is only the total number of regulated genes that had an impact on the performance for the different methods, not the number of upregulated genes compared to the number of downregulated genes.

Four cases, all with different covariance matrices, were studied. In the first case, all of the repetitions had variances equal to 1 and there were no correlations, thus Σ was an identity matrix. The ROC curves produced by the simulated data can be seen in the upper part of Figure 1. WAME and LIMMA performs best, closely followed by the penalized t-statistic. Note that WAME and LIMMA have almost identical performance in this case and, as mention above, this was expected since the weighted moderated t-statistic and the moderated t-statistic are almost equivalent for this setting. Another interesting detail is the weak performance of the t-statistic due to its instability issues when only few repetitions are used.

In the second case, different variances were introduced. Σ was again a diagonal matrix but with the values 0.5, 1, 1.5 and 2 on the diagonal, thus all correlations were again zero. The ROC curves can be seen in the lower part of Figure 1. As before, WAME and LIMMA are the methods that performs best, but in this case, WAME performs better since it put less weight on the repetitions with high variance.

To investigate the impact of correlations, the third case used

$$\Sigma = \begin{pmatrix} 1.0 & 0.4 & 0.2 & 0.0 \\ 0.4 & 1.0 & 0.4 & 0.2 \\ 0.2 & 0.4 & 1.0 & 0.4 \\ 0.0 & 0.2 & 0.4 & 1.0 \end{pmatrix}. \tag{13}$$

This corresponds to a case when there are both high and low correlations between the repetitions. The upper part of Figure 2 shows that WAME

performs slightly better than both LIMMA and the penalized t-statistic since it estimates the correlations and takes them into account.

Finally, in the fourth case both different variances and correlations were included. The variances and correlations were identical to the ones in the second and third cases respectively, i.e. variances of 0.5, 1.0, 1.5, 2.0 and correlations of 0, 0.2 and 0.4, the latter placed according to (13). The result can be seen in the lower part of Figure 2. Here, the largest advantage of using WAME can be seen. For a rejection threshold such that half of the selected genes are true positives, using WAME results in almost a third less false positives which can correspond to hundreds of genes.

All four simulations show that WAME and its weighted moderated t-statistic perform at least as good as the moderated and penalized t-statistics. In the case of both different variances and correlations between the repetitions, WAME performs clearly better than all of the included methods. Both the average fold-change and the ordinary t-statistic have poor performance in the current setting with only four repetitions.

5.2 Evaluation of the point estimator of Σ

The estimation of Σ is one of the crucial steps when applying WAME since errors made will affect estimates of other entities such as α and the weighted mean value $\bar{x}_g^w$. The resulting precision and accuracy when numerical maximum likelihood is applied to the distribution in equation (3) are therefore interesting questions, both when the model assumptions hold and when they are violated. In an attempt to partially answer these questions, Σ was estimated from different simulated datasets and the results were compared to the true values. The datasets were created according to the description in the previous section and the same parameters were used, i.e. $N_G = 10000$, $N_I = 4$ and $\alpha = 2$. Five different cases, listed in Table 1, were examined. As in the previous section, 100 datasets were simulated for each setting and for each such dataset the covariance matrix Σ and the hyperparameter α were estimated according to Section 4. Table 2 summarises the result where the true value of Σ, the mean value of the estimated Σ as well as the standard deviations are listed. It should be noted that in all cases, except for case **III**, α is estimated with high accuracy and precision.

In the first two cases (**I** and **II**), the covariance matrix was estimated without any bias and with low standard deviation showing that the methods are accurate under the model assumptions. In case **III** the normal distribution was substituted against a t-distribution with 5 degrees of freedom, having substantially heavier tails. The estimated Σ seems to be slightly biased to-

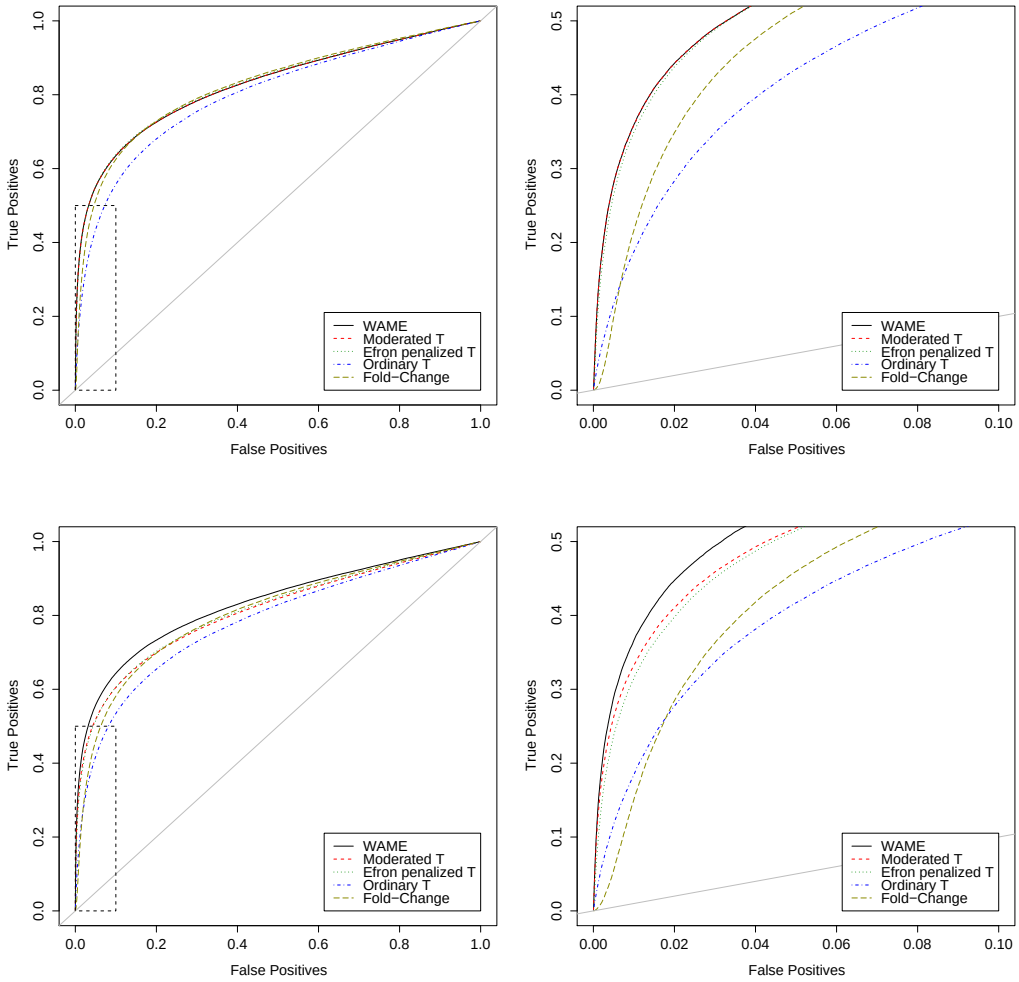


Figure 1: ROC curves from simulated data. The pair at the top, from the first case, show the performance of the evaluated methods on data with equal variances of 1 for all replicates and no correlations. The pair at the bottom, from the second case, analogously show the performance on data with different variances of 0.5, 1, 1.5, 2 and no correlations. The parameters used for these two simulations were as follows. $N_G = 10000$, $N_I = 4$, $\alpha = 2$ and 10% of the genes were regulated. The figures to the right are magnifications of the dashed boxes to the left.

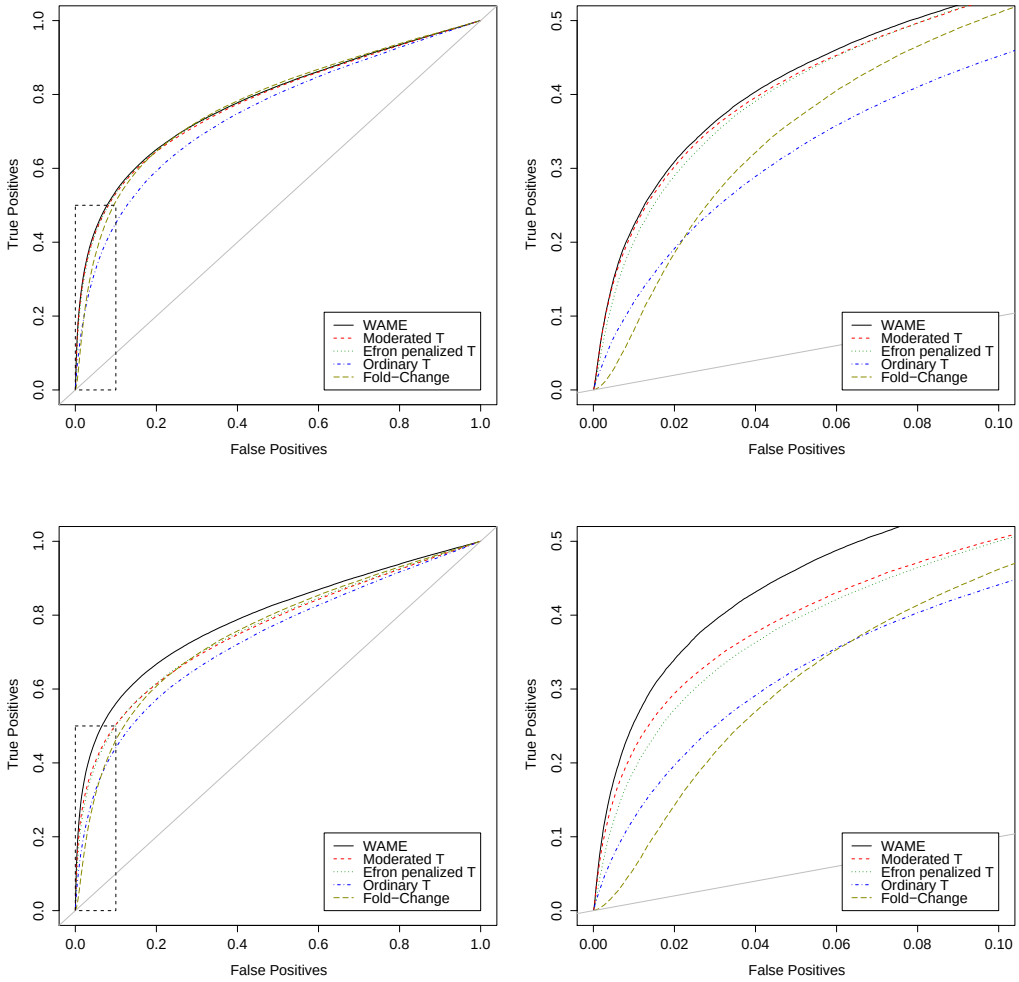


Figure 2: ROC curves from simulated data. The pair at the top, from the third case, show the performance of the evaluated methods on data with equal variances of 1 for all replicates and correlations of 0, 0.2 and 0.4, placed according to (13). The pair at the bottom, from the fourth case, analogously show the performance on data with different variances of 0.5, 1, 1.5, 2 and correlations of 0, 0.2 and 0.4, placed according to (13). The parameters used for these two simulations were as follows. $N_G = 10000$, $N_I = 4$, $\alpha = 2$ and 10% of the genes were regulated. The figures to the right are magnifications of the dashed boxes to the left.

Case	Correlation	Heavy tails	Regulated genes	Filter
I	No	No	None	No
II	Yes	No	None	No
III	Yes	Yes	None	No
IV	Yes	No	Yes, 10%	No
V	Yes	No	Yes, 10%	Yes, 5% removed.

Table 1: Descriptions of the five different settings used in this simulation study. When correlations are used, they follow the structure in equation (13).

ward higher variances and α was estimated to be 1.55 instead of 2. This pattern was also seen when the degrees of freedom were increased to 10 and 15 (results not shown). In case **IV** 10% of the genes were set to be regulated and since no differentially expressed genes are assumed, the regulation leads to positive correlations and increased variance estimates. Having 10% of the genes regulated is a rather high number, but not extreme. Therefore, a filter was applied to minimise the impact of regulated genes on the estimation of the covariance matrix. For each gene g, the filter calculates the minimal absolute value of the fold change, which will be denoted $X_{g,min}$. Removing the top 5% of the genes with highest $X_{g,min}$ gave a much better estimate of Σ, which is included as case **V**. Note that the genes were only removed from the the estimate of Σ^*, i.e. the arbitrarily scaled Σ, and not from the estimates of α and λ. Also note that the number 5% depends on several parameters, such as the total number of regulated genes and the covariance matrix itself. The results of the filtering procedure on real data is presented in the next section.

6 Results from real data

WAME was run on three real data sets: the ischemic part of the dataset of Hall et al. (2004), the dataset of Benson et al. (2004) (henceforth referred to as the *Cardiac* and *Polyp* datasets, respectively) and the *Swirl* dataset (described in Section 3.3 of Dudoit and Yang, 2003). These datasets represent microarray experiments with different characteristics; different laboratories, both two-colour cDNA and one-channel oligonucleotide (Affymetrix) arrays, different tissues and two different species (human and zebrafish). The Cardiac and Swirl datasets are publicly available.

	True Σ				Mean estimated Σ				Sample standard deviation			
I	0.50	0.00	0.00	0.00	0.50	0.00	-0.00	-0.00	0.01	0.01	0.01	0.01
	0.00	1.00	0.00	0.00	*0.00*	1.01	-0.00	0.00	*0.01*	0.04	0.02	0.01
	0.00	*0.00*	1.50	0.00	*-0.00*	*-0.00*	1.51	-0.00	*0.02*	*0.02*	0.05	0.02
	0.00	*0.00*	*0.00*	2.00	*-0.00*	*0.00*	*-0.00*	2.02	*0.01*	*0.01*	*0.01*	0.07
II	0.50	0.28	0.17	0.00	0.50	0.28	0.17	0.00	0.02	0.01	0.01	0.01
	0.40	1.00	0.49	0.28	*0.40*	1.00	0.50	0.29	*0.01*	0.04	0.02	0.03
	0.20	*0.40*	1.50	0.69	*0.20*	*0.40*	1.51	0.70	*0.01*	*0.01*	0.06	0.04
	0.00	*0.20*	*0.40*	2.00	*0.00*	*0.20*	*0.40*	2.00	*0.01*	*0.01*	*0.01*	0.11
III	0.50	0.28	0.17	0.00	0.51	0.29	0.18	-0.00	0.02	0.01	0.01	0.01
	0.40	1.00	0.49	0.28	*0.40*	1.01	0.50	0.28	*0.01*	0.04	0.02	0.02
	0.20	*0.40*	1.50	0.69	*0.20*	*0.40*	1.52	0.70	*0.01*	*0.01*	0.05	0.03
	0.00	*0.20*	*0.40*	2.00	*-0.00*	*0.20*	*0.40*	2.03	*0.01*	*0.01*	*0.01*	0.07
IV	0.50	0.28	0.17	0.00	0.61	0.39	0.28	0.11	0.02	0.02	0.02	0.01
	0.40	1.00	0.49	0.28	*0.48*	1.11	0.60	0.39	*0.01*	0.04	0.03	0.01
	0.20	*0.40*	1.50	0.69	*0.28*	*0.45*	1.61	0.80	*0.01*	*0.01*	0.06	0.04
	0.00	*0.20*	*0.40*	2.00	*0.10*	*0.25*	*0.43*	2.11	*0.01*	*0.01*	*0.01*	0.08
V	0.50	0.28	0.17	0.00	0.46	0.21	0.11	-0.02	0.01	0.01	0.01	0.02
	0.40	1.00	0.49	0.28	*0.33*	0.90	0.38	0.22	*0.01*	0.02	0.02	0.02
	0.20	*0.40*	1.50	0.69	*0.14*	*0.34*	1.39	0.59	*0.01*	*0.02*	0.06	0.03
	0.00	*0.20*	*0.40*	2.00	*-0.02*	*0.16*	*0.36*	1.93	*0.02*	*0.01*	*0.01*	0.07

Table 2: Result from the estimations of Σ from each of the five different cases. Correlations are shown in italic and covariances in non-italic. The parameter values used were $N_G = 10000$, $N_I = 4$ and $\alpha = 2$. The mean values and sample standard deviations were calculated from the results of 100 simulated datasets. Refer to Table 1 for a description of the different cases.

The Cardiac dataset is described to have been strictly quality controlled by a combination of several available methods. The dataset is therefore interesting to examine to see if WAME detects relevant differences in quality even in an example of a quality controlled, publicly available dataset. The Polyp dataset includes one biopsy that was previously thought to be an outlier and therefore discarded, thus providing a case with one seemingly lesser quality to be detected. In the Swirl dataset, two highly differentially expressed genes exist. Therefore, it is of interest to check that those genes are highly ranked by WAME. Furthermore, the Swirl dataset has been analysed previously in Smyth (2004).

6.1 Cardiac dataset

In the public dataset from Hall et al. (2004), heart biopsies from 19 patients with heart failure were harvested before and after mechanical support with a ventricular assist device. The aim of the study was to "define critical regulatory genes governing myocardial remodelling in response to significant reductions in wall stress", where a first step was to identify differentially expressed genes between the two conditions.

Affymetrix one-channel oligonucleotide arrays of type HG-U133A were used in the study, each containing 22283 probe-sets. The quality of the arrays was controlled using quality measures recommended by Affymetrix as well as by the program Gene Expressionist (GeneData, Basel, Switzerland). The quality of the different lab steps leading to the actual hybridisations were controlled using standard methods. The 19 patients were divided into three groups: ischemic (5 patients), acute myocardial infarction (6 patients) and non-ischemic (8 patients). The ischemic group was the smallest and consequently the one where quality variations might make the biggest difference. It was therefore chosen for further examination using WAME, to see if relevant quality variations could be detected despite the close quality monitoring.

The dataset was retrieved in raw .CEL-format from the public repository Gene Expression Omnibus (Edgar et al., 2002). The .CEL-files were subsequently processed using RMA (Irizarry et al., 2003) on all the arrays of the 19 patients simultaneously. Patient-wise $\log_2$-ratios of the five ischemic patients were then formed by taking pairwise differences of the $\log_2$ measurements before and after implant.

Applying WAME to the patient-wise $\log_2$-ratios provided interesting results. The estimated covariance matrix (see Table 3) suggests that two of the five patients (I13 and I7) were substantially more variable than the others,

while the correlations between patients were rather limited. These numbers seem credible when examining Figure 3, where for each pair of patients, the respective $\log_2$-ratios of all genes were plotted against each other. The plots clearly show that the observations of the two patients in question (I13 and I7) are more variable than the others.

The corresponding weights, derived from the estimated covariance matrix Σ, are shown in Table 4. As was discussed in Sections 4.1 and 5.2, when estimating Σ all genes are assumed to be non-differentially expressed. To examine the impact of potentially regulated genes on the estimation of Σ, the analysis was redone, removing genes with high lowest absolute $\log_2$-ratio in the estimation of Σ, as described in Section 5.2. The individual elements of the estimated covariance matrix and of α changed only slightly, even when as much as 50% of the data was removed (data not shown). This is reflected in the weights in Table 4.

	Patient				
Patient	I12	I13	I4	I7	I8
I12	0.046	0.003	0.001	0.012	0.002
I13	*0.033*	0.196	-0.014	0.007	-0.001
I4	*0.023*	*-0.126*	0.065	0.013	0.002
I7	*0.111*	*0.030*	*0.102*	0.258	-0.017
I8	*0.040*	*-0.011*	*0.038*	*-0.152*	0.047

Table 3: Estimated covariance-correlation matrix, Σ, for patients in the Cardiac dataset. (Correlations in italic, covariances in non-italic.)

	Patient				
Removed genes	I12	I13	I4	I7	I8
none	0.297	0.091	0.232	0.053	0.326
5%	0.301	0.089	0.233	0.054	0.323
10%	0.303	0.087	0.235	0.053	0.321
50%	0.323	0.082	0.240	0.046	0.308

Table 4: Weights for patients in the Cardiac dataset. Different numbers of potentially regulated genes were removed in the estimation of Σ, to check their influence. Potential regulation was measured by minimal absolute $\log_2$-ratio among the patients.

The hyperparameter α related to the spread of the gene-specific variance

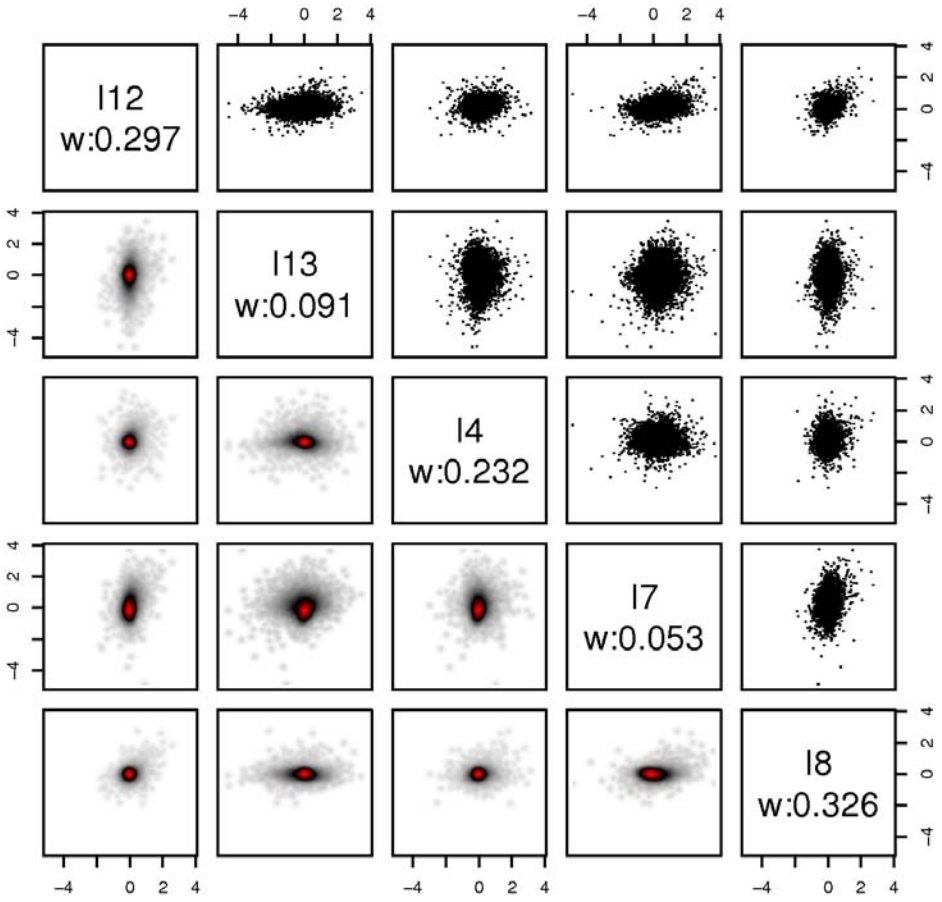


Figure 3: Pair-wise plots of the $\log_2$-ratios of the patients in the Cardiac dataset. The plots to the lower-left show two-dimensional kernel density estimates of the distribution of $\log_2$-ratios in each pair of patients. This provides information in the central areas where the corresponding scatterplots are solid black (cf. Figure 6 in Huber et al., 2003). The colour-scale is, in increasing level of density: white, grey, black and red.

scaling factors, c_g, was estimated to be 1.92, giving a heavy tail for the prior distribution. Thus removing c_g by transformation when estimating Σ (Section 4.1) is justified.

Inspecting the fitted distribution of S_g given $\alpha = 1.92$ against the empirical distribution of S_g reveals a good fit (see Figure 4), implying that the family of inverse gamma prior distributions is rich enough for this dataset.

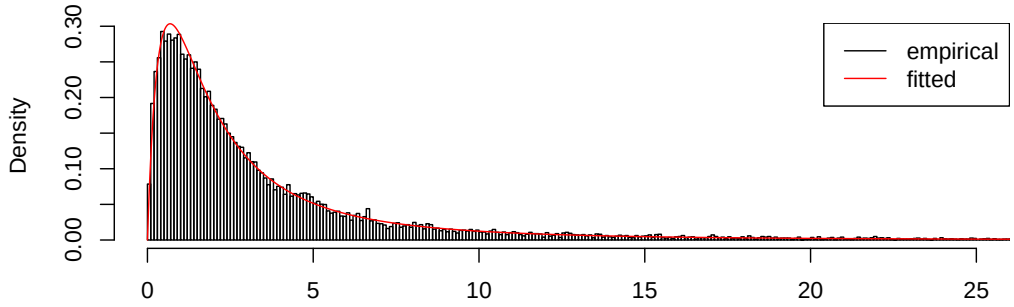


Figure 4: Empirical distribution of S_g in the Cardiac dataset, together with the density of S_g given $\alpha = 1.92$.

Examining the observed values of the statistic, T_g, compared to the expected null distribution reveals a good overall concordance (see Figure 5). Some genes have a larger t_g than can be explained by the null distribution, which points toward some of them being up-regulated by the treatment (see the qq-plot in Figure 5).

6.2 Polyp dataset

In the dataset from Benson et al. (2004), biopsies from nasal polyps of five patients were taken before and after treatment with local glucocorticoids. The goal was to examine closer the mechanisms behind the effect of the treatment and one step was to identify differentially expressed genes. Technical duplicates stemming from the same extracted RNA were run for each biopsy on Affymetrix HG-U133A arrays. This gave a dataset of 20 arrays and 22283 probe-sets.

Comparing each of the arrays in the dataset with all arrays from other patients and/or conditions, by looking at pair-wise scatterplots, the arrays from before treatment of patient 2 consistently showed larger variation than any other. The biopsy in question was found to be considerably smaller than the others, providing possible explanations such as non-representativeness

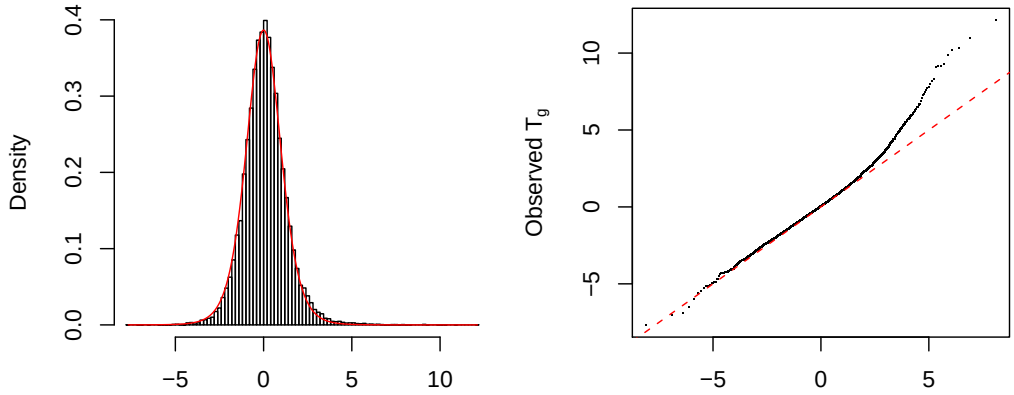


Figure 5: To the left, a histogram of the observed T_g-values together with the density of the null distribution (in red), in the Cardiac dataset. To the right, a quantile-quantile plot where the observed values of T_g are plotted against the quantiles of T_g under the null hypothesis. The central part of the empirical distribution follows the identity line well, showing good concordance with the null distribution. For high positive T_g-values, the observations clearly deviate from the predicted ones, pointing at the existence of up-regulated genes.

in tissue distribution. The data from patient 2 was therefore excluded in Benson et al. (2004).

WAME would preferably identify the patient 2 observation as having larger variation and downweight it. The data was processed using RMA (Irizarry et al., 2003) and the $\log_2$-ratio for each patient was formed by taking differences between the averages over the technical duplicates, before and after treatment, combining 4 arrays for each patient into one set of $\log_2$-ratios. Making one scatter plot of the two sets of $\log_2$-ratios for each pair of patients (Figure 6) clearly indicates that patient 2 is more variable than patients 1,3 and 5. Interestingly, the measurements from patients 1 and 2 seem to be highly correlated and patient 4 seems to have high variability.

Estimating the covariance matrix, Σ, the correlation between patients 1 and 2 is estimated to 0.82 (see Table 5), which is high but not unbelievable when studying Figure 6. The variance of patient 2 is furthermore estimated to be four times that of patient 1. Examining the resulting weights, patient 2 actually receives a weight of -2% (see Table 6). The negativeness is a result of its variance being much higher than that of patient 1, together with them being highly correlated. As negative weights seem questionable, a natural

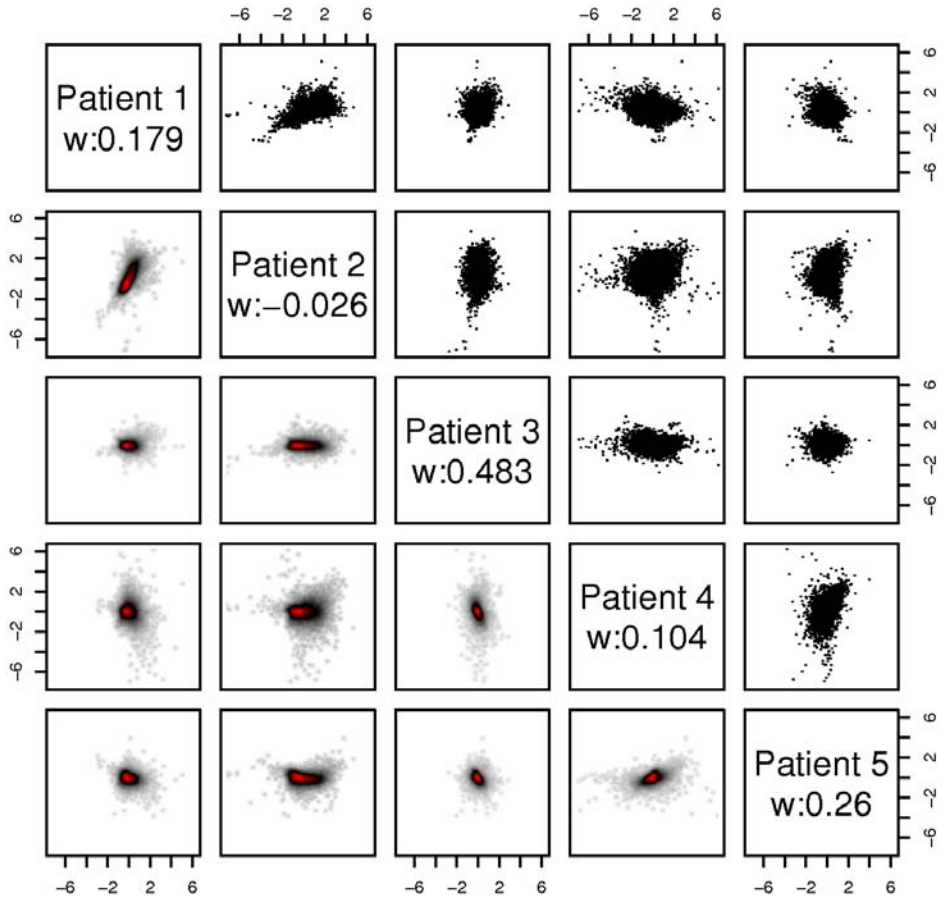


Figure 6: Pair-wise plots of the $\log_2$-ratios of the patients in the Polyp dataset. The plots to the lower-left show two-dimensional kernel density estimates of the distribution of $\log_2$-ratios in each pair of patients. This provides information in the central areas where the corresponding scatterplots are solid black (cf. Figure 6 in Huber et al., 2003). The colour-scale is, in increasing level of density: white, grey, black and red.

solution is to remove patient 2, which was done in (Benson et al., 2004). Beside the result of the very low weight for patient 2, the other patients receive distinctly different weights, which is interesting.

	Patient				
Patient	1	2	3	4	5
1	0.300	0.493	0.000	-0.012	-0.067
2	*0.822*	1.200	0.004	0.041	-0.157
3	*0.002*	*0.012*	0.091	-0.071	-0.055
4	*-0.038*	*0.067*	*-0.417*	0.319	0.102
5	*-0.291*	*-0.340*	*-0.434*	*0.430*	0.178

Table 5: Estimated covariance-correlation matrix, Σ, for patients in the Polyp dataset. (Correlations in italic, covariances in non-italic.)

	Patient				
Removed genes	1	2	3	4	5
none	0.179	-0.026	0.483	0.104	0.260
5%	0.181	-0.025	0.481	0.104	0.259
10%	0.180	-0.024	0.482	0.103	0.259
50%	0.157	-0.015	0.506	0.100	0.252

Table 6: Weights for the patients in the Polyp dataset. Different numbers of potentially regulated genes were removed, to check their potential influence in the estimation of Σ. Potential regulation was measured by minimal absolute $\log_2$-ratio among the patients.

The hyperparameter α, related to the spread of the gene-specific variance scaling factors, c_g, was estimated to 1.97, giving infinite variance for the distribution of c_g. The fit of S_g given $\alpha = 1.97$ was very good (see Figure 10 in the Appendix).

As in the Cardiac dataset, the weights were steadily estimated when potentially regulated genes were removed in the estimation of the covariance matrix Σ (see Table 6). The estimated correlations between patients 3, 4 and 5 were reduced somewhat. Removing 5% of the genes reduced those correlations by 0.03-0.04 and removing 10% reduced them by 0.06-0.07. The high correlation between patient 1 and 2 was only slightly reduced (<0.03), even when 50% of the genes were removed.

Examining the observed values of the statistic, T_g, compared to the expected null distribution (see Figure 7) reveals a good overall concordance. Some genes have a more extreme T_g than can be explained by the null distribution, which points toward many of them being regulated by the treatment (see the qq-plot in Figure 7).

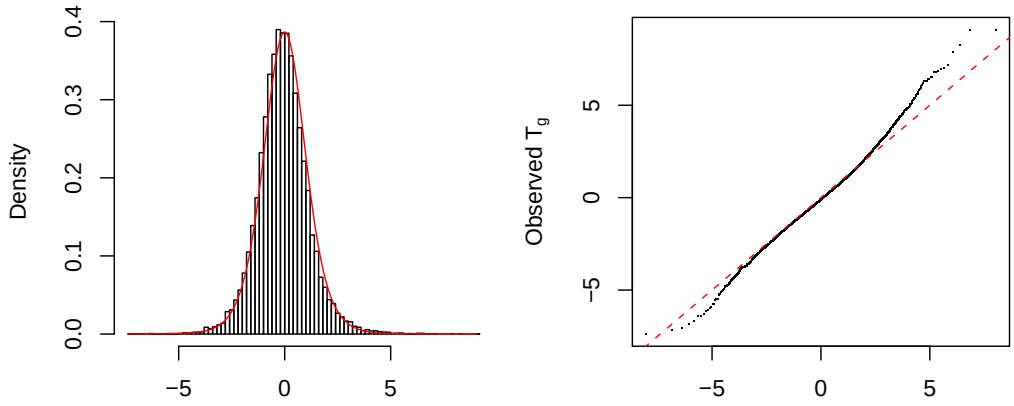


Figure 7: To the left, a histogram of the observed T_g-values together with the density of the null distribution (in red), in the Polyp dataset. To the right, a quantile-quantile plot where the observed values of T_g are plotted against the quantiles of T_g under the null hypothesis. The central part of the empirical distribution follows the identity line well, showing good concordance with the null distribution. For extreme T_g-values, the observations clearly deviate from the predicted ones, pointing at the existence of regulated genes.

6.3 Swirl dataset

In the Swirl experiment (Dudoit and Yang, 2003), one goal was to identify genes that are differentially expressed in zebrafish carrying a point mutated SRB2 gene, compared to ordinary, wild-type zebrafish. SRB2 and one of its known targets, Dlx3 are expected to be highly differentially expressed in this experiment, thus these genes should be highly ranked using WAME. The Swirl dataset has been examined in Smyth (2004).

The dataset consists of four two-colour cDNA microarrays with 8448 spots, with publicly available data. We used standard pre-processing to compensate for effects such as background and dye bias. Background correction *subtract* and within-array normalisation *print tip loess* were used in

the LIMMA package (Smyth et al., 2003). Between-array scale normalisation (Yang et al., 2002) was not performed in contrast to the analysis in Smyth (2004). When including between-array scale normalisation in combination with LIMMA in the simulation study of Section 5.1 the performance was not notably increased (results not shown). However, the model used for simulation leaves the signals unaffected when noise levels varies, which may be questionable for some sources of variation.

Making one scatter plot of the $\log_2$-ratios for each pair of arrays (Figure 8) indicates that array 2 is less variable than the others, while the genes with lowest $\log_2$-ratio on array 1 seem to be outliers, since they are not extreme in any other array. Examining the estimated covariance matrix (see Table 7), array 2 indeed receives the highest variance. In addition, there are substantial correlations between arrays 1 and 3, 2 and 4 and 3 and 4, which is also indicated by the scatter-plots (Figure 8).

	Array			
Array	1	2	3	4
1	0.128	0.007	0.079	0.017
2	*0.066*	0.086	-0.002	0.038
3	*0.489*	*-0.017*	0.203	0.076
4	*0.136*	*0.371*	*0.482*	0.124

Table 7: Estimated covariance-correlation matrix, Σ, for the arrays in the Swirl dataset. (Correlations in italic, covariances in non-italic.)

When re-performing the estimation of Σ after removing potentially regulated genes (in analogy with the analyses of the Polyp and Cardiac datasets), the correlations were decreased somewhat. Removing 5% of the genes decreased the three high correlations by 0.02-0.06, while removing 10% decreased them by 0.04-0.08. However, the corresponding weights only changed marginally (see Table 8).

The hyperparameter α was estimated to 1.89. Further analysis of the dataset shows that the distribution of S_g fits the predicted distribution of S_g well given $\alpha = 1.89$ (see Figure 11 in in the Appendix). The observed values of the statistic, T_g, seem to fit the null distribution well (see Figure 9).

Since the point mutated gene, SRB2 and one of its known targets, Dlx3, are expected to be highly differentially expressed, their actual ranking is of interest. In Table 9 below, the top 20 genes as ranked by WAME are listed. The values of some widely used statistics are included for comparison. The rankings by WAME and the moderated t-statistic (Smyth et al., 2003) are

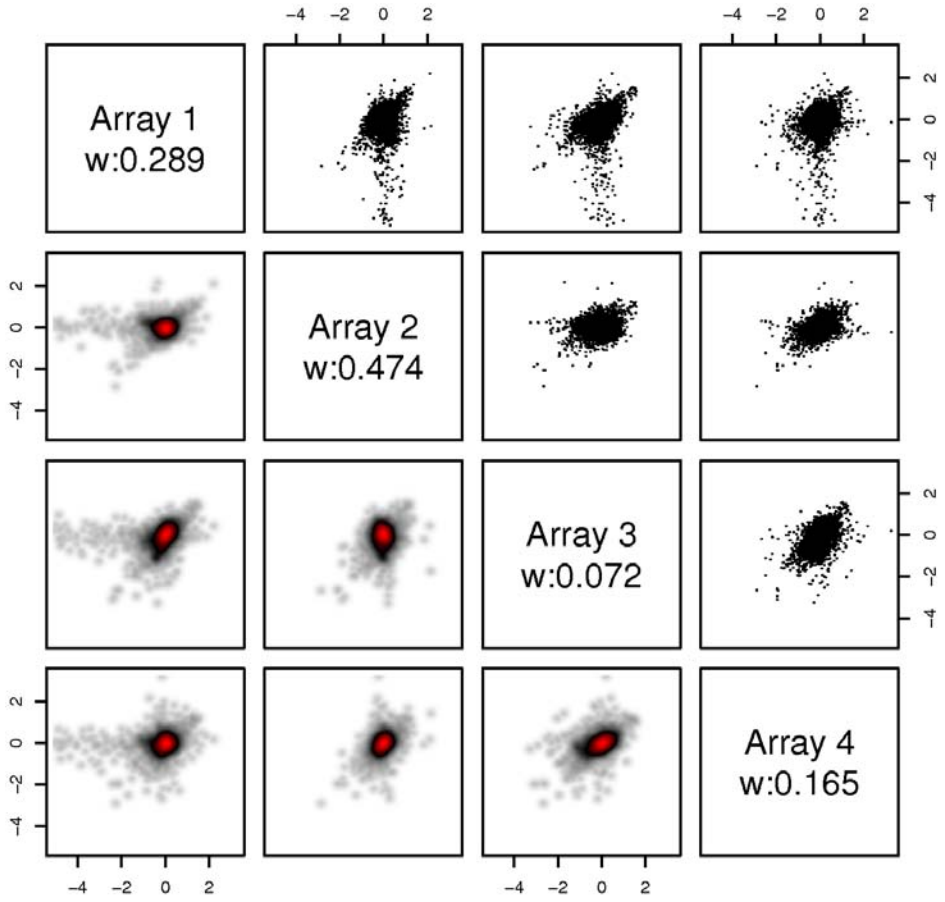


Figure 8: Pair-wise plots of the $\log_2$-ratios of the arrays in the Swirl dataset. The plots to the lower-left show two-dimensional kernel density estimates of the distribution of $\log_2$-ratios in each pair of patients. This provides information in the central areas where the corresponding scatterplots are solid black (cf. Figure 6 in Huber et al., 2003). The colour-scale is, in increasing level of density: white, grey, black and red.

	Array			
Removed genes	1	2	3	4
none	0.289	0.474	0.072	0.165
5%	0.288	0.469	0.076	0.166
10%	0.290	0.462	0.075	0.173
50%	0.282	0.447	0.087	0.184

Table 8: Weights for the arrays in the Swirl dataset. Different numbers of potentially regulated genes were removed, to check their potential influence in the estimation of Σ. Potential regulation was measured by minimal absolute $\log_2$-ratio among the arrays.

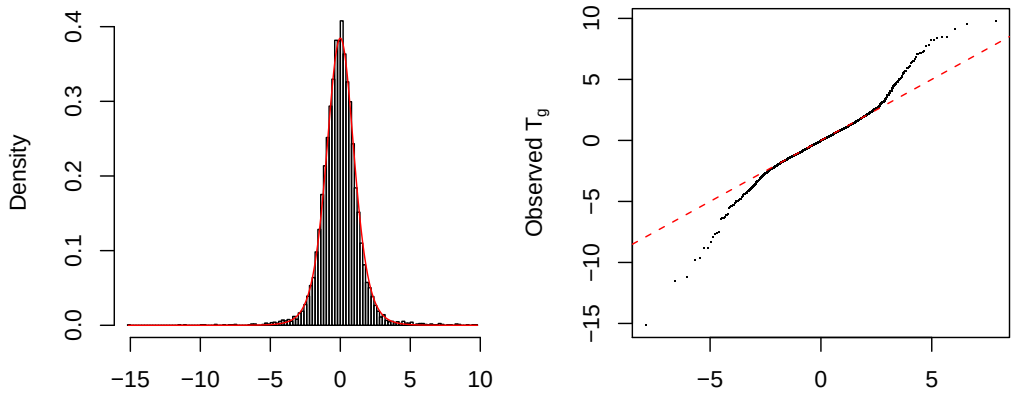


Figure 9: To the left, a histogram of the observed T_g-values together with the density of the null distribution (in red), in the Swirl dataset. To the right, a quantile-quantile plot where the observed values of T_g are plotted against the quantiles of T_g under the null hypothesis. The central part of the empirical distribution follows the identity line well, showing good concordance with the null distribution. For extreme T_g-values, the observations clearly deviate from the predicted ones, pointing at the existence of regulated genes.

quite similar, while the rankings by the ordinary t-statistic and the average $\log_2$-ratio (i.e. fold change) are rather different than the one by WAME, which was expected. All four spots for the two validated genes are included in WAME's top 20 list (see Table 9). It could be noted that the ordinary t-statistics and LIMMA's moderated t-statistics are both generally numerically larger than the WAME values. One reason for this seems to be that the reference distributions of the ordinary t-statistics and LIMMA's moderated t-statistics show q-q plots that differ from the reference distribution by a higher slope also in the central part (data not shown but compare the figure on page 24 in Smyth et al. (2003)) in contrast to the plot in the right part of Figure 9. We have also performed a simulation study with a covariance matrix as in Table 7 and with 10% of the genes differentially expressed. It shows better ROC curves with WAME than with the other two methods (data not shown) in a similar way as in Figure 2.

7 Discussion

A problem with the microarray technology is that it involves several consecutive steps, each exhibiting large quality variations. Thus there is a strong need for quality assessment and quality control to handle occurrences of poor quality. In this paper, we introduce the WAME method for the analysis of paired microarray experiments, which aims at estimating array- or repetition-wide quality deviations and integrates these estimates in the statistical analysis.

The quality deviations are modelled here as different variances for different repetitions (e.g. arrays) as well as correlations between them in a covariance matrix Σ, catching both unequal precision and systematic errors. Genes have different variability (biological and technical), which is modelled by a gene-specific variance scaling factor c_g. Given this structure, the pair-wise measured $\log_2$-ratios for each gene are assumed to be normally distributed.

Estimation of the covariance matrix is complicated by the gene-specific scaling factors and unknown differential expressions μ_g. We assume that most genes are not differentially expressed ($\mu_g = 0$) and the gene-specific scaling factors are removed by a transformation. A scaled version of Σ is estimated by numerical maximum likelihood. The assumed restricted differential expression restrains the experimental setups that can be analysed, but similar assumptions are made in procedures that have become *de facto* standard in the (preceding) normalisation step.

Since most microarray experiments contain only a few repetitions, the

Name	ID	average $\log_2$-ratio	ordinary t-statistic	moderated t-statistic	WAME
fb85d05	18-F10	-2.66	-18.41	-20.79	-15.15
fb58g10	11-L19	-1.60	-14.32	-14.15	-11.51
control	Dlx3	-2.19	-15.91	-17.57	-11.17
control	Dlx3	-2.19	-13.58	-16.08	-9.84
fb24g06	3-D11	1.32	19.52	13.62	9.80
fb54e03	10-K5	-1.20	-25.74	-13.11	-9.66
fc22a09	27-E17	1.26	24.76	13.68	9.50
fb40h07	7-D14	1.35	14.15	12.69	9.12
fb85a01	18-E1	-1.29	-17.35	-13.01	-8.81
fb87f03	18-O6	-1.08	-27.90	-12.06	-8.80
fb37e11	6-G21	1.23	14.37	11.94	8.47
fb94h06	20-L12	1.28	15.41	12.54	8.46
fb87d12	18-N24	1.28	12.96	11.87	8.39
control	BMP2	-2.24	-8.63	-11.78	-8.33
fc10h09	24-H18	1.20	15.05	11.92	8.23
fb85f09	18-G18	1.29	11.50	11.38	8.22
control	BMP2	-2.33	-8.37	-11.58	-7.95
fb26b10	3-I20	1.09	15.50	11.17	7.81
fb37b09	6-E18	1.31	11.57	11.55	7.78
fc22f05	27-G10	-1.19	-10.42	-10.44	-7.70

Table 9: The top 20 most probably regulated genes in the Swirl dataset according to WAME.

estimates of the gene-specific variance scaling factors c_g are imprecise, which may lead to false conclusions. An empirical Bayes approach is used with an inverse gamma prior distribution that moderates extreme estimates similar to (Baldi and Long, 2001; Lönnstedt and Speed, 2002; Smyth, 2004). The hyperparameter α determining the spread of the prior distribution is estimated by numerical maximum likelihood together with the scale of the previously estimated arbitrarily scaled Σ.

In the present paper, quality is modelled in a general manner by the covariance structure matrix Σ. In some microarray experiments, additional information is available, for example, shared sources of variation may be known. Quantitative quality measures may also be available, e.g. spot shape features or residuals from the fitting of probe-level models (Bolstad, 2004). It is possible to explicitly model some such sources of variation, for example using random or fixed effects (cf. Bakewell and Wit (2005)) and to include quality measures as covariates. However, such models would likely focus on some of the clearer sources of variation but leave out more involved and hard modelled sources. One can view our method as an attempt to identify the effects on the single gene level of those variability sources, with the prior distribution modelling the noise structure of a random gene from the whole gene population. An approach combining explicit modelling with a general covariance structure would be interesting as future work.

To identify differentially expressed genes a likelihood-ratio test is derived, resulting in the *weighted moderated t-statistic*, which is a generalisation of the moderated t-statistic in Smyth (2004). The estimated covariance matrix Σ is used to produce both weights for the different repetitions and gene-specific variance estimates. The weighted mean is the estimate of differential expression with minimal variance.

As discussed above, array- or repetition-wide quality deviations in all steps leading to the observed $\log_2$-ratios are estimated and incorporated in the analysis. The current paper is restricted to paired two-sample settings where most genes are non-differentially expressed. A generalisation similar to Smyth (2004) should be possible for experiments with pairwise measurements. The scaled estimate of the covariance matrix Σ could be calculated according to the procedure in the current paper (cf. Section 4.1). The unknown scale of the covariance estimate, as well as the parameter α of the prior distribution for the gene-specific variance scales, could be estimated utilising generalised residual sums of squares for all genes, appropriately defined through the norm determined by Σ (cf. S_g in Section 4.2). Tests for single or multiple identifiable linear combinations of expected values could be derived as in the current paper to get weighted moderated t-statistics and

modified F-statistics. Work on generalisations, with simulated and real data sets is in progress.

A simulation study was done to compare the performance of WAME to four published methods. On data without correlations and with equal variances between repetitions, WAME performs as well as the moderated t-statistic which assumes this structure. When correlations and/or unequal variances are included, WAME performs better than the other methods. In one case, using WAME results in almost a third less false positives which can correspond to hundreds of genes. Evaluating the point estimator of the covariance matrix Σ revealed good precision and accuracy when no regulated genes were present. Including 10% regulated genes resulted in a bias, which was partly handled by removing genes likely to be regulated. In both cases estimation of the hyperparameter α was nearly unbiased and accurate. The estimate of Σ was essentially unbiased when heavy tails were introduced in contrast to the estimate of α which was 1.55 instead of 2. As a practical consideration, filtering of seemingly regulated genes may be appropriate when a relatively large number of genes can be expected to be regulated. However, results from real and simulated data indicate that such filtering results in largely unchanged weights, reducing its importance. Also, in the cases studied the unfiltered statistic is slightly conservative (results not shown).

Three real datasets were analysed: the ischemic part of the dataset of Hall et al. (2004), the dataset of Benson et al. (2004) and the *Swirl* dataset (Dudoit and Yang, 2003). In all cases, relevant correlations and differences in precision between replicates were found, even in the first dataset which had been quality controlled using several available methods. The exact origin of the correlations is an interesting, open question. In the second dataset one previously identified outlier was practically removed by WAME. In the Swirl dataset, expected differentially expressed genes are ranked among the top 20. Relevant empirical distributions showed good fit to the theoretical distributions, indicating that the family of prior distributions for c_g is flexible enough and that the normality assumption is satisfactory.

The model used in WAME is optimistic in several ways. Exact normality is not to be expected and the independence between the genes is hard to fully justify. The noise structure may also be different for the regulated genes, e.g. if there are several normalising procedures involved in the preprocessing step. This may affect the power, which points towards the use of a moderated impact of Σ on the weights in the final analysis. Thus, even if simulations under the model assumptions show highly promising results, there are many experimental situations where the model assumptions may not be justified. We intend to look further into different robustness questions

for model deviations in the future.

The role of microarray experiments is often to test for regulation of tens of thousands of genes as an exploratory tool to derive candidate ranking lists of potentially regulated genes, which in subsequent steps will be biologically interpreted and validated by more precise techniques. We find that our approach competes well with other methods in the production of such lists.

To summarise, WAME estimates and integrates array- or repetition-wide quality deviations in the analysis of paired microarray experiments. An empirical Bayes approach is used to moderate the gene-specific variance estimates, resulting in a weighted moderated t-statistic with a derived distribution. The performance of WAME has been evaluated on both simulated and real microarray data. The simulations show a considerable advantage relative to four other methods studied, particularly for data with unequal variances or correlations among repetitions. The three real datasets studied indicate that data with unequal variances or correlations should be quite common. The model controls with diagnostic plots also show satisfactory results for all three real datasets.

Appendix

Additional Figures

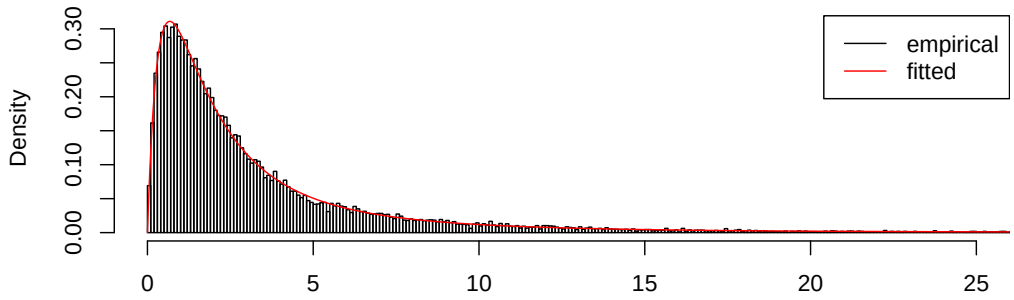


Figure 10: Empirical distribution of S_g in the Polyp dataset, together with the density of S_g given $\alpha = 1.97$.

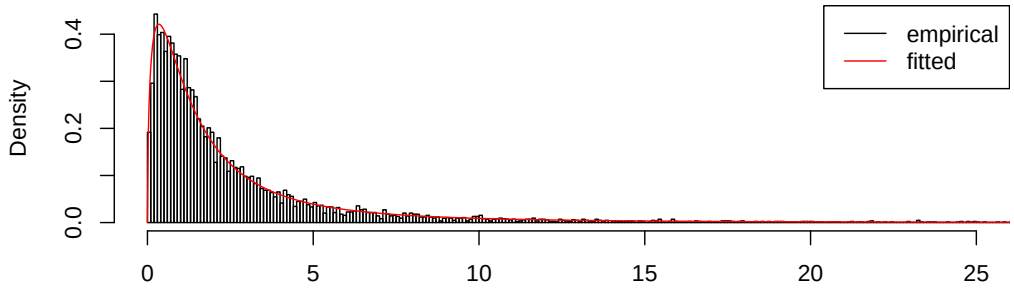


Figure 11: Empirical distribution of S_g in the Swirl dataset, together with the density of S_g given $\alpha = 1.89$.

Mathematical details

We observe $\mathbf{X}_g = (X_{g1}, \ldots, X_{gN_I})$ where $g = 1, \ldots, N_G$. Let Σ be a covariance structure matrix for the N_I repetitions, c_g a set of gene-specific variance scaling factors and α a hyperparameter determining the shape of the prior distribution for c_g. Then for fixed μ_g, Σ and α,

$$\begin{aligned} c_g &\sim \Gamma^{-1}(\alpha, 1), \\ \mathbf{X}_g \mid c_g &\sim \mathbf{N}_{N_I}(\mu_g \mathbf{1}, c_g \Sigma) \end{aligned}$$

and all variables corresponding to different genes are assumed independent.

Estimation of a scaled version of the matrix Σ

Assume that $\mu_g = 0$ for all g. Under this assumption, it is possible to derive a scale independent estimate of the covariance matrix Σ by a transformation of the vector $\mathbf{X}_g$. This is done as follows (the index g is dropped to increase the readability). Let $\mathbf{U} = (U_1, \ldots, U_{N_I})$ where

$$U_i = \begin{cases} X_1 & \text{if } i = 1 \\ X_i / X_1 & \text{if } 2 \leq i \leq N_I \end{cases}.$$

The inverse becomes

$$X_i = \begin{cases} U_1 & \text{if } i = 1 \\ U_i U_1 & \text{if } 2 \leq i \leq N_I \end{cases}$$

and the Jacobian can be derived to

$$J(u_1, \ldots, u_{N_I}) = u_1^{N_I - 1},$$

so for $\mathbf{U} \in \mathbb{R}^{N_I}$ the density becomes

$$\begin{aligned} f_{\mathbf{U} \mid c, \Sigma}(\mathbf{u}) &= f_{\mathbf{X} \mid c, \Sigma}\left(\mathbf{x}(\mathbf{u})\right) |J(\mathbf{u})| \\ &= (2\pi)^{-N_I/2} c^{-N_I/2} |\Sigma|^{-1/2} |u_1|^{N_I - 1} e^{-\frac{u_1^2}{2c} v^{\mathrm{T}} \Sigma^{-1} v}. \end{aligned}$$

where $v = (1, u_2, \ldots, u_{N_I})^{\mathrm{T}}$. Integration over u_1 yields

$$\begin{aligned} f_{U_2, \ldots, U_{N_I} \mid \Sigma}(u_2, \ldots, u_{N_I} \mid \Sigma) &= \int_{-\infty}^{\infty} f_{\mathbf{U} \mid c, \Sigma}(\mathbf{u} \mid c, \Sigma)\, du_1 \\ &= C |\Sigma|^{-1/2} \left[v^{\mathrm{T}} \Sigma^{-1} v \right]^{-N_I/2}, \end{aligned} \tag{14}$$

where C is a normalisation constant and v is defined as above. This density is scale invariant with respect to the parameter Σ in the sense that for any scalar λ,

$$f_{U_2,\ldots,U_{N_I} \mid \Sigma}(u_2,\ldots,u_{N_I}|\lambda\Sigma) = f_{U_2,\ldots,U_{N_I} \mid \Sigma}(u_2,\ldots,u_{N_I}|\Sigma).$$

Thus, it is also independent of c and under the assumption of independent genes, the log-likelihood function becomes

$$l(\Sigma) = C' - \frac{N_G}{2}\log\left(|\Sigma|\right) - \frac{N_I}{2}\sum_{g=1}^{N_g}\log\left(v_g^{\mathrm{T}}\Sigma^{-1}v_g\right),$$

where C' is a constant that is independent of Σ. Numerical maximisation yields a scaled version of Σ, denoted Σ^*. Here the first element in the first row of Σ^* is fixed to one.

Estimation of the hyperparameter α and the scale λ

From the model assumptions, we know that

$$c_g \sim \Gamma^{-1}(\alpha, 1).$$

Assume that Σ is known and define

$$S_g = (A\mathbf{X}_g)^{\mathrm{T}}(A\Sigma A^{\mathrm{T}})^{-1}A\mathbf{X}_g,$$

where A is a contrast matrix, i.e. a matrix of dimension $N_I - 1 \times N_I$, with full rank and with each row sum equal to 0. It follows that

$$S_g \sim c_g \times \chi^2_{N_I-1}\,.$$

The unconditional distribution of S_g can be derived by integrating over c_g, i.e.,

$$\begin{aligned} f_{S_g \mid \alpha}(s_g) &= \int_0^\infty f_{S_g \mid c_g}(s) f_{c_g \mid \alpha}(c_g)\, dc_g \\ &= \frac{1}{2}\frac{(s/2)^{(N_I-1)/2-1}}{\Gamma(\alpha)\,\Gamma((N_I-1)/2)}\int_0^\infty c^{-\alpha-(N_I-1)/2-1}e^{-(s/2+1)/c_g}\, dc_g \\ &= \frac{1}{2}\frac{\Gamma(\alpha+(N_I-1)/2)}{\Gamma(\alpha)\Gamma((N_I-1)/2)}\frac{(s/2)^{(N_I-1)/2-1}}{[1+s/2]^{\alpha+(N_I-1)/2}}\,. \end{aligned}$$

This is a beta prime distribution (also called a beta distribution of the second kind) (Johnson et al., 1995) with parameters $N_I - 1$ and α which is denoted

$\beta'(N_I-1,\alpha)$. Since only a scaled version of Σ, denoted Σ^*, is assumed known from the primary estimation step, the following entities are defined. Let

$$\begin{aligned}\Sigma^* &= \lambda\Sigma \\ S_g^* &= (A\mathbf{X}_g)^{\mathrm{T}}(A\Sigma^* A^{\mathrm{T}})^{-1}A\mathbf{X}_G = S_g/\lambda,\end{aligned}$$

where λ is the unknown scale for Σ^*. It follows that

$$S_g^* \sim 2/\lambda \times \beta'(N_I - 1, \alpha).$$

The log likelihood function for S_g^* can be derived to

$$\begin{aligned}l(\alpha,\lambda|\{s_g\}_{g=1}^{N_G}) = C + N_G\left[(N_I-1)/2\log(\lambda) + \log\Gamma(\alpha+(N_I-1)/2) - \log\Gamma(\alpha)\right] \\ - (\alpha+(N_I-1)/2)\sum_{g=1}^{N_G}\log(s_g\lambda/2+1).\end{aligned}$$

Numerical maximum likelihood is used to estimate α and λ, which together with Σ^* can be used to calculate an estimate for Σ.

Inference about μ_g

The hypotheses that are interesting to test are if different genes are regulated or not, that is for each g,

$$\begin{aligned}H_0 &: \text{ gene } g \text{ is not regulated } (\mu_g = 0) \\ H_A &: \text{ gene } g \text{ is regulated } (\mu_g \neq 0).\end{aligned}$$

To test these hypotheses a maximum likelihood ratio (LRT) test is derived. For each g, we reject H_0 if

$$\frac{\sup\limits_{\mu_g\neq 0} L\left(\mu_g|\mathbf{x}_g\right)}{L\left(0|\mathbf{x}_g\right)} \geq k,$$

where $1 \leq k < \infty$. The likelihood L can be calculated by integration over c_g, i.e.

$$\begin{aligned}L\left(\mu_g|\mathbf{x}\right) &= \int f_{\mathbf{X}\,|\,\mu_g,c_g,\Sigma}(\mathbf{x}) f_{c_g\,|\,\alpha}(c_g)\,dc_g \\ &= (2\pi)^{-N_I/2}\left|\Sigma\right|^{-1/2}\frac{\Gamma(N_I/2+\alpha)}{\Gamma(\alpha)}\left[\frac{\left(\mathbf{x}_g-\mu_g\mathbf{1}\right)^{\mathrm{T}}\Sigma^{-1}\left(\mathbf{x}_g-\mu_g\mathbf{1}\right)}{2}+1\right]^{-N_I/2-\alpha}.\end{aligned}$$

To calculate the numerator in the likelihood ratio we need to maximise L over μ_g, which is the same as minimising

$$(\mathbf{x}_g - \mu_g \mathbf{1})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_g - \mu_g \mathbf{1}) .$$

A little algebra shows that this optimum corresponds to the argument

$$\hat{\mu}_g = \frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\mathbf{x}_g .$$

We will use $\bar{x}_g^w$ to denote this weighted sum and it can be shown to be the weighted mean with least variance. The maximum value of the likelihood function becomes

$$L(\bar{x}_g^w|\mathbf{x}_g) = (2\pi)^{-N_I/2} |\Sigma|^{-1/2} \frac{\Gamma(N_I/2+\alpha)}{\Gamma(\alpha)} \left[\frac{\mathbf{x}_g^{\mathrm{T}}\Sigma^{-1}\mathbf{x}_g - (\bar{x}_g^w)^2\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}{2} + 1\right].$$

Using this, the likelihood ratio test statistic can be rewritten as

$$\begin{aligned}
\frac{L\left(\bar{x}_g^w|\mathbf{x}_g\right)}{L\left(0|\mathbf{x}_g\right)} &= \left[\frac{\mathbf{x}_g^{\mathrm{T}}\Sigma^{-1}\mathbf{x}_g + 2}{\mathbf{x}_g^{\mathrm{T}}\Sigma^{-1}\mathbf{x}_g - \left(\bar{x}_g^w\right)^2 \mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1} + 2}\right]^{N_I/2+\alpha} \\
&= \left[1 + \frac{\left(\bar{x}_g^w\right)^2 \mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}{\mathbf{x}_g\Sigma^{-1}\mathbf{x}_g - \left(\bar{x}_g^w\right)^2 \mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1} + 2}\right]^{N_I/2+\alpha} \\
&= \left[1 + \frac{\left(\bar{x}_g^w\right)^2 \mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}{\left(\mathbf{x}_g - \left(\bar{x}_g^w\right)\mathbf{1}\right)^{\mathrm{T}}\Sigma^{-1}\left(\mathbf{x}_g - \left(\bar{x}_g^w\right)\mathbf{1}\right) + 2}\right]^{N_I/2+\alpha} \\
&= \left[1 + \frac{(\bar{x}_g^w)^2 \mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}{\left(A_{\mathbf{w}}\mathbf{x}_g\right)^{\mathrm{T}}\Sigma^{-1}\left(A_{\mathbf{w}}\mathbf{x}_g\right) + 2}\right]^{N_I/2+\alpha}
\end{aligned}$$

where $A_{\mathbf{w}}$ is the contrast matrix

$$A_{\mathbf{w}} = \begin{pmatrix} 1-w_1 & -w_2 & -w_3 & \dots & -w_{N_I} \\ -w_1 & 1-w_2 & -w_3 & \dots & -w_{N_I} \\ \dots & \dots & \dots & \dots & \dots \\ -w_1 & -w_2 & -w_3 & \dots & 1-w_{N_I} \end{pmatrix}$$

and w_i is the i:th element of the vector

$$\frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}.$$

The next step is to show that

$$(A_{\mathbf{w}}\mathbf{x}_g)^{\mathrm{T}}\,\Sigma^{-1}\,(A_{\mathbf{w}}\mathbf{x}_g) = s_g\ . \tag{15}$$

To do that, we first note that for any pair of contrast matrices A_1 and A_2 with N_I columns and of rank $N_I - 1$, with each row sum equal to zero,

$$(A_1\mathbf{x}_g)^{\mathrm{T}}\,(A_1\Sigma A_1^{\mathrm{T}})^{-}\,(A_1\mathbf{x}_g) = (A_2\mathbf{x}_g)^{\mathrm{T}}\,(A_2\Sigma A_2^{\mathrm{T}})^{-}\,(A_2\mathbf{x}_g)\,.$$

Here a generalised inverse is used, defined as $BB^{-}B = B$, which gives

$$B^{-1} = B^{-}$$

when B is invertible. Now,

$$s_g = (A\mathbf{x}_g)^{\mathrm{T}}\,(A\Sigma A^{\mathrm{T}})^{-1}\,(A\mathbf{x}_g) = (A_{\mathbf{w}}\mathbf{x}_g)^{\mathrm{T}}\,(A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}})^{-}\,(A_{\mathbf{w}}\mathbf{x}_g)\,,$$

so we can prove (15) by showing that

$$(A_{\mathbf{w}}\mathbf{x}_g)^{\mathrm{T}}\,\Sigma^{-1}\,(A_{\mathbf{w}}\mathbf{x}_g) = (A_{\mathbf{w}}\mathbf{x}_g)^{\mathrm{T}}\,(A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}})^{-}\,(A_{\mathbf{w}}\mathbf{x}_g)\,.$$

Since $A_{\mathbf{w}}$ is idempotent, this is the same as proving that

$$(A_{\mathbf{w}}\Sigma A_{\mathbf{w}^{\mathrm{T}}})^{-} = A_{\mathbf{w}}^{\mathrm{T}}\Sigma^{-1}A_{\mathbf{w}}\ .$$

Writing $A_{\mathbf{w}}$ as

$$A_{\mathbf{w}} = I - \mathbf{1}\frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}$$

it follows that

$$\begin{aligned}
A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}}\left(A_{\mathbf{w}}^{\mathrm{T}}\Sigma^{-1}A_{\mathbf{w}}\right)A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}} =& A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}}\Sigma^{-1}A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}} \\
=& \left[I - \mathbf{1}\frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right]\Sigma\left[I - \mathbf{1}\frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right]^{\mathrm{T}}\Sigma^{-1} \\
&\times\left[I - \mathbf{1}\frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right]\Sigma\left[I - \mathbf{1}\frac{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right]^{\mathrm{T}} \\
=& \left[\Sigma - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right]\Sigma^{-1}\left[\Sigma - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right] \\
=& \Sigma - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}} = A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}}\ .
\end{aligned}$$

Thus,

$$(A_{\mathbf{w}}\Sigma A_{\mathbf{w}}^{\mathrm{T}})^{-} = A_{\mathbf{w}}^{\mathrm{T}}\Sigma^{-1}A_{\mathbf{w}}$$

and (15) is proved.

Using this result, we can write the LRT as

$$\frac{|\bar{x}_g^w|}{\sqrt{s_g + 2}} \geq k', \tag{16}$$

where $0 \leq k' < \infty$ is a new constant. To derive the distribution of the statistic that corresponds to (16) under the null hypothesis, we proceed as follows. Let

$$T_g = \sqrt{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}\,(N_I - 1 + 2\alpha)}\;\frac{\bar{X}_g^w}{\sqrt{S_g + 2}}\;.$$

Then since

$$\bar{X}_g^w \sim \mathbf{N}\left(0, \frac{c_g}{\mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}\right)$$

it can be shown that $\bar{X}_g^w$ is independent to all elements of $A_{\mathbf{w}}\mathbf{X}_g$ and thus to S_g. Furthermore,

$$T_g = \frac{\bar{X}_g^w / \sqrt{c_g / \mathbf{1}^{\mathrm{T}}\Sigma^{-1}\mathbf{1}}}{\sqrt{S_g/c_g + 2/c_g}/\sqrt{N_I - 1 + 2\alpha}},$$

where the numerator is independent of S_g and has the same normal distribution conditionally on all c_g (and thus also unconditionally), showing that the denominator in this ratio expression is independent of the numerator. A similar argument shows that S_g/c_g and $2/c_g$ are independent, and since they are chi-square distributed with $N_I - 1$ and 2α degrees of freedom respectively, the sum is chi-square distributed with $N_I - 1 + 2\alpha$ degrees of freedom. Hence, under the null hypothesis, T_g is a t-distribution with $N_I - 1 + 2\alpha$ degrees of freedom,

$$T_g \mid \Sigma, \alpha \;\sim\; t_{N_I - 1 + 2\alpha}\;.$$

We call T_g the weighted moderated t-statistic.

References

D.J. Bakewell and E. Wit. Weighted analysis of microarray gene expression using maximum-likelihood. *Bioinformatics*, 21(6):723–729, 2005.

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. Benson, L. Carlsson, M. Adner, M. Jernås, M. Rudemo, A. Sjögren, P.A. Svensson, R. Uddman, and Cardell L.O. Gene profiling reveals increased expression of uteroglobin and other anti-inflammatory genes in glucocorticoid-treated nasal polyps. *Journal of Allergy and Clinical Immunology*, 113(6):1137–1143, 2004.

B.M. Bolstad. *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. PhD thesis, University of California, Berkeley, California, 2004.

D.-T. Chen. A graphical approach for quality control of oligonucleotide array data. *Journal of Biopharmaceutical Statistics*, 14(3):591–606, 2004.

S. Dudoit and J.Y.H. Yang. Bioconductor R packages for exploratory data analysis and normalization of cDNA microarray data. In G. Parmigiani, E.S. Garett, R.A. Irizarry, and S.L. Zeger, editors, *The Analysis of Gene Expression Data*. Springer, 2003.

C.I. Dumur, S. Nasim, A.M. Best, K.J. Archer, A.C. Ladd, V.R. Mas, D.S. Wilkinson, C.T. Garret, and A. Ferreira-Gonzalez. Evaluation of quality-control criteria for microarray gene expression analysis. *Clinical Chemistry*, 50(11):1994–2002, 2004.

R. Edgar, M. Domrachev, and A.E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

J.L. Hall, S. Grindle, X. Han, D. Fermin, S. Park, Y. Chen, R.J. Bache, A. Mariash, Z. Guan, S. Ormaza, J. Thompson, J. Graziano, S.E. de Sam Lazaro, S. Pan, R.D. Simari, and L.W. Miller. Genomic profiling of the human heart before and after mechanical support with a ventricular assist device reveals alterations in vascular signaling networks. *Journal of Physiological Genomics*, 17(3):283–291, 2004. URL `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GDS558`.

S. Hautaniemi, H. Edgren, P. Vesanen, M. Wolf, A.-K. Jarvinen, O. Yli-Harja, J. Astola, K. Olli, and O. Monni. A novel strategy for microarray quality control using bayesian networks. *Bioinformatics*, 19(16):2031–2038, 2003.

W. Huber, A. von Heydebreck, and M. Vingron. Analysis of microarray gene expression data. In M. et al. Bishop, editor, *Handbook of Statistical Genetics, 2nd Edition.* John Wiley & Sons, 2003.

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

K. Johnson and S. Lin. QA/QC as a pressing need for microarray analysis: meeting report from CAMDA'02. *BioTechniques*, 34(suppl):S62–S63, 3 2003.

N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions Volume 2.* Wiley, 1995.

C. Li and W. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science U S A*, 98:31–36, 2001.

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

J.S. Maritz. *Empirical Bayes Methods.* Methuen & Co. Ltd., 1970.

T. Park, S.-G. Yi, S. Lee, and J.K. Lee. Diagnostic plots for detecting outlying slides in a cDNA microarray experiment. *BioTechniques*, 38(3): 463–471, 2005.

R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2004. URL `http://www.R-project.org`.

H. Robbins. An empirical Bayes approach to statistics. In J. Neyman, editor, *Third Berkeley Symposium on Mathematics and Probability*, pages 157–163, 1956.

C.P. Robert. *The Bayesian Choice.* Springer, 2003.

M.M. Ryan, S.J. Huffaker, M.J. Webster, M. Wayland, T. Freeman, and S. Bahn. Application and optimization of microarray technologies for human postmortem brain studies. *Biological Psychiatry*, 55(4):329–336, 2004.

L. Shi, W. Tong, F. Goodsaid, F.W. Frueh, H. Fang, T. Han, J.C. Fuscoe, and D.A. Casciano. QA/QC: Challenges and pitfalls facing the microarray community and regulatory agencies. *Expert Review of Molecular Diagnostics*, 4(6):761–777, 2004.

G.K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

G.K. Smyth, N.P. Thorne, and J. Wettenhall. *LIMMA: Linear Models for Microarray Data User's Guide*, 2003. URL `http://www.bioconductor.org`.

H. Tomita, M.P. Vawter, D.M. Walsh, S.J. Evans, P.V. Choudary, J. Li, K.M. Overman, M.E. Atz, R.M. Myers, E.G. Jones, S.J. Watson, H. Akil, and W.E. Bunney Jr. Effect of agonal and postmortem factors on gene expression profile: Quality control in microarray analyses of postmortem human brain. *Biological Psychiatry*, 55(4):346–352, 2004.

W. Tong, S. Harris, X. Cao, H. Fang, L. Shi, H. Sun, J. Fuscoe, A. Harris, H. Hong, Q. Xie, R. Perkins, and Casciano D. Development of public toxicogenomics software for microarray data management and analysis. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 549(1-2):241–253, 2004.

X. Wang, S. Ghosh, and S.W. Guo. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, 29(15): e75, 2001.

X. Wang, M.J. Hessner, Y. Wu, N. Pati, and S. Ghosh. Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bioinformatics*, 19(11): 1341–1347, 2003.

Y.H. Yang, S. Dudoit, P. Luu, D. Lin, V. Peng, J. Ngai, and T.P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.

Paper II

Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 10

Quality Optimised Analysis of General Paired Microarray Experiments

Erik Kristiansson* Anders Sjögren†

Mats Rudemo‡ Olle Nerman**

*Mathematical Statistics, Chalmers University of Technology, erikkr@math.chalmers.se

†Mathematical Statistics, Chalmers University of Technology, anders.sjogren@math.chalmers.se

‡Mathematical Statistics, Chalmers University of Technology, rudemo@math.chalmers.se

**Mathematical Statistics, Chalmers University of Technology, nerman@math.chalmers.se

Quality Optimised Analysis of General Paired Microarray Experiments*

Erik Kristiansson, Anders Sjögren, Mats Rudemo, and Olle Nerman

Abstract

In microarray experiments, several steps may cause sub-optimal quality and the need for quality control is strong. Often the experiments are complex, with several conditions studied simultaneously. A linear model for paired microarray experiments is proposed as a generalisation of the paired two-sample method by Kristiansson et al. (2005). Quality variation is modelled by different variance scales for different (pairs of) arrays, and shared sources of variation are modelled by covariances between arrays. The gene-wise variance estimates are moderated in an empirical Bayes approach. Due to correlations all data is typically used in the inference of any linear combination of parameters. Both real and simulated data are analysed. Unequal variances and strong correlations are found in real data, leading to further examination of the fit of the model and of the nature of the datasets in general. The empirical distributions of the test-statistics are found to have a considerably improved match to the null distribution compared to previous methods, which implies more correct p-values provided that most genes are non-differentially expressed. In fact, assuming independent observations with identical variances typically leads to optimistic p-values. The method is shown to perform better than the alternatives in the simulation study.

KEYWORDS: quality control, generalised linear model, experimental design, empirical Bayes, DNA Microarray

---

*We would like to thank Claus Ekstrøm for valuable comments on the manuscript. Erik Kristiansson and Anders Sjögren wish to thank the National Research School in Genomics and Bioinformatics for funding. Olle Nerman wishes to thank University of Canterbury and John Angus Erskine Bequest, New Zeeland for support by an Erskine Visiting Fellowship in spring 2005.

1 Introduction

Microarray experiments involve a series of steps, ranging from selection of biological samples to hybridisation and scanning of arrays, each producing data with varying quality. There is therefore a pressing need for quality control (Johnson and Lin, 2003).

In Kristiansson et al. (2005) an analysis procedure called *Weighted Analysis of Paired Microarray Experiments* (WAME) was proposed for paired two-sample microarray experiments. Quality was modelled as a common covariance structure for all genes, giving each pair of observations a variance estimate and catching shared sources of variation by covariances. To reflect the different variability of different genes, gene-specific scaling factors for the covariance structure matrix were introduced, having inverse gamma prior distribution (cf. Lönnstedt and Speed (2002) and Smyth (2004)). A weighted moderated t-test was derived to identify differentially expressed genes between the two conditions. For three real datasets both distinctly different variances and high correlations were estimated, rendering substantial differences between the array- or patient-specific weights. Furthermore, the empirical distributions of the respective resulting p-values were considerably improved compared to the examined alternative methods.

In the present paper, a generalisation of Kristiansson et al. (2005) is suggested, allowing for general paired experiments. This is now stated in a generalised linear model framework (Arnold, 1980; Smyth, 2004). The covariance structure is general, allowing correlations between all pairs. Tests for contrast type linear combinations of parameters are derived, analogous to the test for differential expression between conditions in the two-sample case. This results in moderated t- and F-tests. Results from analyses of simulated data are presented briefly to assess the benefit of the method in cases where the model assumptions are true.

Two real datasets with multiple conditions are investigated, with interesting results. In one case, a two-colour cDNA microarray experiment is investigated, comparing gene expression of wild-type and knock-out mice to a common reference pool. Here correlations may be expected, since one channel originates from the very same mRNA pool, sharing multiple sources of variation. In the other case, 19 human patients divided into 3 groups are investigated before and after treatment with a ventricular assist device. Some relatively high correlations between measurements from patients in different groups are detected. In tests for differential expression within one group data from other groups is therefore included. The results are compared with the corresponding results of Kristiansson et al. (2005).

2 The generalised linear model

2.1 Model assumptions and formulation

The model introduced in this paper is designed for microarray experiments with paired observations from s different conditions ($s \geq 2$). For each gene $g = 1, \ldots, N_G$ let the vector $\boldsymbol{\gamma}_g = (\gamma_{g1}, \ldots, \gamma_{gs})^{\mathrm{T}}$ contain the expectation of the logarithm (base 2) of the amount of mRNA from each of the s conditions. Assume that $n \geq 2$ pair-wise differences of some of these conditions are observed, denoted by the vector

$$\mathbf{X}_g = (X_{g1}, \ldots, X_{gn}).$$

Let $\boldsymbol{\mu}_g$ be the expectation of the vector $\mathbf{X}_g$ and let D be an $n \times s$ design matrix with rank p such that

$$\boldsymbol{\mu}_g = D\boldsymbol{\gamma}_g.$$

Since all observations are pair-wise differences, D will have row sums equal to zero.

As discussed in Kristiansson et al. (2005), there may exist both differences in precisions and systematic effects between the paired observations and therefore, a gene-independent unknown covariance structure matrix Σ is introduced. The gene-specific variability is modelled by scaling Σ with a factor c_g, which is assumed to be independent for different genes. The vectors $\mathbf{X}_g$ are also assumed to be independent and, conditional on c_g, normally distributed, i.e.,

$$\mathbf{X}_g \mid c_g \sim \mathbf{N}\left(\boldsymbol{\mu}_g, c_g\Sigma\right). \tag{1}$$

The subspace $V \subset \mathbb{R}^n$ will denote the p-dimensional vector space spanned by the columns of D, thus $\boldsymbol{\mu}_g \in V$. Conditional on c_g, this model is sometimes referred to as a generalised linear model (Arnold, 1980).

Many microarray experiments consist of few observations for each gene, which makes gene-specific variance estimates imprecise. Therefore, a prior distribution for c_g is introduced and assumed to be an inverse gamma distribution with unknown shape parameter α and the scale parameter fixed to 1, i.e.

$$c_g \sim \Gamma^{-1}(\alpha, 1).$$

This choice is motivated by the fact that the inverse gamma distribution is a conjugate prior for the variance of a normal distribution. An empirical Bayes approach will be used to estimate the hyperparameter α, a method that has been proven successful in the context of microarray analysis (Baldi and Long, 2001; Lönnstedt and Speed, 2002; Smyth, 2004; Kristiansson et al., 2005).

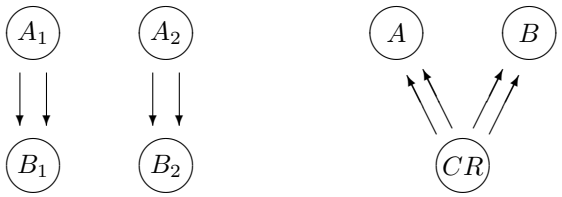
2.2 Examples of parametrisation

The model described above is suitable for a vast number of experimental setups and two examples will now be given. Figure 1(a) shows an illustration of a direct comparison (Churchill, 2002) where the conditions A_1 and A_2 are compared against B_1 and B_2 respectively. Two observations are used for each pair of conditions resulting in four observations in total. One way to parametrise such a design is to let

$$\boldsymbol{\gamma}_g = (\gamma_{A_1}, \gamma_{A_2}, \gamma_{B_1}, \gamma_{B_2})^{\mathrm{T}}$$

and use the design matrix

$$D = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & -1 \end{pmatrix}.$$



(a) (b)

Figure 1: Two examples of experimental setup in microarray analysis; (a) is a direct comparison and (b) a common reference design. Circles corresponds to different conditions and arrows corresponds to pair-wise observations between the conditions. The heads of the arrows indicate which of the conditions that are numerators in the pair-wise log-ratios (i.e. colored "red").

Note that the model suggested in Kristiansson et al. (2005) only works for direct comparisons between two conditions and is a special case of the more general model described here.

Another widely used experimental setup is the common reference design where two or more conditions are compared through one or more references

(Churchill, 2002; Steibel and Rosa, 2005). In Figure 1(b) conditions A and B are compared through a single reference called CR. A suitable parametrisation for this setup is obtained by putting

$$\boldsymbol{\gamma}_g = (\gamma_A, \gamma_B, \gamma_{CR})^{\mathrm{T}}$$

and then choosing the design matrix D as

$$D = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 1 & -1 \end{pmatrix}.$$

2.3 Notation

We end this section with some words about notation. In this paper, $\mathbb{R}^n$ will be regarded as a vector space and $\|\mathbf{X}\|$ will denote the Euclidean norm,

$$\|\mathbf{X}\|^2 = \sum_{i=1}^{n} X_i^2,$$

for a random vector $\mathbf{X} \in \mathbb{R}^n$. For any subspace V, the projection on V based on the Euclidean norm will be denoted $\mathcal{P}_V\mathbf{X}$. This projection is by definition the unique element $\mathbf{Y} \in V$ such that $\|\mathbf{X} - \mathbf{Y}\|$ is minimised. Moreover, $V^{\perp}$ will denote the subspace consisting of all elements in $\mathbb{R}^n$ orthogonal to all the elements in V.

3 Theory

In this section, estimators of the parameters Σ and α are derived together with a test procedure for linear restrictions of the elements in $\boldsymbol{\gamma}_g$. Many of the details, especially in Section 3.1 and 3.2, are parallel to Section 4 of Kristiansson et al. (2005) and are therefore excluded. Implementations in the statistical language R (R Development Core Team, 2004) for all methods presented here are available from `http://wame.math.chalmers.se`.

3.1 Estimation of the covariance matrix

The estimation of the covariance matrix Σ is complicated for a number of reasons. First there is a scale factor c_g that for each gene g scales Σ uniquely.

To remove this dependence of c_g, a scale-independent method is used. Moreover, estimating a covariance matrix when the mean value is unknown is generally not straight forward since there are trivial solutions that give infinite likelihood (e.g. take the mean value equal to one observation and the corresponding variance equal to zero). To circumvent this problem, no regulated genes between any pair of conditions is assumed, i.e. $\boldsymbol{\mu}_g = \mathbf{0}$ for all g. This assumption is temporary for this section and of course not true in general, but it turns out to be good enough to generate results for data where a clear majority of the genes are not differentially expressed.

The estimator of Σ can now be derived in a similar way to Kristiansson et al. (2005). Fix a gene g and put

$$U_{gi} = \begin{cases} X_{g1} & \text{if } i = 1 \\ X_{gi}/X_{g1} & \text{if } 2 \leq i \leq n. \end{cases}$$

The distribution of $\mathbf{U}_g = (U_{g1}, \ldots, U_{gn})$ can be calculated, and by integration over U_{g1}, the distribution of $(U_{g2}, \ldots, U_{gn})$ is obtained,

$$f_{U_{g2},\ldots,U_{gn}}(u_{g2}, \ldots, u_{gn}) = C|\Sigma|^{-1/2} \left(v_g^{\mathrm{T}} \Sigma^{-1} v_g\right)^{-n/2},$$

where $v_g = (1, u_{g2}, \ldots, u_{gn})$. This is a multivariate Cauchy distribution, which is a special case of the multivariate t-distribution (Tong, 1990). Note that the covariance matrix multiplied by an arbitrary scalar determines the parameters of this distribution uniquely. Numerical maximum likelihood can therefore be used to estimate a positive definite matrix Σ^*, which is Σ scaled by an unknown scalar λ, i.e.,

$$\Sigma^* = \lambda \Sigma. \tag{2}$$

To make Σ^* and λ unique the upper left element in Σ^* is fixed to one.

3.2 Estimation of the shape and scale parameters

In this section, estimators of the scale λ introduced in the previous section and the hyperparameter α are derived. The provisional assumption of $\boldsymbol{\mu}_g = \mathbf{0}$ made in the last section is henceforth dropped. Moreover, the scaled covariance matrix Σ^* is assumed to be known.

Since both λ and α are associated with the variance of the genes, the estimator will be based on the information available in the vectors independent (conditionally on c_g) of the projection of $\mathbf{X}_g$ on V (the maximum likelihood

estimate of μ_g). To simplify understanding, we start by transforming $\mathbf{X}_g$ with the square-root of the scaled covariance matrix, i.e.,

$$\mathbf{X}_g^* = \Sigma^{*-1/2}\mathbf{X}_g,$$

where $\Sigma^{*-1/2}$ is a positive definite matrix such that $\Sigma^{*-1/2}\Sigma^{*-1/2} = \Sigma^{*-1}$ holds. This results in a new model where the variance heterogeneity and correlations are removed, i.e,

$$\mathbf{X}_g^* \mid c_g \sim \mathbf{N}\left(\boldsymbol{\mu}_g^*, \frac{c_g}{\lambda}\mathbf{I}\right),$$

$\boldsymbol{\mu}_g^* = \Sigma^{*-1/2}\boldsymbol{\mu}_g \in V^* = \{\Sigma^{*-1/2}v : v \in V\}$ for all g. Let S_g^* be the square length of the projection of $\mathbf{X}_g^*$ on $V^{*\perp}$, i.e.,

$$S_g^* = \|\mathcal{P}_{V^{*\perp}}\mathbf{X}_g^*\|^2.$$

Conditional on c_g, the distribution of S_g^* will be a scaled chi-squared distribution with $n-p$ degrees of freedom,

$$S_g^* \mid c_g \sim c_g/\lambda \times \chi^2_{n-p}$$

(Theorem 3.12, Arnold (1980)). Using the model assumption that c_g follows an inverse Γ-distribution, the unconditional distribution of S_g^* can be shown to be a scaled β'-distribution with parameters $(n-p)/2$ and α (Johnson et al., 1995, page 248), i.e.,

$$S_g^* \sim 2/\lambda \times \beta'\left((n-p)/2, \alpha\right).$$

From this, the parameters α and λ can be estimated by numerical maximum likelihood and Σ can then be estimated from (2). Due to the large number of genes, the estimates of Σ and α are expected to be precise, so Σ and α are from now on assumed to be known.

3.3 Inference about $\boldsymbol{\gamma}_g$

Statistical tests of linear hypotheses based on $\boldsymbol{\gamma}_g$ will now be derived. For a fixed gene g, such a hypothesis H_0 and the corresponding alternative hypothesis H_A can be written as

$$\begin{aligned} H_0 &: C\boldsymbol{\gamma}_g = \mathbf{0} \\ H_A &: C\boldsymbol{\gamma}_g \neq \mathbf{0}, \end{aligned} \tag{3}$$

where C is a matrix of rank k. We assume that this hypothesis is *testable*, i.e., for each row $\mathbf{c}$ in C there should exist a vector $\mathbf{a} \in \mathbb{R}^n$ such that $\mathbf{a}^{\mathrm{T}} X_g$ is an unbiased estimator of $\mathbf{c}\boldsymbol{\gamma}_g$. In other words, it should be possible to estimate the linear combinations of parameters that will be tested. From testability, it follows that there exists a matrix A with rank k such that

$$C\boldsymbol{\gamma}_g = A\boldsymbol{\mu}_g.$$

Furthermore, let $V_0 \subset V$ be the null space of A, that is the space of all possible $\boldsymbol{\mu_g} \in V$ such that $A\boldsymbol{\mu_g} = \mathbf{0}$. The hypotheses in (3) can then be stated as $H_0 : \boldsymbol{\mu_g} \in V_0$, $H_A : \boldsymbol{\mu_g} \in V \setminus V_0$.

Two likelihood ratio tests will be derived. First a weighted moderated F-test, which under the previously described assumptions will work for any testable hypothesis. Moreover, a weighted moderated t-statistic will be developed for the case when C only has a single row. In this case, the hypothesis concerns one linear combination of the elements of the vector $\boldsymbol{\mu}_g$. This t-test will of course generate p-values equivalent to the F-test but the form of the statistic itself has some advantages, such as a sign to indicate up and down regulation just like the ordinary t-statistic. We start with the more generally applicable F-statistic.

As in the previous section, the model will be transformed to make the theory more straight forward. This time Σ is known, so we let $\tilde{\mathbf{X}}_g = \Sigma^{-1/2}\mathbf{X}_g$. It follows that

$$\tilde{\mathbf{X}}_g \mid c_g \sim \mathbf{N}_n\left(\tilde{\boldsymbol{\mu}}_g, c_g I\right)$$

where $\tilde{\boldsymbol{\mu}}_g = \Sigma^{-1/2}\boldsymbol{\mu_g}$. Moreover, let $\tilde{D} = \Sigma^{-1/2}D$ and define the spaces $\tilde{V}$ and $\tilde{V}_0$ analogous to V and V_0, i.e., let $\tilde{V}$ be the space spanned by the columns in $\tilde{D}$ and $\tilde{V}_0$ be the space with all $\tilde{\boldsymbol{\mu}}_g \in \tilde{V}$ such that $\tilde{A}\tilde{\boldsymbol{\mu}}_g = 0$. As before, $\tilde{A} = A\Sigma^{1/2}$ is a matrix with rank k such that $C\boldsymbol{\gamma}_g = \tilde{A}\tilde{\boldsymbol{\mu}}_g$ holds.

The likelihood ratio test will be derived as in Kristiansson et al. (2005). The likelihood function L of the transformed model is calculated by integration over the prior distribution, resulting in

$$L(\tilde{\boldsymbol{\mu}}_g|\tilde{\mathbf{x}}_g) = \int f_{\tilde{\mathbf{X}}_g \mid c_g}(\tilde{\mathbf{x}}_g) f_{c_g}(c_g)\, dc_g = K\left[\frac{\|\tilde{\mathbf{x}}_g - \tilde{\boldsymbol{\mu}}_g\|^2}{2} + 1\right]^{-n/2-\alpha},$$

where K is a normalisation constant not depending on $\tilde{\mathbf{x}}_g$ or $\tilde{\boldsymbol{\mu}}_g$. The likelihood ratio test can now be formed; we reject $\tilde{\boldsymbol{\mu}} \in \tilde{V}_0$ if

$$\frac{\sup\limits_{\tilde{\boldsymbol{\mu}}_g \in \tilde{V}} L(\tilde{\boldsymbol{\mu}}_g|\tilde{\mathbf{X}}_g)}{\sup\limits_{\tilde{\boldsymbol{\mu}}_g \in \tilde{V}_0} L(\boldsymbol{\mu}_g|\tilde{\mathbf{X}}_g)} > \kappa \tag{4}$$

for a suitable constant κ. The suprema are achieved when $\tilde{\boldsymbol{\mu}}_g$ is the projection of $\tilde{X}_g$ on the spaces $\tilde{V}$ and $\tilde{V}_0$ respectively,

$$\begin{aligned}\arg\max_{\tilde{\boldsymbol{\mu}}_g \in \tilde{V}} L(\tilde{\boldsymbol{\mu}}_g|\tilde{\mathbf{X}}_g) &= \mathcal{P}_{\tilde{V}}\,\tilde{\mathbf{X}}_g \quad \text{and} \\ \arg\max_{\tilde{\boldsymbol{\mu}}_g \in \tilde{V}_0} L(\tilde{\boldsymbol{\mu}}_g|\tilde{\mathbf{X}}_g) &= \mathcal{P}_{\tilde{V}_0}\,\tilde{\mathbf{X}}_g.\end{aligned}$$

Using some algebra and the Pythagorean theorem it is possible to write the likelihood ratio test (4) as

$$\frac{\|\,\mathcal{P}_{\tilde{V}}\,\tilde{\mathbf{X}}_g - \mathcal{P}_{\tilde{V}_0}\,\tilde{\mathbf{X}}_g\|^2}{\|\tilde{\mathbf{X}}_g - \mathcal{P}_{\tilde{V}}\,\tilde{\mathbf{X}}_g\|^2 + 2} = \frac{\|\,\mathcal{P}_{\tilde{V}_0^{\perp}\cap\tilde{V}}\,\tilde{\mathbf{X}}_g\|^2}{\|\,\mathcal{P}_{\tilde{V}^{\perp}}\,\tilde{\mathbf{X}}_g\|^2 + 2} > \kappa', \tag{5}$$

where κ' is a new constant. The subspace $\tilde{V}_0^{\perp} \cap \tilde{V}$ consists of all elements that are in $\tilde{V}$ and are orthogonal (in the metric induced by the Euclidean norm) to the elements in $\tilde{V}_0$, and the subspace $\tilde{V}^{\perp}$ consists of all vectors in $\mathbb{R}^n$ orthogonal to the vectors in $\tilde{V}$.

Under the null hypothesis, the distribution of the statistic (5) can be deduced. Using Theorem 3.11 and Theorem 3.12 in Arnold (1980) it follows that conditionally on c_g, the squared norms of the projections in (5) are independent and χ^2 distributed. The space $\tilde{V}_0^{\perp} \cap \tilde{V}$ has dimension k and $\tilde{V}^{\perp}$ has dimension $n-p$ so, conditionally on c_g,

$$\begin{aligned}\|\,\mathcal{P}_{\tilde{V}_0^{\perp}\cap\tilde{V}}\,\tilde{\mathbf{X}}_g\|^2 &\sim c_g \times \chi^2_k \\ \|\,\mathcal{P}_{\tilde{V}^{\perp}}\,\tilde{\mathbf{X}}_g\|^2 &\sim c_g \times \chi^2_{n-p}.\end{aligned}$$

If we let

$$T = \frac{n-p+2\alpha}{k}\frac{\|\,\mathcal{P}_{\tilde{V}_0^{\perp}\cap\tilde{V}}\,\tilde{\mathbf{X}}_g\|^2}{\|\,\mathcal{P}_{\tilde{V}^{\perp}}\,\tilde{\mathbf{X}}_g\|^2 + 2},$$

and divide both the numerator and the denominator by c_g and use the facts that $2/c_g \sim \chi^2_{2\alpha}$ and that the sum of two independent χ^2 distributed random variables is χ^2 distributed itself, it follows that

$$T \sim F_{k,n-p+2\alpha}.$$

Explicit formulas for $\|\,\mathcal{P}_{\tilde{V}_0^{\perp}\cap\tilde{V}}\,\tilde{\mathbf{X}}_g\|^2$ and $\|\,\mathcal{P}_{\tilde{V}^{\perp}}\,\tilde{\mathbf{X}}_g\|^2$ are straight forward to derive, i.e,

$$\begin{aligned}&\|\,\mathcal{P}_{\tilde{V}_0^{\perp}\cap\tilde{V}}\tilde{\mathbf{X}}_g\|^2 = \\ &\qquad \mathbf{X}_g^{\mathrm{T}}\Sigma^{-1}D(D^{\mathrm{T}}\Sigma^{-1}D)^{-}C^{\mathrm{T}}\left[C(D^{\mathrm{T}}\Sigma^{-1}D)^{-}C^{\mathrm{T}}\right]^{-}C(D^{\mathrm{T}}\Sigma^{-1}D)^{-}D^{\mathrm{T}}\Sigma^{-1}\mathbf{X}_g\end{aligned}$$

and

$$\| \mathcal{P}_{\tilde{V}^\perp} \tilde{\mathbf{X}}_g \|^2 = \mathbf{X}_g^{\mathrm{T}} (\Sigma^{-1} - \Sigma^{-1} D (D^{\mathrm{T}} \Sigma^{-1} D)^- D^{\mathrm{T}} \Sigma^{-1}) \mathbf{X}_g,$$

where for any matrix M, M^- is used to denote the generalised inverse.

When the matrix C has a single row and thus $k = 1$, it is also possible to derive a t-test equivalent to the F-test. Define

$$\bar{X}_g^w = C (D^{\mathrm{T}} \Sigma^{-1} D)^- D^{\mathrm{T}} \Sigma^{-1} \mathbf{X}_g.$$

Since C only has one row, $\bar{X}_g^w$ is a weighted mean value and hence, conditionally on c_g, normally distributed,

$$\bar{X}_g^w \sim \mathbf{N} \left(C \boldsymbol{\gamma}_g, c_g C (D^{\mathrm{T}} \Sigma^{-1} D)^- C^{\mathrm{T}} \right). \tag{6}$$

Define the weighted moderated t-statistic as

$$T' = \sqrt{\frac{n - p + 2\alpha}{C (D^{\mathrm{T}} \Sigma^{-1} D)^- C^{\mathrm{T}}}} \frac{\bar{X}_g^w}{\sqrt{S_g + 2}},$$

where

$$S_g = \| \mathcal{P}_{\tilde{V}^\perp} \tilde{\mathbf{X}}_g \|^2. \tag{7}$$

Under the null hypothesis where $C\boldsymbol{\gamma} = 0$, it is possible to show by using similar arguments as for the F-statistic that

$$T' \sim t_{n-p+2\alpha}.$$

4 Simulations

In this section, a simulated time course experiment is used to evaluate the derived statistics. First, the performance is compared with two other common methods; the moderated F-statistic (Smyth, 2004) and the ordinary F-statistic. Then, effects of correlation between arrays at different time-points are examined.

4.1 Description of the simulated dataset

The simulated experiment consists of three conditions that each is compared to a common reference condition by three replicates. We will think of this setup as a time course with three time points and we call the conditions T1, T2 and T3. An illustration of the design can be seen in Figure 2.

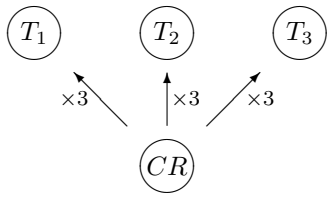


Figure 2: The experimental design used for the simulation studies.

Each time point had 1% of the genes regulated exclusively. Genes regulated at more than one time point were also chosen; time point one and two had 1% genes regulated, time point two and three had 1% regulated genes and finally, time point one, two and three had 1% regulated genes. All groups of regulated genes were mutually exclusive and the genes were selected randomly. Thus, totally 3% of the genes at time point one, 4% of the genes at time point two and 3% of the genes at time point three were chosen to be regulated. The expected values of the regulated genes were sampled independently from a uniform distribution between -2 and 2. For genes regulated at several time points, the expected values were the same for all those time points.

	T1 1	T1 2	T1 3	T2 1	T2 2	T2 3	T3 1	T3 2	T3 3
T1 1	1.00	0.57	0.61	0.14	0.38	0.34	0.00	0.12	0.34
T1 2	*0.40*	2.00	0.73	0.28	0.64	0.54	0.10	0.23	0.56
T1 3	*0.35*	*0.30*	3.00	0.44	0.90	0.73	0.24	0.35	0.78
T2 1	*0.10*	*0.14*	*0.18*	2.00	0.98	0.49	0.24	0.31	0.68
T2 2	*0.22*	*0.26*	*0.30*	*0.40*	3.00	0.52	0.42	0.45	0.93
T2 3	*0.34*	*0.38*	*0.42*	*0.35*	*0.30*	1.00	0.31	0.30	0.59
T3 1	*0.00*	*0.04*	*0.08*	*0.10*	*0.14*	*0.18*	3.00	0.69	0.86
T3 2	*0.12*	*0.16*	*0.20*	*0.22*	*0.26*	*0.30*	*0.40*	1.00	0.42
T3 3	*0.24*	*0.28*	*0.32*	*0.34*	*0.38*	*0.42*	*0.35*	*0.30*	2.00

Table 1: The covariance matrix Σ used for the simulation studies. Variances are underlined and correlations are written in italic below the diagonal.

The total number of genes used was 10000 and the hyperparameter α was fixed to 2. These values are typical for real datasets. The covariance matrix for the experiment was chosen so that there are moderate correlations between the arrays within time points and low to moderate correlations between arrays from different time points (Table 1). Observations for each gene g were then simulated according to the model.

4.2 Comparison with other methods

To investigate if the assumption of a general covariance matrix results in a significantly improved performance, WAME was compared to two other methods; the moderated F-statistic (Smyth, 2004) and the ordinary F-statistic. The moderated F-statistic is based on a linear model with an empirical Bayes approach similar to WAME but variance homogeneity and uncorrelated arrays are assumed. This method is available as an R-package called LIMMA (Smyth et al., 2005) and can be retrieved from the Bioconductor repository (Gentleman et al., 2004). Since this package contains a function to calculate weights used in the estimation of the expected values, LIMMA was used both with and without this feature and will be referred to as weighted and unweighted LIMMA respectively.

First, a test of regulation in time point two was used, i.e. a contrast matrix with a single row was used. All methods were applied to the simulated time course experiment (as described in Section 4.1) and by counting the number of true positives as a function of the number of false positives, Receiver Operating Characteristic (ROC) curves were plotted. The simulation results (see Figure 3) show that WAME clearly performs better than the other methods. Moreover, the performance in LIMMA is only marginally improved when weights are used. The ordinary F-statistic has the worst performance.

Next, the test of no regulation at any time point was performed and the corresponding ROC plots were plotted (Figure 4). As in the previous simulation, WAME performs better than the other methods.

4.3 The effects of correlations between conditions

To investigate the impact of correlations between arrays from different time points, the simulated time course experiment was used once more. This time, three different versions of WAME were compared, each with the scaled covariance matrix Σ^* given instead of estimated. The scale λ and the hyperparameter α were estimated as usual.

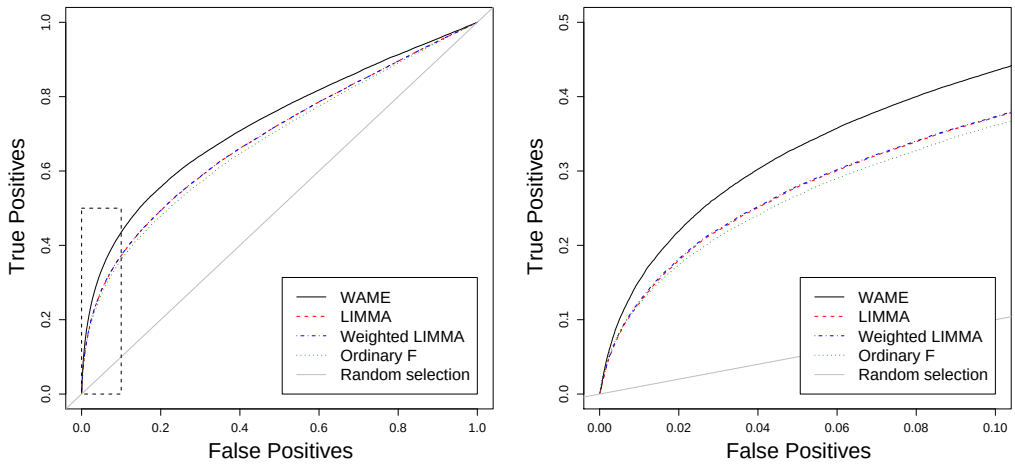


Figure 3: Receiver Operating Characteristic curves for testing regulation at time point two. Four methods are compared on simulated data; WAME, LIMMA with and without weights and the ordinary F-statistic. The figure to the right is a magnification of the dashed box to the left.

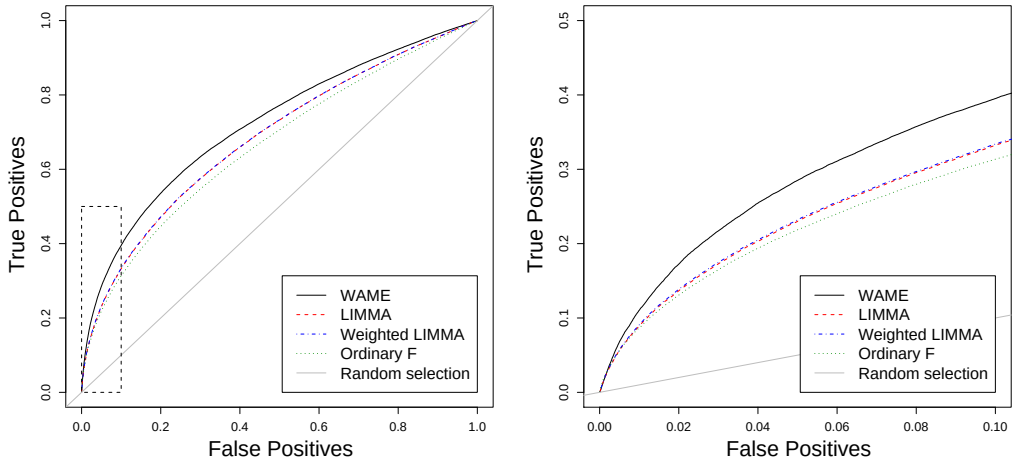


Figure 4: Receiver Operating Characteristic curves for testing regulation at all time points. Four methods are compared on simulated data; WAME, LIMMA with and without weights and the ordinary F-statistic. The figure to the right is a magnification of the dashed box to the left.

The first version used the true matrix shown in Table 1, the second version used the true matrix altered by setting the correlations between groups to zero. Finally, the third version used an identity matrix which results in a model equivalent to the model presented in Smyth (2004) and will thus perform similar to LIMMA. The setup of this simulation is summarised in Table 2.

WAME version 1	True covariance matrix including both different variances for different arrays and correlations between all pairs of arrays.
WAME version 2	True covariance matrix but all correlations between pairs of arrays from different time points set to zero.
WAME version 3	Identity matrix, i.e. same variance for all arrays and no correlations. Equivalent to LIMMA.

Table 2: The experimental setup used in the simulation study to investigate the impact of the correlations. The result can be seen in Figure 5.

The test used is the same as in the first simulation study, i.e. a test for regulation in time point two and the results can be seen in Figure 5. As expected, the version with the true covariance matrix performs best, followed by the version with independence between time points and finally the version which uses an identity matrix. It is interesting to see that the performance loss when correlations between the time points are ignored, is relatively large. This shows that a potential method that focuses on each of the groups independently will be far from optimal.

5 Results from real data

Here, two real datasets are examined. First, the apoAI dataset (Callow et al., 2000) is analysed, where two-colour spotted cDNA microarrays are used to compare eight knockout mice to eight control mice through a common reference. Then the Cardiac dataset (Hall et al., 2004) is investigated. In this experiment heart biopsies was harvested from 19 patients before and after treatment with a ventricular assist device. One-channel oligonucleotide microarrays from Affymetrix Inc. have been used to create this dataset.

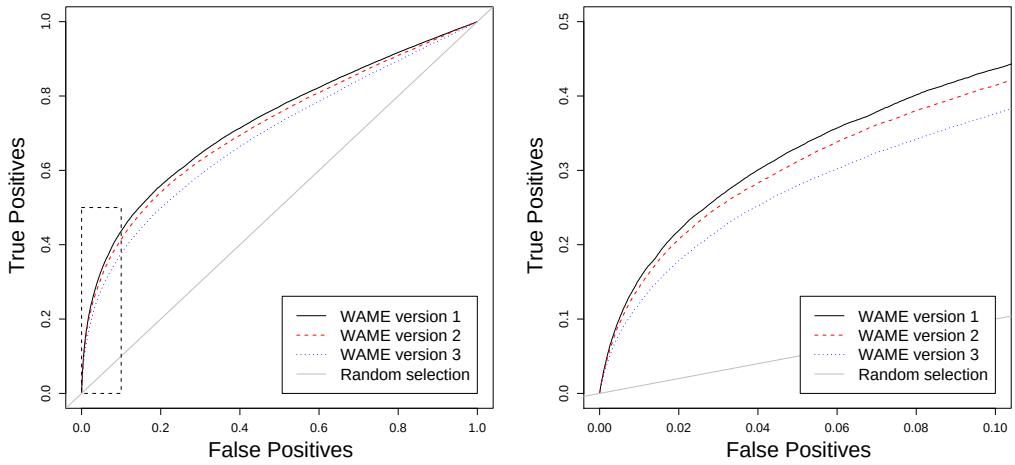


Figure 5: Simulated Receiver Operating Characteristic curves for testing three versions of WAME, all using given covariance matrices. Version 1 uses the correct covariance matrix, version 2 uses the correct covariance matrix but with all correlations between the time points set to zero and finally, version 3 uses an identity matrix.

5.1 The apoAI dataset

The apoAI dataset (Callow et al., 2000) comes from a study of high-density lipoprotein (HDL) metabolism in mice. In the study, mRNA from eight mice with the apolipoprotein AI gene inactivated were compared to mRNA from eight control mice through a common reference, which was created by pooling mRNA from the controls (see Figure 6). The samples from the knockouts and controls were labeled with Cy3 and the samples from the common reference were labeled with Cy5. In total, 16 two-channel cDNA microarrays were hybridised.

The pre-processing of this data is described in Callow et al. (2000) and is summarised here. ScanAlyze was used to analyse the scanned arrays and the background estimation and correction were done in Spot (Buckley, 2000). The resulting files containing background corrected raw intensities are publicly available at

`http://www.stat.berkeley.edu/users/terry/zarray/Html/apodata.html`

Array elements with missing values were removed (totally 158 genes out of 6226) and print-tip loess normalisation (Yang et al., 2002) was used to remove

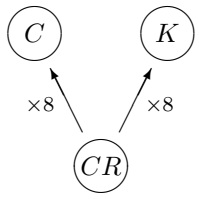


Figure 6: Experimental design for the apoAI dataset. The conditions C and K correspond to the control mice and knockout mice respectively. CR is the common reference which was created by pooling the mRNA from the controls. The mRNA from the control mice and the knockout mice were labeled with Cy5 and the mRNA from the common reference were labelled with Cy3. In total 16 arrays were used in this experiment.

systematic errors.

In the present paper, the linear model developed in Section 3 will be used to analyse this dataset. Three conditions are used to parametrise this experiment; the control (C), the knockout (K) and the common reference (CR). The contrast of interest is the difference between the knockout and control. Note that due to the experimental design, this dataset could not be analysed by the simpler model presented in Kristiansson et al. (2005).

WAME was applied to the normalised values and the estimated covariance matrix (see Table 3) reveals differences in the variance structure between the two groups. In the first group, which consists of the eight control mice, all arrays have fairly similar variances (between 0.16 and 0.10). In the knockout group however, there are several arrays with quite high variances, e.g. array 1 (0.29), array 5 (0.24) and array 6 (0.23).

The estimated covariances are positive for all pairs of arrays. In fact, only a few pair of arrays have a correlation lower than 0.10 and the majority have a correlation above 0.20. These correlations can be verified by examination of the high density parts of the clouds in Figure 7. These moderate correlations can be a result of the common reference design that is used in the experiment. By hybridising mRNA from the same pool on all the arrays, sources of variation will undoubtedly be shared.

The weights that correspond to the contrast comparing the knockouts to the controls are shown in Table 4. As mentioned earlier, the variance in the control group is homogeneous which results in fairly equal weights. In the knock-out group however, the variance and thus the weights differ

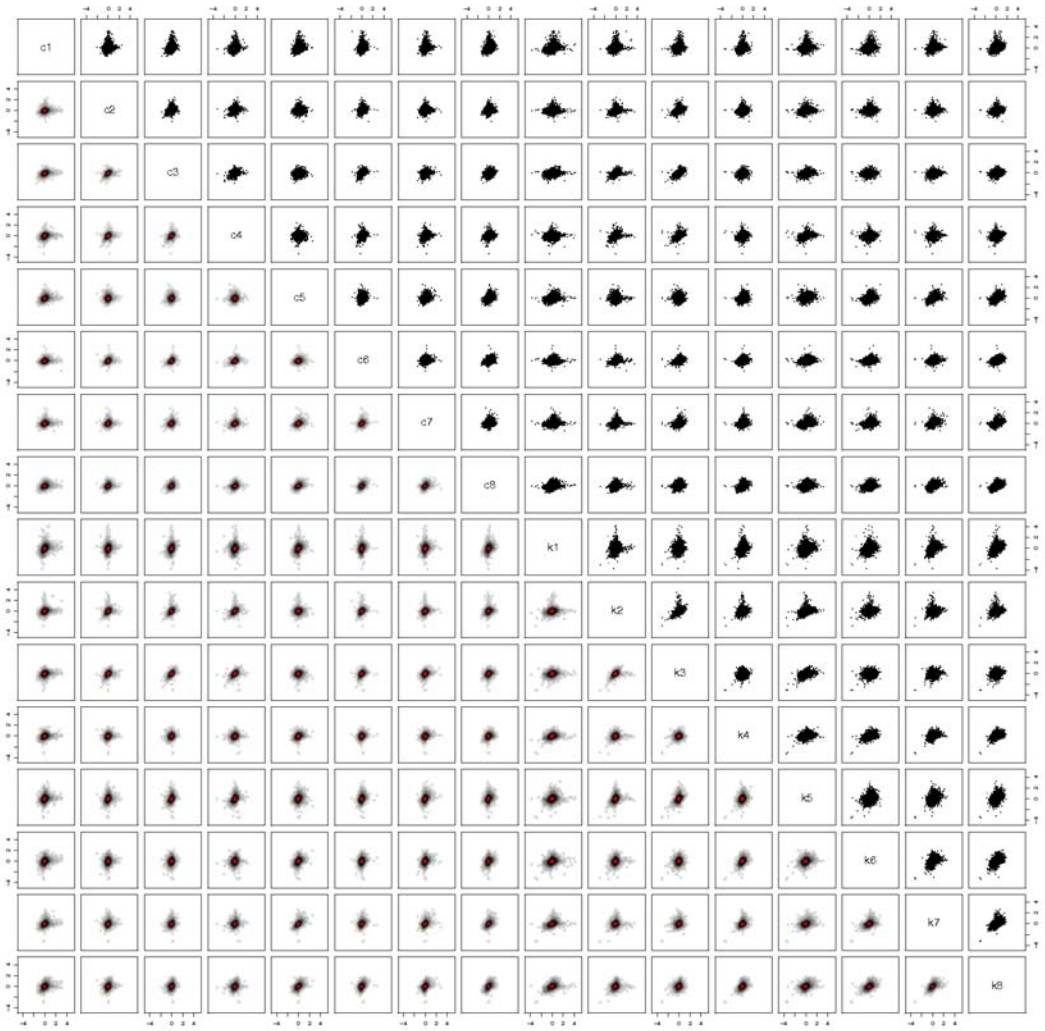


Figure 7: Pair-wise plots of the $\log_2$-ratios for all the 16 arrays in the apoAI dataset. The lower left half shows kernel density estimates of the two-dimensional distribution according to the colour-scale: white, grey, black and red (in increasing level of density). A standalone image in the lossless PNG format can be found at `http://wame.math.chalmers.se`.

	c1	c2	c3	c4	c5	c6	c7	c8	k1	k2	k3	k4	k5	k6	k7	k8
c1	0.16	0.03	0.04	0.04	0.05	0.02	0.03	0.04	0.08	0.03	0.03	0.04	0.05	0.05	0.05	0.05
c2	*0.23*	0.11	0.05	0.06	0.02	0.04	0.03	0.03	0.04	0.06	0.06	0.02	0.05	0.02	0.02	0.03
c3	*0.32*	*0.50*	0.11	0.06	0.02	0.05	0.03	0.03	0.04	0.06	0.08	0.02	0.06	0.02	0.02	0.03
c4	*0.26*	*0.45*	*0.52*	0.14	0.02	0.04	0.04	0.03	0.04	0.07	0.08	0.02	0.07	0.03	0.02	0.04
c5	*0.31*	*0.13*	*0.15*	*0.12*	0.14	0.03	0.05	0.05	0.07	0.01	0.02	0.04	0.04	0.06	0.08	0.07
c6	*0.19*	*0.42*	*0.44*	*0.37*	*0.26*	0.10	0.04	0.04	0.03	0.05	0.06	0.02	0.05	0.03	0.03	0.04
c7	*0.21*	*0.28*	*0.29*	*0.33*	*0.40*	*0.39*	0.11	0.04	0.04	0.04	0.05	0.02	0.06	0.04	0.05	0.05
c8	*0.27*	*0.28*	*0.28*	*0.26*	*0.45*	*0.41*	*0.40*	0.11	0.06	0.03	0.04	0.03	0.05	0.05	0.06	0.07
k1	*0.38*	*0.20*	*0.25*	*0.18*	*0.34*	*0.19*	*0.21*	*0.32*	0.29	0.05	0.04	0.06	0.07	0.08	0.08	0.09
k2	*0.22*	*0.43*	*0.50*	*0.50*	*0.09*	*0.42*	*0.30*	*0.26*	*0.24*	0.15	0.08	0.03	0.07	0.04	0.02	0.04
k3	*0.21*	*0.45*	*0.59*	*0.54*	*0.12*	*0.47*	*0.36*	*0.29*	*0.21*	*0.56*	0.15	0.02	0.08	0.03	0.02	0.03
k4	*0.27*	*0.18*	*0.19*	*0.17*	*0.25*	*0.12*	*0.16*	*0.23*	*0.29*	*0.20*	*0.13*	0.15	0.05	0.08	0.04	0.05
k5	*0.24*	*0.29*	*0.38*	*0.38*	*0.22*	*0.30*	*0.34*	*0.28*	*0.27*	*0.36*	*0.40*	*0.28*	0.24	0.07	0.05	0.06
k6	*0.24*	*0.14*	*0.15*	*0.19*	*0.35*	*0.22*	*0.25*	*0.30*	*0.30*	*0.21*	*0.13*	*0.43*	*0.28*	0.23	0.07	0.08
k7	*0.32*	*0.18*	*0.20*	*0.15*	*0.60*	*0.28*	*0.37*	*0.48*	*0.41*	*0.14*	*0.17*	*0.29*	*0.28*	*0.38*	0.14	0.08
k8	*0.34*	*0.25*	*0.21*	*0.24*	*0.49*	*0.35*	*0.37*	*0.51*	*0.44*	*0.25*	*0.21*	*0.32*	*0.29*	*0.43*	*0.55*	0.16

Table 3: Estimated covariance matrix for the apoAI dataset. The variances are underlined and correlations are written in italic below the diagonal.

substantially. Here, the arrays 1, 5 and 6, which all have a large variance, get heavily down-weighted.

Array	c1	c2	c3	c4	c5	c6	c7	c8	Sum
Weights	-0.08	-0.09	-0.16	-0.13	-0.15	-0.13	-0.11	-0.16	-1
Array	k1	k2	k3	k4	k5	k6	k7	k8	Sum
Weights	0.01	0.17	0.25	0.10	0.05	0.03	0.23	0.15	1

Table 4: The weights used in testing difference in gene expression between the knockouts and the controls in the apoAI dataset.

For all array elements p-values were calculated for three different methods; the weighted moderated t-statistic (WAME), the moderated t-statistic (LIMMA) and the ordinary t-statistic. The 15 most extreme elements according to WAME can be seen in Table 5. The first entry in this list corresponds to the removed apoAI gene and it is found to be heavily down-regulated as one would expect. The next seven entries correspond to three other genes, which all have been verified to be differentially expressed (Callow et al., 2000).

Two quantile-quantile plots are shown in Figure 8. To the left, the observed t-values from WAME are plotted against the theoretical values under the null hypothesis. The eight verified elements are marked with crosses and they clearly stand out compared to the other elements. To the right, a blow up of the dashed box is shown. In this figure, quantile-quantile curves for the weighted moderated t-statistic (WAME) and the moderated t-statistic (LIMMA) are plotted. WAME seems to follow the diagonal line well while

LIMMA deviate slightly towards higher observed absolute t-values. A similar deviation can also be seen in the quantile-quantile curve of the ordinary t-statistic (results not shown).

Name	Average $\log_2$-ratio	Ord. t-test p-value	LIMMA p-value	WAME p-value
Apo AI, lipid-Img	-3.18	1.51×10^{-12}	4.98×10^{-15}	6.11×10^{-15}
Est, highly similar to apolipoprotein A-I precursor, lipid-UG	-2.96	1.21×10^{-8}	1.63×10^{-10}	2.53×10^{-10}
Catechol O-Methylthansferase, membrane-bound, brain-Img	-1.76	1.21×10^{-8}	3.14×10^{-10}	1.95×10^{-9}
Est, Weakly similar to C-5 Sterol Desaturase, lipid-UG	-0.96	3.39×10^{-9}	7.52×10^{-10}	5.24×10^{-9}
Est, Highly similar to Apolipoprotein C-III precursor, lipid-UG	-1.01	3.30×10^{-7}	4.56×10^{-8}	2.47×10^{-8}
Apo CIII, lipid-Img	-0.90	5.53×10^{-8}	1.22×10^{-8}	2.99×10^{-8}
Est	-0.92	3.01×10^{-7}	4.65×10^{-8}	6.27×10^{-8}
Similar to yeast sterol desaturase, lipid-Img	-0.94	4.50×10^{-6}	7.09×10^{-7}	3.79×10^{-7}
Similar to Hypothetical protein 1 - fruit fly	-0.57	4.62×10^{-3}	2.63×10^{-3}	2.52×10^{-4}
Fatty acid-binding protein, epidermal, lipid-UG	-0.48	5.66×10^{-4}	2.49×10^{-4}	3.06×10^{-4}
BLANK	0.44	5.65×10^{-3}	4.65×10^{-3}	4.13×10^{-4}
estrogen rec	0.42	1.43×10^{-3}	1.03×10^{-3}	7.10×10^{-4}
Cy5RT	0.71	4.14×10^{-3}	1.68×10^{-3}	1.03×10^{-3}
Tbx6	-0.33	5.62×10^{-3}	7.93×10^{-3}	1.18×10^{-3}
Est	0.47	1.30×10^{-2}	9.63×10^{-3}	1.32×10^{-3}

Table 5: The 15 genes in the ApoAI dataset with smallest p-values according to WAME. The table also shows the corresponding fold-changes and p-values for the ordinary and moderated t-statistics.

To highlight the impact of the generalised linear model used in WAME, the empirical distributions of the weighted mean (WAME) and the ordinary (i.e. unweighted) mean (LIMMA) were compared. Under the model assumptions, the weighted mean should have a smaller variance and thus the distribution should have a smaller spread. This is in fact observed when the empirical distributions are estimated by a kernel density estimator (see Figure 9). This suggests that the estimator of the fold-changes in WAME has higher precision than the corresponding estimator in LIMMA.

5.2 The Cardiac dataset

In Hall et al. (2004) heart biopsies from 19 patients with heart failure were harvested to investigate differences in gene expression before and after treatment with a ventricular assist device. The role of the study was to identify genes involved in vascular signaling networks. The patients were divided into three groups; the ischemic group (I) with 5 patients that had evidence of coronary artery disease, the acute myocardial infarction group (IM) with 6 patients that had an acute myocardial infarction within 10 days of the implant and finally the nonischemic group (N) where the 8 patients did not show any evidence for coronary artery disease. Each mRNA sample was prepared and hybridised to one Affymetrix one-channel oligonucleotide array

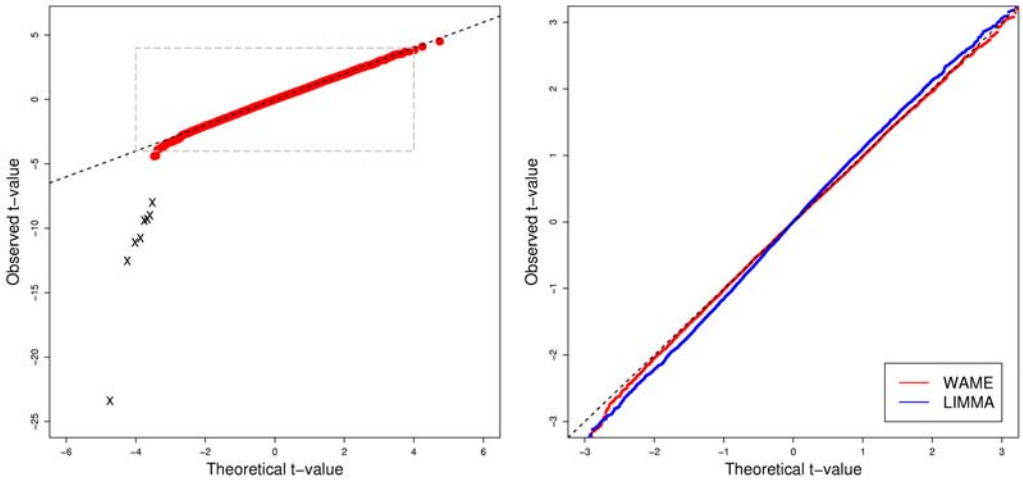


Figure 8: To the left a quantile-quantile plot comparing the estimated weighted moderated t-statistic (WAME) to its corresponding theoretical null distribution. The eight elements marked with a cross corresponds to the four genes that were verified to be differentially expressed. To the right is a blow up of the dashed box. This plot also contains the quantile-quantile curve for the moderated t-statistic (LIMMA).

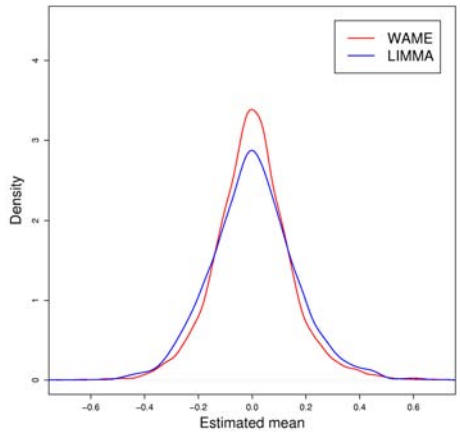


Figure 9: Kernel density estimates of the distributions of the mean values from WAME and LIMMA. The spread of the former is smaller which suggests that the estimator in WAME is more precise.

(HG-U133A) resulting in two arrays for each patient, i.e. 38 arrays in total. The resulting raw data was made publically available by the authors and can be found at the Gene Expression Omnibus repository (Edgar et al., 2002).

The .CEL-files for all 38 arrays were retrieved and then pre-processed and normalised by RMA (Irizarry et al., 2003). The paired observations were used to form $\log_2$-ratios according to the experimental design shown in Figure 10. The ischemic group (I) of this dataset was analysed in Kristiansson et al. (2005) where patients with different variances and substantial correlations were identified.

In the present paper the ischemic group will be analysed once more, but this time using the more general model. In this framework, all patients will be incorporated, even if we only test for differential expression for patients in a single group. This is a major difference to the model in Kristiansson et al. (2005) which can only take advantage of the arrays from one group of interest, i.e. it can only analyse direct comparisons with two conditions. In this section, the results from the new analysis are presented and compared to the corresponding old results.

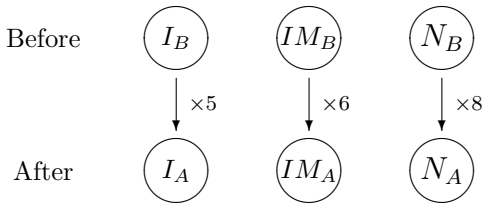


Figure 10: Experimental design for the cardiac dataset. All three groups have a condition before the treatment (I_B, IM_B and N_B) and a condition after the treatment (I_A, IM_A and N_A). The dataset consists of 19 observations in total.

The estimated covariance matrix for all 19 pairs of arrays is shown in Table 6. The variances in this dataset differ considerable, both between and within the three groups. For example, the IM group has three arrays with high variance and three arrays with rather low variance, while the variances in I and N group are more homogeneous.

The correlations in this dataset are in general small, but there are exceptions. The first patient in the IM group (IM1) has a high positive correlation with IM4 (0.54) but also negative correlations with I7 (-0.64) and IM7 (-0.55). These correlations can also be seen in Figure 11. A closer

	I12	I13	I4	I7	I8	IM1	IM3	IM4	IM5	IM6	IM7	N2	N22	N3	N4	N6	N7	N8	N9
I12	0.04	0.00	0.00	0.01	0.00	-0.01	0.01	-0.00	-0.01	0.00	0.01	-0.01	0.00	0.01	0.00	0.01	-0.00	0.00	0.01
I13	*0.03*	0.17	-0.01	0.01	-0.00	0.10	0.11	0.08	0.07	0.02	-0.05	0.01	-0.06	-0.01	0.03	-0.02	-0.06	-0.05	0.01
I4	*0.04*	*-0.13*	0.06	0.01	0.00	-0.03	0.00	-0.01	-0.01	0.00	0.02	-0.01	0.01	0.00	0.01	0.02	-0.01	0.02	-0.00
I7	*0.11*	*0.07*	*0.09*	0.23	-0.02	-0.26	0.07	-0.04	0.03	-0.01	0.04	-0.06	0.01	-0.00	0.02	0.05	-0.03	0.05	0.01
I8	*0.03*	*-0.02*	*0.04*	*-0.16*	0.04	0.04	-0.02	0.01	-0.00	0.00	-0.00	0.01	0.00	0.01	-0.00	0.00	0.01	-0.01	0.00
IM1	*-0.04*	*0.29*	*-0.15*	*-0.64*	*0.22*	0.73	-0.06	0.18	0.10	0.06	-0.14	0.14	-0.07	0.02	-0.02	-0.10	-0.04	-0.18	0.00
IM3	*0.11*	*0.41*	*0.02*	*0.22*	*-0.12*	*-0.11*	0.43	0.04	-0.04	0.02	-0.02	-0.07	-0.04	-0.01	0.06	-0.01	-0.06	-0.01	0.03
IM4	*-0.01*	*0.50*	*-0.06*	*-0.24*	*0.12*	*0.54*	*0.17*	0.14	0.07	0.02	-0.04	0.02	-0.04	-0.01	0.00	-0.00	-0.04	-0.05	0.01
IM5	*-0.09*	*0.29*	*-0.05*	*0.09*	*-0.00*	*0.19*	*-0.09*	*0.32*	0.35	0.01	-0.03	0.05	-0.03	-0.01	-0.00	0.03	-0.03	-0.05	-0.00
IM6	*0.10*	*0.20*	*0.02*	*-0.13*	*0.08*	*0.29*	*0.15*	*0.22*	*0.06*	0.06	-0.01	0.00	-0.01	0.01	0.00	-0.01	-0.01	-0.02	0.00
IM7	*0.09*	*-0.41*	*0.27*	*0.31*	*-0.06*	*-0.55*	*-0.09*	*-0.39*	*-0.20*	*-0.20*	0.08	-0.03	0.03	0.01	0.00	0.04	0.03	0.05	-0.00
N2	*-0.12*	*0.06*	*-0.08*	*-0.29*	*0.10*	*0.40*	*-0.25*	*0.15*	*0.22*	*0.05*	*-0.28*	0.17	-0.01	0.00	-0.02	-0.02	-0.01	-0.05	-0.01
N22	*0.04*	*-0.49*	*0.12*	*0.04*	*0.02*	*-0.28*	*-0.20*	*-0.36*	*-0.19*	*-0.08*	*0.38*	*-0.10*	0.09	0.01	-0.01	0.01	0.03	0.02	-0.00
N3	*0.11*	*-0.10*	*0.04*	*-0.01*	*0.14*	*0.07*	*-0.04*	*-0.06*	*-0.05*	*0.10*	*0.09*	*0.03*	*0.20*	0.06	-0.00	0.00	0.01	-0.00	0.01
N4	*0.03*	*0.20*	*0.08*	*0.12*	*-0.03*	*-0.08*	*0.28*	*0.04*	*-0.01*	*0.04*	*0.05*	*-0.11*	*-0.08*	*-0.02*	0.10	0.01	-0.01	0.01	0.00
N6	*0.09*	*-0.14*	*0.21*	*0.30*	*0.03*	*-0.35*	*-0.03*	*-0.04*	*0.14*	*-0.13*	*0.36*	*-0.15*	*0.13*	*0.03*	*0.10*	0.11	0.02	0.05	0.01
N7	*-0.06*	*-0.41*	*-0.07*	*-0.18*	*0.07*	*-0.15*	*-0.27*	*-0.27*	*-0.15*	*-0.14*	*0.28*	*-0.04*	*0.31*	*0.14*	*-0.05*	*0.16*	0.12	0.03	-0.01
N8	*0.01*	*-0.28*	*0.20*	*0.21*	*-0.06*	*-0.46*	*-0.03*	*-0.27*	*-0.17*	*-0.17*	*0.41*	*-0.28*	*0.16*	*-0.04*	*0.04*	*0.36*	*0.21*	0.20	-0.01
N9	*0.26*	*0.13*	*-0.01*	*0.09*	*0.06*	*0.02*	*0.22*	*0.10*	*-0.01*	*0.00*	*-0.01*	*-0.08*	*-0.04*	*0.09*	*0.06*	*0.10*	*-0.12*	*-0.05*	0.05

Table 6: Estimated covariance matrix for the cardiac dataset. The variances are underlined and correlations are written in italic below the diagonal.

examination of the $\log_2$-ratios revealed that both IM1 and I7 have skewed distributions, IM1 towards positive values and I7 towards negative values. This may be a result from either the experiment itself or from pre-processing steps and can explain the unexpected negative correlations. However, we will still keep these arrays in the further analysis, since their high variance will result in a low weight and thus a low impact on the final result. The hyperparameter α was estimated to 1.73 which means that the resulting t-distribution will gain approximately 3.5 degrees of freedom.

The weights used in the test for differential expression in group I can be seen in Table 7. Not surprisingly, the patient I7 gets the lowest weight, while I12 and I8, which both have low variance and small correlations, get the highest weights. Note that these weights sum to 1.

It is also interesting to examine the weights in the IM and N group. Due to correlations between patients from different groups, these weights will be non-zero but sum to zero. Due to the high variances in the IM group, the corresponding weights are relatively low. In the N group, N7 and N6 have relatively high weights (0.10 and -0.06), which stems from that the fact that they both have a low variance and correlates with group I (N7 mostly negative and N6 mostly positive).

To test for differential expression in group I, the weighted moderated t-statistic was calculated. Figure 12 show the resulting quantile-quantile plot where the theoretical t-distribution has approximately 19.5 degrees of freedom. Most genes follow the diagonal line well which suggest relatively few regulated genes and a good model fit. A few genes have a larger absolute t-value than can be explained by the null distribution which points towards

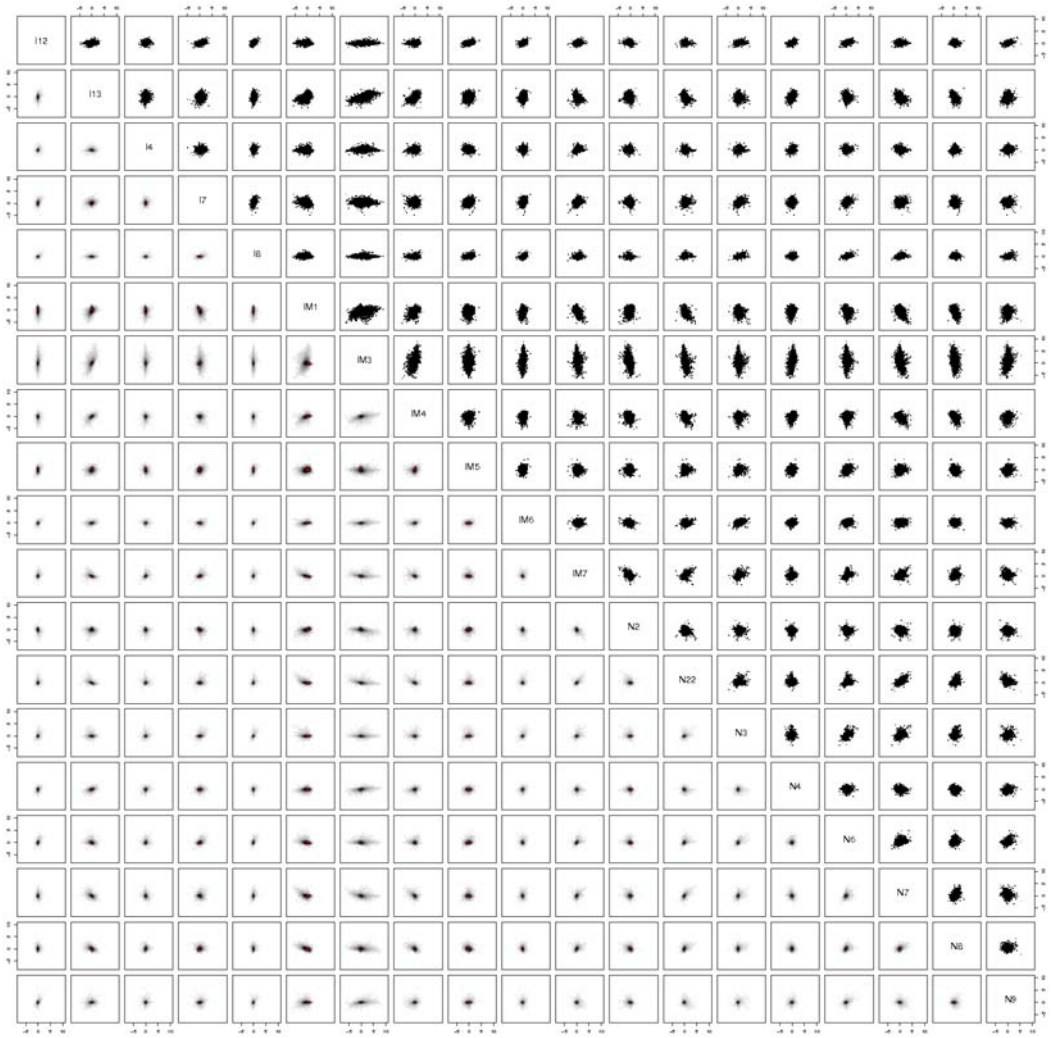


Figure 11: Pair-wise plots of the $\log_2$-ratios for all the 19 patients in the cardiac dataset. The lower left half shows kernel density estimates of the two-dimensional distribution according to the colour-scale: white, grey, black and red (in increasing level of density). A standalone image in the lossless PNG format can be found at `http://wame.math.chalmers.se`.

Patient	I12	I13	I4	I7	I8				Sum
Weight	0.28	0.13	0.21	0.10	0.28				1
Patient	IM1	IM3	IM4	IM5	IM6	IM7			Sum
Weight	0.03	-0.01	-0.02	-0.00	-0.01	0.01			0
Patient	N2	N22	N3	N4	N6	N7	N8	N9	Sum
Weight	0.03	0.05	-0.05	-0.02	-0.06	0.10	0.01	-0.05	0

Table 7: Weights for the patients in the cardiac dataset when differential expression in the ischemic (I) group is studied.

some of them being regulated. Note that the gene expression in this experiment is asymmetric with more genes seemingly up-regulated than seemingly down-regulated.

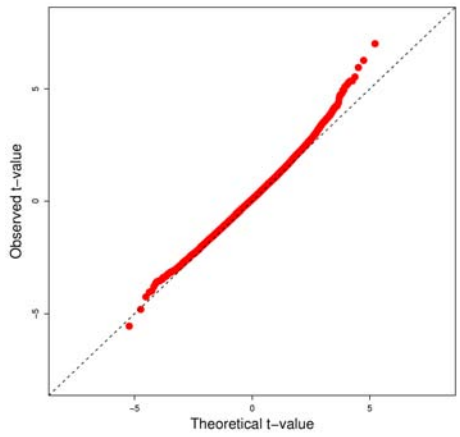


Figure 12: This figure shows the quantile-quantile plots of the weighted moderated t-statistic for the test of differential expression in group I.

When the results are compared to the corresponding analysis restricted to the arrays from group I (the analysis in Kristiansson et al. (2005)) there are both similarities and differences. The estimated scaled covariance matrix for the 5 patients in group I are similar and both the scaled variances and the correlations above 0.10 change less than 10%. The hyperparameter α is estimated to 1.92 when only the patients from group I are used instead of 1.73 when all 19 patients are used. This results in a slightly different scaling factor λ (0.046 instead of 0.042) and thus a different covariance matrix.

Moreover, there are interesting differences between the weights from the two models. Under the general model (Table 7), the weights are more conservative than under the restricted model (Table 8). For example, the weight for patient I7 increases from 0.05 to 0.10.

Patient	I12	I13	I4	I7	I8	Sum
Weight	0.30	0.09	0.23	0.05	0.33	1

Table 8: The weights from the I group calculated by the old model in Kristiansson et al. (2005). These weights are less conservative than the weights in Table 7.

Ranking lists sorted by the p-values were produced for both models. Among the top 100 and 500 genes, 38% and 52% respectively, appear on both lists. To compare the deviating genes, quantile-quantile plots of the p-values on a logarithmic scale (base 10) were made (Figure 13). In the plot to the left (general model), fewer genes deviate from the diagonal line than in the plot to the right (restricted model). Since the number of regulated genes is unknown it is hard to say which plot that is more correct but a better overall fit makes the extreme genes more distinct.

6 Discussion

The experimental design of microarray gene expression assays are often complex and several conditions are usually involved. The arrays in these experiments are produced through a series of steps, each inducing differences in precision and systematic effects. This generates data with varying quality, which is desirable to take into account.

In this paper, a generalisation of the paired two-sample analysis method introduced in Kristiansson et al. (2005) is described. The new method, which as its predecessor, is referred to as WAME, is based on a linear model capable of analysing paired microarray experiments with any number of conditions. The observations are assumed to measure the differences in mRNA levels on a logarithmic scale (log-ratios) between pairs of conditions. This means that the method will be applicable to most experiments using two-channel cDNA microarrays and many experiments with one-channel oligonucleotide microarrays from Affymetrix.

For each gene, the vector of all the pair-wise log-ratios are assumed to follow a multidimensional normal distribution. A covariance matrix Σ is used

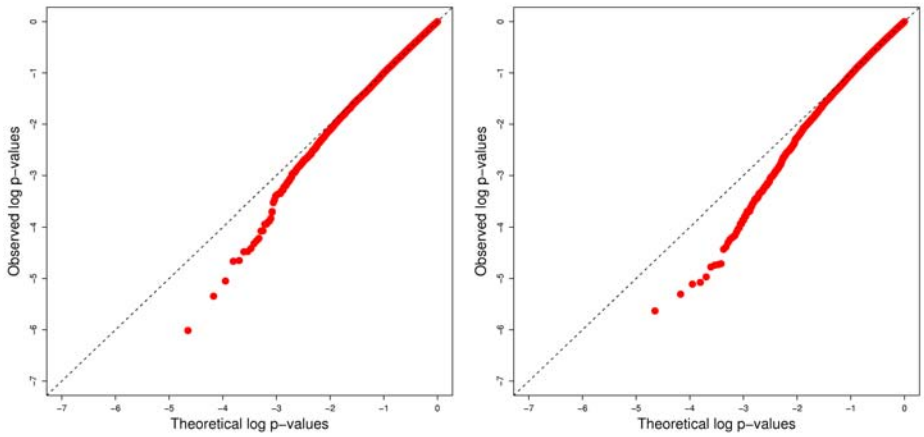


Figure 13: This figure shows two quantile-quantile plots of the $\log_{10}$ p-values for the test of differential expression in group I. The plot to the left is based on the general paired model using all 19 patients and the plot to the right is based on the restricted model from Kristiansson et al. (2005) using only the patients in the I group.

to catch the differences in quality of the different pairs, such as correlations and unequal precision. Gene-specific variances are modelled through a factor c_g, scaling the covariance matrix uniquely for each gene. Since microarray experiments often consist of few gene-wise repetitions, c_g is modelled by an inverse gamma distribution random variable with shape parameter α and fixed scale parameter β. This is analogous to Lönnstedt and Speed (2002), Smyth (2004) and Kristiansson et al. (2005).

To estimate the covariance matrix Σ, an assumption that most genes are not differentially expressed is made. Then, after c_g is removed by a transformation, Σ scaled with an unknown scale λ can be estimated by numerical maximum likelihood. Point estimators for λ and α are also derived based on the residual sum of squares. All these steps parallel Kristiansson et al. (2005).

For any testable linear hypotheses, a likelihood ratio test is derived, resulting in a weighted moderated F-statistic. In the special case of a one-dimensional null hypothesis restriction, a weighted moderated t-statistic is formed. In both cases, correlations give rise to array-specific weights, that can be non-zero for parts of the data that would not be included under the assumption of independent arrays. The weighted moderated statistics can be

seen as generalisations of the moderated F- and t-statistics found in Smyth (2004).

The model was evaluated on a simulated time course experiment with three time points. The improvement in performance, compared to LIMMA (Smyth et al., 2005) was shown to be substantial. Moreover, when the time points were (wrongly) assumed to be independent, the effect on the performance was shown to be relatively large.

Two real datasets were analysed. The first was the apoAI dataset (Callow et al., 2000) which consists of eight knockout mice that are compared to eight controls through a common reference. The estimated covariance matrix contains moderate positive correlations for almost all arrays. This is probably a result of the common reference design, where sources of variation undoubtedly are shared between the arrays. Quantile-quantile plots of the t-statistic revealed that WAME fitted the diagonal line well, while LIMMA tended to over-estimate the t-values. A similar effect was observed for the model in (Kristiansson et al., 2005). Moreover, the precision for the estimated fold-changes was shown to increase when variances and covariances were taken into account.

The other dataset investigated was the cardiac dataset (Hall et al., 2004) which contains paired measurements for 19 patients, divided into three groups (I, IM and N) based on their medical condition. One-channel oligonucleotide microarrays from Affymetrix were used to produce the data. The covariance matrix revealed differences between the groups. The variances were homogeneous in both the I and N group in contrast to the IM group where both high and low variances were found. The IM group also contained several high correlations, both to patients within the group and to patients in the other groups.

Differential expression in the first group was investigated and the results compared to Kristiansson et al. (2005). The weights with the general model suggested in this paper resulted in more conservative weights and quantile-quantile plots for the p-values showed that fewer genes deviated from the diagonal.

When using WAME, it is important to keep in mind that the model is far from perfect. The noise structure may be different for different genes and the assumption of normality may not be valid. The gene-specific variance might also be different for different groups of conditions, leading to erroneous variance estimates. The effect of a potential dependence between expression level and variance is unknown.

In principle we may adapt the weighted moderated F- or t-tests to test null hypotheses that the linear combinations in question are equal to arbitrary

constants. Confidence intervals and ellipsoids can then be constructed based on the correspondence theorem between hypothesis testing and confidence sets (Casella and Berger, 2002, Theorem 9.2.2). However, since the assumed variance structure is hard to validate for regulated genes, we have refrained from developing this topic at this stage.

It should also be noted that the restriction to models with pairwise observations in Section 2.1 may be relaxed. A crucial step, preceding the hypothesis testing in Section 3.3, is the estimation of the covariance matrix in Section 3.1 where we assume that the expected value of the observation vector $\mathbf{X}_g$ has approximately mean zero for a majority of genes. This seems reasonable if $\mathbf{X}_g$ consists of pair-wise differences as assumed in Section 2.1 leading to a design matrix with row sums zero, but the assumption may also be satisfied in other designs such as in some regression models.

A further generalisation of WAME is currently being developed, extending the procedure from paired to general microarray experiments. A linear transformation can there be used to first remove the information explainable by the null hypothesis. Assuming that the null hypothesis is true for most genes, the transformed covariance structure can then be estimated and the weighted statistics formed from the transformed data similar to the methods in the current paper. Work is underway to derive the properties of this procedure and to verify its usefulness on real data. An R-package providing easy access to the WAME procedure is also under development.

References

S.F. Arnold. *The Theory of Linear Models and Multivariate Analysis*. John Wiley & Sons, 1980.

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. Buckley. Spots User's Guide. *CSIRO Mathematical and Information Sciences, Sydney, Australia. http://www.cmis.csiro.au/iap/Spot/spotmanual.htm*, 2000.

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 Research*, 10(12):2022–2029, 2000.

G. Casella and R.L. Berger. *Statistical Inference*. Duxbury, 2002.

G.A. Churchill. Fundamentals of experimental design for cdna microarrays. *Nature Genetics Supplement*, 32:490–495, 2002.

R. Edgar, M. Domrachev, and A.E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, L. Friedrich, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y.H. Yang, and J. Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

J.L. Hall, S. Grindle, X. Han, D. Fermin, S. Park, Y. Chen, R.J. Bache, A. Mariash, Z. Guan, S. Ormaza, J. Thompson, J. Graziano, S.E. de Sam Lazaro, S. Pan, R.D. Simari, and L.W. Miller. Genomic profiling of the human heart before and after mechanical support with a ventricular assist device reveals alterations in vascular signaling networks. *Journal of Physiological Genomics*, 17(3):283–291, 2004. URL `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GDS558`.

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

K. Johnson and S. Lin. QA/QC as a pressing need for microarray analysis: meeting report from CAMDA'02. *BioTechniques*, 34(suppl):S62–S63, 3 2003.

N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions Volume 2*. Wiley, 1995.

E. Kristiansson, A. Sjögren, M. Rudemo, and O. Nerman. Weighted analysis of paired microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 30, 2005.

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

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL `http://www.R-project.org`.

G.K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

G.K. Smyth, N.P. Thorne, and J. Wettenhall. *LIMMA: Linear Models for Microarray Data User's Guide*, 2005. URL `http://www.bioconductor.org`.

J.P. Steibel and J.M. Rosa. On reference designs for microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

Y.L. Tong. *The Multivariate Normal Distribution*. Springer, 1990.

Y.H. Yang, S. Dudoit, P. Luu, D. Lin, V. Peng, J. Ngai, and T.P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.

Paper III

Weighted analysis of general microarray experiments

Anders Sjögren*[1,2], Erik Kristiansson[1,2] , Mats Rudemo[1,2] , Olle Nerman[1,2]

[1]Mathematical Statistics, Chalmers University of Technology, 412 96 Göteborg, Sweden
[2]Mathematical Statistics, Göteborg University, 412 96 Göteborg, Sweden

Email: Anders Sjögren*- anders.sjogren@math.chalmers.se; Erik Kristiansson - erikkr@math.chalmers.se; Mats Rudemo - rudemo@math.chalmers.se; Olle Nerman - nerman@math.chalmers.se;

*Corresponding author

Abstract

Background: In DNA microarray experiments, measurements from different biological samples are often assumed to be independent and to have identical variance. For many datasets these assumptions have been shown to be invalid and typically lead to too optimistic p-values. A method called WAME has been proposed where a variance is estimated for each sample and a covariance is estimated for each pair of samples. The current version of WAME is, however, limited to experiments with paired design, e.g two-channel microarrays.

Results: The WAME procedure is extended to general microarray experiments, making it capable of handling both one- and two-channel datasets. Two public one-channel datasets are analysed and WAME detects both unequal variances and correlations. WAME is compared to other common methods: fold-change ranking, ordinary linear model with t-tests, LIMMA and weighted LIMMA. The p-value distributions are shown to differ greatly between the examined methods. In a resampling-based simulation study with few regulated genes, the p-values generated by WAME are found to be substantially more correct than the alternatives. WAME is also shown to have higher power than the other methods. WAME is available as an R-package.

Conclusions: The WAME procedure is generalized and the limitation to paired-design microarray datasets is removed. WAME is shown to have higher power and more accurate p-values when few genes are regulated compared to four other common methods.

Background

Introduction

The DNA microarray technique involves a series of steps, from the harvesting of cells or biopsies to the preprocessing of the scanned arrays, before analysable data are obtained. During several of these steps the quality can be affected by random factors. For instance, depending on the handling of a biological sample the mRNA can be more or less degraded [1], and the cell-type composition of a biopsy can be more or less representative for the tissue in question. When arrays share sources of variation the deviations from the nominal value will be correlated. For example, two arrays from sources with degraded RNA will both tend to underestimate the expression of easily degradable genes, and two biopsies with a similar and non-representative cell-type composition will deviate in a similar fashion from the average expression for the ideal cell-type composition.

The procedure *Weighted Analysis of Microarray Experiments* (WAME) [2, 3] introduced a model where a covariance-structure matrix common for all genes aims at catching differences in quality by differences in variances and covarying deviations by correlations between arrays. For computations of test statistics and estimators this resulted in weighting of observations according to the estimated covariance-structure matrix, giving lower weight to imprecise or positively correlated arrays.

In order for the estimation of the covariance matrix to work in the current WAME method, the measurements of most genes must only measure noise, i.e. have an expected value of zero. This is the case in experiments where pair-wise log-ratios are observed and where few genes are differentially expressed between any of the pairwise measured conditions. In the present paper, this crucial constraint will be relaxed to only require that most genes are non-differentially expressed between the conditions actually being compared. Thus, non-paired experiments can be analysed, e.g. many additional ones based on one-channel microarray data. The relaxation is realised by transforming the data to remove irrelevant information in a manner yielding transformed data with expectation zero for non-differentially expressed genes, after which the current WAME method is applied. The transformed data are shown to give equivalent tests and estimates to those of the original data, given the corresponding covariance-structure matrices.

Problem formulation and current methods

Given a microarray experiment with n arrays and m genes, we observe for each gene g an n-dimensional vector $\mathbf{X}_g$ of $\log_2$ transformed values measuring mRNA abundance. In WAME the vector $\mathbf{X}_g$ is assumed to have expectation $\boldsymbol{\mu}_g$ described by a design matrix D and a gene-specific parameter vector $\boldsymbol{\gamma}_g$, typically having one dimension per studied condition. A covariance-structure matrix Σ, common for all genes, is used to model differences in quality between arrays as different variances and shared sources of variation between arrays as correlations. A gene-specific variance-scaling factor c_g is assumed to have inverse gamma prior distribution with a global shape parameter α. Conditional on c_g the vector $\mathbf{X}_g$ is assumed to have a normal distribution with covariance matrix $c_g\Sigma$. A matrix C specifies the differential expression vector $\boldsymbol{\delta}_g$, describing the linear combinations of the parameters that are of main interest. Formally,

$$\begin{aligned} \boldsymbol{\mu}_g &= D\,\boldsymbol{\gamma}_g \ , \\ \mathbf{X}_g \mid c_g &\sim \mathrm{N}(\boldsymbol{\mu}_g, c_g\Sigma) \ , \\ c_g &\sim \Gamma^{-1}(\alpha, 1) \ , \end{aligned} \tag{1}$$

and variables corresponding to different genes are assumed independent. We want to estimate the differential expression

$$\boldsymbol{\delta}_g = C\,\boldsymbol{\gamma}_g \tag{2}$$

or we want to test for differential expression

$$\begin{aligned} H_0 &: \boldsymbol{\delta}_g = \mathbf{0} \\ H_A &: \boldsymbol{\delta}_g \neq \mathbf{0} \ . \end{aligned} \tag{3}$$

In the current version of WAME [2,3] the estimation of the covariance-structure matrix Σ is based on a temporary assumption of expectation zero, $\boldsymbol{\mu}_g = \mathbf{0}$, for all genes, which is shown to give reasonable results if the expectation is close to zero for most genes. Thus, this is a suitable assumption for data with paired observations and few regulated genes between the pair-wise measured conditions.

The WAME model can be compared with the ordinary linear model (OLM) [4],

$$\mathbf{X}_g \sim \mathrm{N}(\boldsymbol{\mu}_g, c_g I) \tag{4}$$

which gives rise to the ordinary t- or F-tests, and with a widely used empirical Bayes model proposed in [5] and implemented in the LIMMA package [6],

$$\begin{aligned} \mathbf{X}_g \mid c_g &\sim \mathrm{N}(\boldsymbol{\mu}_g, c_g I) \ , \\ c_g &\sim \Gamma^{-1}(\alpha, \beta) \ . \end{aligned} \tag{5}$$

The novel feature of WAME was thus the introduction of the quality modelling covariance-structure matrix Σ.

After the introduction of WAME, a weighted version of LIMMA was proposed [7], which we will refer to as wLIMMA. There, a model with array-wise variance scales but no correlations is used,

$$\begin{aligned} \mathbf{X}_g \mid c_g &\sim \mathrm{N}(\boldsymbol{\mu}_g, c_g \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)) \ , \\ c_g &\sim \Gamma^{-1}(\alpha, \beta) \ . \end{aligned} \tag{6}$$

The parameters are estimated using a restricted maximum-likelihood (REML) approach.

A widely used approach is to only consider the ordinary least-squares estimated differential expression, often referred to as the log fold-change, here abbreviated as FC, or as the average M-value. In the present paper, the ranking of the genes imposed by this method will be included in comparisons, when applicable.

Results

The new version of WAME

In the current version of WAME [2, 3] the covariance-structure matrix Σ is estimated using a temporary assumption that $\boldsymbol{\mu}_g = \mathbf{0}$ for most genes, i.e. that the measurements of most genes consist solely of biological and technical noise. In the new version of WAME we relax this to only assume that most genes are non-differentially expressed, i.e. $\boldsymbol{\delta}_g = \mathbf{0}$. This allows a much larger class of experimental designs and design matrices D, most notably unpaired designs.

The trick used is to transform the data and consider

$$\mathbf{Y}_g = \mathbf{X}_g - \tilde{\boldsymbol{\mu}}_g^0 \tag{7}$$

where $\tilde{\boldsymbol{\mu}}_g^0$ is a suitable linear estimator of $\boldsymbol{\mu}_g$ which is unbiased under H_0 and which preserves the estimability of the differential expression $\boldsymbol{\delta}_g$, based on only the transformed data (see Methods for details). An example is (8) below where for each gene the mean value of all arrays is subtracted.

Since the transformed data contain only noise for non-differentially expressed genes by construction, the current version of WAME can essentially be applied to the transformed data $\mathbf{Y}_g$. As before, the covariance-structure matrix (now Σ_Y) and the hyperparameter α are first estimated under a provisional assumption (now $\boldsymbol{\delta}_g = \mathbf{0}$). The maximum likelihood estimates of $\boldsymbol{\delta}_g$ and the likelihood ratio test statistics of (3) are then computed. The tests and estimators are in fact unchanged by the transformation (7), if the

covariance-structure matrices for the transformed and untransformed data are known (details given in Methods). WAME is implemented as a package for the R language [8] and is available at `http://wame.math.chalmers.se/`.

Evaluation on real and resampled data

To investigate the properties of the new version of WAME, two real datasets are examined. Briefly, they are analysed both using WAME and the current methods described in Background. Array-specific weights, p-value distributions and rankings are produced showing clear differences between the procedures, most notably in the p-value distributions. To investigate the power of the different procedures and to look at p-value distributions in a controlled but realistic setting, we also analyse simulated data with real noise from the studied datasets and synthetic signal.

Description of the real datasets

Two public one-channel microarray datasets are analysed. The datasets are selected from the NCBI GEO database [9] with the criteria of having unpaired design and being sufficiently large to allow for the resample-based simulations in Resampled data below.

In the first dataset [10], biopsies were taken from the left atrium from 20 human hearts with normal sinus rythm and 10 hearts with permanent atrial fibrillation. It is here referred to as Atrium. In the second dataset [11], mechanisms in chronic obstructive pulmonary disease, COPD, were investigated by taking lung tissue biopsies from 12 smokers with mild or no emphysema and from 18 smokers with severe emphysema. In both datasets one Affymetrix HGU-133A array was used for each patient. In the present paper RMA [12] is used to obtain expression measures from the raw probe-wise intensities. The analyses are performed using the R language and the Bioconductor framework [13] .

Analysis of the real datasets

A natural parameterisation of the included datasets is to have one parameter per condition, yielding design and hypothesis matrices

$$D = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \text{ and } C = \begin{bmatrix} -1 & 1 \end{bmatrix} .$$

Under the null hypothesis, for each gene g and array i, an unbiased estimator of the expected value of the measurement X_{ig} is obtained by the gene-wise mean value over all arrays from both groups. The transformation then becomes a subtraction of that mean value, cf. (7),

$$Y_{ig} = X_{ig} - \frac{1}{n} \sum_{j=1}^{n} X_{jg} \ . \tag{8}$$

Note how the transformation preserves the difference in mean value between the two groups of arrays.

If the elements in $\mathbf{X}_g$ from the different arrays had in fact independent and identically distributed noise for each fixed gene g as assumed in OLM and unweighted LIMMA, the noise in $\mathbf{Y}_g$ would have equal variances for all arrays. In Figure 1 array-wise density estimates for the transformed expression values are shown. For arrays from the same condition the distributions should be identical, reflecting the combined variability of signal and noise. For unregulated genes the expectation of $\mathbf{Y}_g$ is zero, so if the assumption of few regulated genes holds the densities from all arrays should furthermore be essentially equal. Examination of Figure 1 reveals that neither of these statements are true, indicating that some variances are highly unequal.

Analogously, all pairs of arrays within each condition should have a common joint distribution and when few genes are regulated all pairs of arrays should essentially have a common joint distribution with a small negative correlation of $-1/(n-1)$. Examination of scatter plots for all pairs of arrays shows that this is clearly not the case (some obvious examples are shown in Figure 2, all pairs are included in the web supplement).

As expected from the observations above, unequal variances and non-zero correlations are estimated in the analyses with WAME, giving rise to highly unequal weights in the estimates of the differential expressions (shown in Table 1). In fact, the sign of the weight for some arrays even get switched compared to the sign of the weight of the other arrays from the same condition. This is an effect of strong correlations combined with unequal variances. It is an issue which is further addressed in Discussion.

The analysis methods described in Background are applied to the data and p-values and ranks computed. The respective probability plots are shown in Figure 3, demonstrating that there are substantial differences in the distribution of p-values between the different statistics. Since correlations and unequal variances are observed, the model assumptions of the alternative standard methods do not seem to hold. The p-values could thereby have become optimistic. On the other hand, it cannot be ruled out that the temporary assumption in WAME of no regulated genes makes its p-values conservative, which could also partly explain the differences. These problems are studied below by use of resampled data.

A common alternative to using the p-values as measures of significance is to consider the ranking of the genes, induced by the p-values or test statistics, and to select a fixed number of top ranked genes for further investigations. In Table 2 the concordance of the ranked lists are shown. The results from the included methods differ, for instance those from WAME compared to the other methods. This is not surprising since high correlations and highly unequal variances were identified by WAME, giving rise to highly unequal weights.

Resampled data

To examine closer the effect of violated assumptions of independence and identical distribution, we repeatedly selected two random subgroups of four arrays from within one group in the original data and performed tests between those groups. This was performed 100 times for the largest group in each of the two real datasets. Differentially expressed genes have unequal expected values in the two populations being sampled (cf. (2)). Since we now sample twice from the same condition, no differentially expressed genes exist.

Figure 4 shows the empirical p-value distributions for the resampled COPD data analysed with the four methods, together with the respective average empirical distribution,

$$\overline{F}(p) = \frac{1}{100}\sum_{i=1}^{100} F_i(p) \ ,$$

where F_i denotes the empirical CDF from the ith of the 100 resamples. For WAME, the p-value distributions are very close to the expected uniform. For OLM, LIMMA and weighted LIMMA there is a high variability between the p-value distributions and they are in many cases substantially different from the expected uniform. For WAME, OLM and LIMMA, the respective average empirical distribution is approximately correct, while for weighted LIMMA it is clearly optimistic. The results for the Atrium

dataset (see the web supplement) are very similar.

Evaluation of power

To evaluate the power of the tests in the studied datasets, a known regulation is added to randomly selected genes in one of the resampled groups, created according to the previous section. Thus, the noise is obtained from the real data and only the signal is synthetic. Ideally, the power can then be estimated by the proportion of differentially expressed genes that have a computed p-value less than a fixed level. However, valid p-values of the test statistics cannot be obtained from the respective models since, as demonstrated above, the corresponding assumptions are typically not valid. Ideally, the p-values would be determined by the true null distribution of the respective test statistics, given the array-wise quality deviations. In the simulation study, the critical value of the test statistics are therefore estimated from the empirical distribution of the test statistic for the unregulated genes. This is used to estimate the power of the different statistics (details are given in Methods).

The power estimates for the different methods are shown in Figure 6, for a level 0.1% test. The 0.1% level yields approximately 22 false positives if relatively few genes are in fact differentially expressed. For WAME, Σ is estimated both before and after adding a signal to 2228 genes (10%), thereby substantially affecting the estimate of Σ (cf. Figure 5). The powers of the two versions are nevertheless very similar (difference less than 0.003) and only the latter version is included in the plot.

When the covariance-structure matrix Σ is estimated in WAME it is assumed that no genes are differentially expressed. Figure 5 includes the average empirical distribution for the p-values of the unregulated genes when different proportions of the genes have a $\log_2$ differential expression of 1. It is clear that the distributions are biased for high proportions, giving conservative p-values, which should be an effect of biased estimates of Σ.

The results from the studied datasets indicate (i) that WAME offers a relevant power increase compared to the included alternatives, (ii) that weighted LIMMA does not offer an advantage compared to LIMMA and (iii) that the moderated statistics (WAME, LIMMA and wLIMMA) are superior to the traditional methods of ranking by ordinary t-statistic (OLM) or estimated differential expression (FC).

Discussion

The WAME model and the simulations

The WAME model aims at catching quality deviations by one covariance-structure matrix common for all genes. This is certainly simplistic in some cases, e.g. when only certain physical parts of an array or certain types of mRNAs are of decreased quality. The estimated covariance structure can then only be expected to reflect a mixture of the qualities of the different genes. However, examining the simulations (Figure 6), we see a clear power gain in the WAME model compared to the other models. Also, WAME succeeds in catching enough of the quality deviations to make the p-value distributions more correct, thus providing increased usefulness of the p-values (Figure 3).

The models of LIMMA, weighted LIMMA and WAME are nested, where weighted LIMMA adds unequal variances and WAME adds unequal variances and correlations. Examination of Figure 1 shows that there are evident differences in variability between arrays. It is therefore interesting that we have not found a power increase of weighted LIMMA compared to LIMMA. Further, the p-values of weighted LIMMA turned out to be too optimistic (Figure 4). Comparison with the results of the WAME method, where the power increases and the p-value distributions get substantially more correct, suggests that the correlations are crucial in the model.

In the simulations, noise is taken from real data through resampling within a fixed group. This procedure provides data with fewer assumption on the noise structure compared to a fully parameterised simulation and should hopefully better reflect realistic situations. To evaluate the power of the different methods, a synthetic signal which is constant within each condition is added to the resample-based noise. This follows the assumption in the models of both WAME, OLM, LIMMA and weighted LIMMA, that the noise structure is equal for genes that are differentially expressed and non-differentially expressed. However, the biological variability of the expression of differentially expressed genes might be different under the different conditions due to the changed rôle of those genes. For complicated conditions such as complex diseases, the problem is more severe (cf. [14–16]) since crucial genes might only be differentially expressed in a subset of the studied arrays. Further work is needed to evaluate the performance of WAME in such settings, as well as to possibly expand it to better fit these situations.

A relevant question regarding the modelling of quality deviations by the covariance-structure matrix Σ is whether biologically interesting features may be hidden by this model. In the present datasets, the

question can partly be answered by examining the pairwise plots (cf. Figure 2) and noticing that a large proportion of the genes show similar deviations, which should speak against a specific interesting biological explanation. Also, the estimated covariance structure matrix Σ can be inspected with the goal of finding relevant correlations between arrays and thus highlighting interesting features in the data. Possible future work is to use such inspections to reveal unwanted features in normalisation or in preprocessing wet-lab steps that give rise to correlated errors for a large proportion of the genes.

Weights with switched signs

In the studied datasets, strong correlations combined with unequal variances make some weights within a group switch sign, in essence meaning that it is beneficial to partly subtract some arrays within a group in the estimate to be able to add more of the others in the same group (cf. Table 1). Since this might seem counter-intuitive, an elucidating example of possible mechanisms behind such weights follows.

Consider an example where two two-colour arrays are observed, X_1 and X_2. Let the two arrays have two sources of variation, one that is mutually independent (ϵ_1, ϵ_2) and one consisting of different proportions, a_1 and a_2, of one common source of variation η. Let ϵ_1, ϵ_2 and η be independent and normally distributed with expectation 0 and variances σ_ϵ^2 and σ_η^2, respectively. Furthermore, let μ be the parameter to be estimated. The model becomes

$$X_i = \mu + a_i\eta + \epsilon_i \ , i \in \{1, 2\} \ .$$

Then, X_1 gets a negative weight if and only if

$$a_1 > a_2 + \frac{\sigma_\epsilon^2}{a_2\sigma_\eta^2} \ ,$$

i.e. if array 1 includes a large enough contribution from the common source of variation. When a negative weight is allowed instead of removing the array, a smaller proportion of the common source of variation is included in the final estimate. Its precision is thus increased.

Validity of the p-values and derived entities

Varying quality of arrays and correlated errors were demonstrated in [2, 3] and in the present paper through examination of the data. These questions are typically neglected in microarray analyses, since independence and identical distribution or exchangeability are generally assumed under the null hypothesis. Thus, the validity is questionable of the corresponding p-values and their derived entities, e.g.

false discovery rates and estimates of proportions of differentially expressed genes. This problem is obvious in the resample based simulations.

A number of experiments have been analysed (data not shown) in addition to those published in the present paper and in [2,3]. In almost all cases relevant unequal variances and correlations have been identified, indicating that the problem is common.

In the resample based simulations with added signal, WAME is shown to be conservative, which is an effect of the biased estimate of Σ. Further work on an estimator of Σ with better characteristics under regulation is therefore needed. However, the simulations indicate (i) that the power of the test is basically unaffected by the bias and (ii) that hundreds of genes may be differentially expressed (two-fold) with only mildly conservative p-values as result.

Correlations between genes or between arrays?

It has recently been argued that the expression of different genes are highly dependent, making the law of large number normally inapplicable [17] and standard estimators of e.g. the false discovery rate (FDR) imprecise [18]. In [18], a latent FDR is introduced, being the conditional FDR given a random effect b that captures the correlation effects between genes. The FDR is then the marginal latent FDR, that is the average over the random effect b.

For the datasets examined in the present paper, the model assumptions of e.g. the ordinary linear model are shown not to hold (cf. Figure 1 and Figure 2). This can be expected to result in invalid p-values, which is indeed observed in Figure 4. Interestingly, the p-value distribution seem to be valid marginally, i.e. on average over the different resamples, which would yield valid but imprecise estimates of the FDR. This type of failed model assumptions is not taken into account in e.g. [17,18]. Since for a performed experiment, the p-values from the ordinary t-statistic (OLM) share a common bias conditional on the experiment (see Figure 4), the different p-values may be highly dependent. However, this dependency is due to failure of taking array-wide quality deviations into account in the model and not due to the nature of microarray data *per se*, e.g. through substantial long-range gene-gene interactions.

Consequently, the strong observed dependencies between statistics from different genes might largely be explainable by quality deviations between the arrays in the experiment, e.g. correlations between arrays. Since WAME models these deviations such that the p-values are essentially correctly distributed when few

genes are differentially expressed in the studied datasets, the dependency between genes should be greatly decreased. The covariance structure matrix Σ is therefore in a sense a parallel to the random factor b in [18]. It remains as future work to evaluate the gene-gene dependencies and estimates of e.g. the FDR in the context of the WAME model.

In the WAME model, the data from different genes are assumed independent, which is unrealistic, e.g. since genes act together in pathways. However, this is only used in the derivation of the maximum likelihood estimaties of the covariance structure matrix Σ and the shape parameter α. The assumption could thus be relaxed to a dependence between the different genes that is weak enough that the estimates of Σ and α become precise, and accurate under H_0. This holds if the law of large numbers is applicable for averages of certain functions of the gene-wise observed data (cf. the likelihood functions in [2, 3]). Given the large number of genes and the observed p-value distributions in Figure 4, this relaxed assumption seems plausible.

It can be noted that for the studied data, WAME has higher power and considerably more valid p-values than weighted LIMMA. Since the difference between the weighted LIMMA and WAME models is the inclusion of correlations between arrays, this emphasises the importance of the correlations in the model.

Conclusions

Statistical methods in microarray analysis are typically based on the often erroneous assumption that the data from different arrays are independent and identically distributed. An exception is Weighted Analysis of Microarray Experiment (WAME) where heteroscedasticity and correlations between arrays are modelled by a covariance-structure common for all genes. In the present paper, WAME has been extended to handle datasets without a natural pairing, e.g. data from one-channel microarrays, and corresponding estimates and test statistics have been derived. In the examined one-channel microarray datasets WAME detected unequal variances and nonzero correlations.

WAME was compared with four other common methods: an ordinary linear model with t-tests, LIMMA, weighted LIMMA, and fold-change ranking. The comparison was performed using resampling of the different arrays within the datasets. Here, WAME had the highest power. When a relatively small proportion of the genes are regulated, WAME also generates close to correct p-value distributions while the p-value distributions from the other methods are highly variable. However, when many genes are

differentially expressed, the p-values from WAME tend to be conservative.

In conclusion, p-values from the standard methods for microarray analysis should in general not be trusted and any result based on p-values, e.g. estimates of the number of regulated genes and false discovery rates, should be interpreted with care. The analyses of the examined datasets showed that WAME gives a powerful approach for finding differentially expressed genes and that it produces more trustworthy p-values when a relatively small proportion of genes are differentially expressed.

Methods

Details on the new version of WAME

Model Framework

For $g = 1, \ldots, m$, let $\mathbf{X}_g$ be an n-dimensional vector with expectation $\boldsymbol{\mu}_g = D\,\boldsymbol{\gamma}_g$, where D is the design matrix, having rank k, and $\boldsymbol{\gamma}_g \in \mathbb{R}^q$ is the parameter vector. Furthermore, let

$$\mathbf{X}_g \mid c_g \sim \mathrm{N}(\boldsymbol{\mu}_g, c_g\Sigma)\ ,$$
$$c_g \sim \Gamma^{-1}(\alpha, 1)\ ,$$

where Σ is the non-singular covariance-structure matrix, c_g is the variance-scaling factor, α is the shape parameter for c_g and $(c_1, \mathbf{X}_1), \ldots, (c_m, \mathbf{X}_m)$ are assumed independent. The differential expression vector is defined as

$$\boldsymbol{\delta}_g = C\,\boldsymbol{\gamma}_g\ ,$$

where C is a matrix of rank p such that $\boldsymbol{\delta}_g$ is estimable. Here, an estimator of $\boldsymbol{\delta}_g$ and a test for

$$\begin{aligned} H_0 &: \boldsymbol{\delta}_g = \mathbf{0} \\ H_A &: \boldsymbol{\delta}_g \neq \mathbf{0} \end{aligned} \tag{9}$$

are in focus.

As mentioned in Background, one main obstacle is that Σ is hard to estimate. In fact, Σ and $\boldsymbol{\delta}_g$ cannot be maximum likelihood estimated simultaneously, since there are trivial infinite suprema of the likelihood, e.g. when the variance of one observation is set to zero and the corresponding mean is selected so that it equals that observation.

The current WAME method

In the current version of WAME [3], Σ is estimated as follows. First, temporarily assume that $\boldsymbol{\mu}_g = \mathbf{0}$ for all genes, which is reasonable for paired experimental designs with few differentially expressed genes between any pairwise measured conditions. For each gene, the variance scaling factor c_g is removed by dividing the n measurements with the first measurement, yielding a random vector distributed according to a multivariate generalisation of the Cauchy distribution. A scaled version of Σ is then maximum likelihood estimated numerically. Second, the unknown scale and the hyperparameter α of the prior distribution of c_g are maximum likelihood estimated numerically without the assumption of $\boldsymbol{\mu}_g = \mathbf{0}$. The parameters Σ and α are subsequently treated as known in the maximum-likelihood estimates and likelihood-ratio tests for the different genes.

The new WAME method

The new version of WAME relaxes the assumption from $\boldsymbol{\mu}_g = \mathbf{0}$ to $\boldsymbol{\delta}_g = \mathbf{0}$, which incorporates many designs without a natural pairing. This is performed by subtracting an arbitrary estimator $\tilde{\boldsymbol{\mu}}_g^0$ of $\boldsymbol{\mu}_g$, which is unbiased under H_0 and has as image the space $\mathcal{V}_0$ of possible values for $\boldsymbol{\mu}_g$ under H_0,

$$\mathbf{Y}_g = \mathbf{X}_g - \tilde{\boldsymbol{\mu}}_g^0 \, . \tag{10}$$

It can be shown that this transformation preserves the estimability of $\boldsymbol{\delta}_g$.

By construction, the transformed data $\mathbf{Y}_g$ will have expectation zero for non-differentially expressed genes and the current WAME method can be applied on $\mathbf{Y}_g$, including the estimation of the covariance-structure matrix Σ_Y for $\mathbf{Y}_g$. It will now be proved that the likelihood ratio tests of (9) and the maximum likelihood estimates of $\boldsymbol{\delta}_g$ based on $\mathbf{X}_g$ or $\mathbf{Y}_g$ are identical, if α and Σ or Σ_Y respectively are considered known.

We shall henceforth consider a fixed gene g and drop the g index.

Equality of tests and estimators

Before beginning, some further definitions are needed. Define the Mahalanobis inner product corresponding to a symmetric n by n matrix A as

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_A = \mathbf{x}_1^{\mathrm{T}} \, A^{-} \, \mathbf{x}_2 \ , \tag{11}$$

and the norm $\|\cdot\|_A$ as

$$\|\,\mathbf{x}\,\|_A^2 = \langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^{\mathrm{T}}\, A^{-}\, \mathbf{x}\ ,$$

where $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2$ lies in the rowspace of A and the generalised inverse A^- is any matrix satisfying $AA^-A = A$. Let $\mathcal{X}$ denote the n-dimensional inner product space with $\langle \cdot, \cdot \rangle_\Sigma$ as inner product. Define $\mathcal{V} \subset \mathcal{X}$ as the space of possible values for $\boldsymbol{\mu}_g$,

$$\mathcal{V} = \{\boldsymbol{\mu} : \boldsymbol{\mu} = D\,\boldsymbol{\gamma},\ \boldsymbol{\gamma} \in \mathbb{R}^q\}$$

and let $\mathcal{V}_0 \subset \mathcal{X}$ denote the corresponding space restricted by the null hypothesis,

$$\mathcal{V}_0 = \{\boldsymbol{\mu} : \boldsymbol{\mu} = D\,\boldsymbol{\gamma},\ C\,\boldsymbol{\gamma} = \mathbf{0},\ \boldsymbol{\gamma} \in \mathbb{R}^q\}\ .$$

Proposition *Let $\tilde{\boldsymbol{\mu}}^0$ be an arbitrary linear estimator of $\boldsymbol{\mu}$, which is unbiased under H_0 and which has image $\mathcal{V}_0$. Let*

$$\mathbf{Y} = \mathbf{X} - \tilde{\boldsymbol{\mu}}^0\ ,$$

and let Σ_Y be the covariance-structure matrix of $\mathbf{Y}$. Then the likelihood ratio test of (9) and the maximum likelihood estimate of $\boldsymbol{\delta}$ based on $\mathbf{X}$ with Σ and α known are identical to the ones based on $\mathbf{Y}$ with Σ_Y and α known.

Proof of the Proposition

The proof is divided into two steps which combined conclude the proof.

1. The likelihood ratio test (LRT) of (9) and the maximum likelihood estimator (MLE) of $\boldsymbol{\delta}$ does not depend on the choice of $\tilde{\boldsymbol{\mu}}^0$.
2. The proposition holds when $\tilde{\boldsymbol{\mu}}^0$ is the MLE of $\boldsymbol{\mu}$ under H_0.

Proof of step 1

Let $\boldsymbol{\mu}'$ and $\boldsymbol{\mu}''$ be two valid choices of $\tilde{\boldsymbol{\mu}}^0$, i.e. they are both unbiased estimators of $\boldsymbol{\mu}$ under H_0 and have $\mathcal{V}_0$ as image. Let $\mathbf{Y}' = \mathbf{X} - \boldsymbol{\mu}'$ and $\mathbf{Y}'' = \mathbf{X} - \boldsymbol{\mu}''$. Recall that a matrix P is a projection matrix projecting on $\mathcal{V}_0$ if and only if for all $\mathbf{x} \in \mathbb{R}^n$, $Px \in \mathcal{V}_0$ and for all $\mathbf{x}_0 \in \mathcal{V}_0$, $P\,\mathbf{x}_0 = \mathbf{x}_0$. It can be shown that $\boldsymbol{\mu}'$ and $\boldsymbol{\mu}''$ can be written as $\boldsymbol{\mu}' = P'X$ and $\boldsymbol{\mu}'' = P''X$ for some projection matrices P' and P'' projecting on $\mathcal{V}_0$.

Since P' and P'' project on the same space it follows that $P'P'' = P''$ and $P''P' = P'$, and thus $(I - P')\,\mathbf{Y}'' = \mathbf{Y}'$ and $(I - P'')\,\mathbf{Y}' = \mathbf{Y}''$. Hence there is an invertible map between $\mathbf{Y}'$ and $\mathbf{Y}''$ and likelihood methods based on $\mathbf{Y}'$ and $\mathbf{Y}''$ respectively will give equal results. Consequently, the MLE of (9) and the LRT of $\boldsymbol{\delta}$ will not depend on the choice of $\tilde{\boldsymbol{\mu}}^0$

Proof of step 2

Since $\boldsymbol{\delta}$ is estimable based on $\mathbf{X}$, there exist a matrix A such that $C = AD$ and thus $\boldsymbol{\delta} = A\,\boldsymbol{\mu}$. The likelihood of $\boldsymbol{\mu}$ can therefore be examined instead of the likelihood of $\boldsymbol{\delta}$.

The likelihood of $\boldsymbol{\mu}$ based on $\mathbf{X}$ can be shown to be

$$\begin{aligned} L\left(\boldsymbol{\mu} \,|\, \mathbf{X}\right) &= \int_0^\infty f(\mathbf{X} \,|\, \boldsymbol{\mu}, c) \cdot f(c)\; dc \\ &\propto \left[\| \,\mathbf{X} - \boldsymbol{\mu}\, \|^2_\Sigma / 2 + 1\right]^{-n/2-\alpha} \;, \end{aligned} \tag{12}$$

where $\propto$ denotes proportionality. Using the Projection Theorem [19], the MLE of $\boldsymbol{\mu}$ is the orthogonal projection of $\mathbf{X}$ on $\mathcal{V}$,

$$\hat{\boldsymbol{\mu}} = \mathcal{P}_{\mathcal{V}}\,\mathbf{X} \;,$$

where the orthogonality is according to the inner product of $\mathcal{X}$. When H_0 is true, $\boldsymbol{\mu}$ is restricted to $\mathcal{V}_0$ and thus the MLE of $\boldsymbol{\mu}$ becomes

$$\hat{\boldsymbol{\mu}}^0 = \mathcal{P}_{\mathcal{V}_0}\,\mathbf{X} \;.$$

Note that $\hat{\boldsymbol{\mu}}^0$ is a valid choice for $\tilde{\boldsymbol{\mu}}^0$, i.e. $\hat{\boldsymbol{\mu}}^0$ is unbiased under H_0 and has $\mathcal{V}_0$ as image. Let

$$\mathbf{Z} = \mathbf{X} - \hat{\boldsymbol{\mu}}^0 \;,$$

which gives $\mathbf{Z} = \mathcal{P}_{\mathcal{V}_0^\perp}\,\mathbf{X}$, where $\mathcal{V}_0^\perp$ denotes the orthogonal complement of $\mathcal{V}_0$ in $\mathcal{X}$. Standard properties of the normal distribution gives

$$\mathbf{Z} \,|\, c \sim \mathrm{N}(\boldsymbol{\mu}_z, c\Sigma_z) \;,$$

where $\boldsymbol{\mu}_z = D_z\,\boldsymbol{\gamma}$ with $\mathrm{D}_z = \mathcal{P}_{\mathcal{V}_0^\perp}\,D$, and where $\Sigma_z = \mathcal{P}_{\mathcal{V}_0^\perp}\,\Sigma\,\mathcal{P}^{\mathrm{T}}_{\mathcal{V}_0^\perp}$.

The likelihood function of $\boldsymbol{\mu}_z$ (with respect to the Lebesgue measure on the space of possible values of $\mathbf{Z}$ spanned by the column vectors of Σ_z) can be written as

$$L\left(\boldsymbol{\mu} \,|\, \mathbf{Z}\right) \propto \left[\| \,\mathbf{Z} - \boldsymbol{\mu}_z\, \|^2_{\Sigma_z} / 2 + 1\right]^{-n/2-\alpha} \;.$$

Since, $\boldsymbol{\delta}$ is estimable based on $\mathbf{Z}$, the likelihood of $\boldsymbol{\mu}_z$ can be examined instead of the likelihood of $\boldsymbol{\delta}$.

The likelihood ratio statistic of (9) based on $\mathbf{X}$ is defined by

$$T = \frac{\sup_{\boldsymbol{\mu}\in\mathcal{V}} L(\boldsymbol{\mu}\,|\,\mathbf{X})}{\sup_{\boldsymbol{\mu}\in\mathcal{V}_0} L(\boldsymbol{\mu}\,|\,\mathbf{X})} ,$$

which can be rewritten (cf. [3]) as a strictly increasing function of

$$\begin{aligned} T' &= \frac{n-p+2\alpha}{k} \frac{\|\,\mathcal{P}_{\mathcal{V}}\,\mathbf{X} - \mathcal{P}_{\mathcal{V}_0}\,\mathbf{X}\,\|^2_{\Sigma}}{\|\,\mathbf{X} - \mathcal{P}_{\mathcal{V}}\,\mathbf{X}\,\|^2_{\Sigma} + 2} \\ &= \frac{n-p+2\alpha}{k} \frac{\|\,\mathcal{P}_{\mathcal{V}\cap\mathcal{V}_0^{\perp}}\,\mathbf{X}\,\|^2_{\Sigma}}{\|\,\mathcal{P}_{\mathcal{V}^{\perp}}\,\mathbf{X}\,\|^2_{\Sigma} + 2} , \end{aligned} \tag{13}$$

where $\mathcal{V}^{\perp}$ and $\mathcal{V}_0^{\perp}$ are the orthogonal complements of $\mathcal{V}$ and $\mathcal{V}_0$ respectively.

Note that the space of possible values for $\boldsymbol{\mu}_z$ is $\mathcal{V}\cap\mathcal{V}_0^{\perp}$ and that $\boldsymbol{\mu}_z = 0$ under H_0. Let $\mathcal{P}^z$ denote the orthogonal projection according to $\langle\cdot,\cdot\rangle_{\Sigma_z}$. Then, the likelihood ratio statistic of (9) based on $\mathbf{Z}$ can in analogy with (13) be shown to be a strictly increasing function of

$$T'_z = \frac{n-p+2\alpha}{k} \frac{\|\,\mathcal{P}^z_{\mathcal{V}\cap\mathcal{V}_0^{\perp}}\,\mathbf{Z}\,\|^2_{\Sigma_z}}{\|\,\mathbf{Z} - \mathcal{P}^z_{\mathcal{V}\cap\mathcal{V}_0^{\perp}}\,\mathbf{Z}\,\|^2_{\Sigma_z} + 2} . \tag{14}$$

The Lemma below yields that for all $\mathcal{W}\subseteq\mathcal{V}_0^{\perp}$ and all $\mathbf{z}\in\mathcal{V}_0^{\perp}$, $\|\,\mathbf{z}\,\|^2_{\Sigma_z} = \|\,\mathbf{z}\,\|^2_{\Sigma}$ and $\mathcal{P}^z_{\mathcal{W}}\,\mathbf{z} = \mathcal{P}_{\mathcal{W}}\,\mathbf{z}$. The equivalence of the test statistics (13) and (14) is now straight-forward,

$$\begin{aligned} T'_z &= \frac{n-p+2\alpha}{k} \frac{\|\,\mathcal{P}^z_{\mathcal{V}\cap\mathcal{V}_0^{\perp}}\,\mathbf{Z}\,\|^2_{\Sigma_z}}{\|\,\mathbf{Z} - \mathcal{P}^z_{\mathcal{V}\cap\mathcal{V}_0^{\perp}}\,\mathbf{Z}\,\|^2_{\Sigma_z} + 2} \\ &= \frac{n-p+2\alpha}{k} \frac{\|\,\mathcal{P}_{\mathcal{V}\cap\mathcal{V}_0^{\perp}}\,\mathcal{P}_{\mathcal{V}_0^{\perp}}\,\mathbf{X}\,\|^2_{\Sigma}}{\|(\mathcal{P}_{\mathcal{V}} + \mathcal{P}_{\mathcal{V}^{\perp}})(\mathcal{P}_{\mathcal{V}_0^{\perp}}\,\mathbf{X} - \mathcal{P}_{\mathcal{V}\cap\mathcal{V}_0^{\perp}}\,\mathcal{P}_{\mathcal{V}_0^{\perp}}\,\mathbf{X})\|^2_{\Sigma} + 2} \\ &= \frac{n-p+2\alpha}{k} \frac{\|\,\mathcal{P}_{\mathcal{V}\cap\mathcal{V}_0^{\perp}}\,\mathbf{X}\,\|^2_{\Sigma}}{\|\,\mathcal{P}_{\mathcal{V}^{\perp}}\,\mathbf{X}\,\|^2_{\Sigma} + 2} = T' . \end{aligned} \tag{15}$$

Lemma *Let $\mathcal{W}$ be a subspace of $\mathcal{X}$ and let $\mathcal{P}_{\mathcal{W}}$ be the orthogonal projection from $\mathcal{X}$ onto $\mathcal{W}$. Then for any $\mathbf{x}_1,\ \mathbf{x}_2\in\mathcal{W}$,*

$$\langle\mathbf{x}_1,\mathbf{x}_2\rangle_{\Sigma} = \langle\mathbf{x}_1,\mathbf{x}_2\rangle_{\Sigma_{\mathcal{W}}} ,$$

where $\Sigma_{\mathcal{W}} = \mathcal{P}_{\mathcal{W}}\,\Sigma\,\mathcal{P}_{\mathcal{W}}$.

Proof Let A be a matrix of a change of basis [19] from the standard basis to an orthonormal basis of $\mathcal{X}$ such that the first l basis vectors span $\mathcal{W}$. Let $I_{(l)}$ denote the identity matrix with all but the l top left

diagonal elements set to zero. It follows that $A^{\mathrm{T}}A = \Sigma^{-1}$ and $A\,\mathcal{P}_{\mathcal{W}} = I_{(l)}A$ and therefore,

$$\begin{aligned}
\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_{\Sigma} &= \mathbf{x}_1^{\mathrm{T}}\,\Sigma^{-1}\,\mathbf{x}_2 \\
&= \mathbf{x}_1^{\mathrm{T}}\,A^{\mathrm{T}}A\,\mathcal{P}_{\mathcal{W}}\,\mathbf{x}_2 \\
&= \mathbf{x}_1^{\mathrm{T}}\,A^{\mathrm{T}}\left(I_{(l)}\right)^{-}A\,\mathbf{x}_2 \\
&= \mathbf{x}_1^{\mathrm{T}}\,A^{\mathrm{T}}\left(I_{(l)}A\Sigma A^{\mathrm{T}}I_{(l)}\right)^{-}A\,\mathbf{x}_2 \\
&= \mathbf{x}_1^{\mathrm{T}}\,A^{\mathrm{T}}\left(A\,\mathcal{P}_{\mathcal{W}}\,\Sigma\,\mathcal{P}_{\mathcal{W}}^{\mathrm{T}}\,A^{\mathrm{T}}\right)^{-}A\,\mathbf{x}_2 \\
&= \mathbf{x}_1^{\mathrm{T}}\left(\mathcal{P}_{\mathcal{W}}\,\Sigma\,\mathcal{P}_{\mathcal{W}}^{\mathrm{T}}\right)^{-}\mathbf{x}_2\ ,
\end{aligned}$$

where the last equality uses the fact that $(AB)^{-} = B^{-}A^{-1}$ when A is invertible .

□

The next step is to show that the MLE of $\boldsymbol{\delta}$ when $\mathbf{X}$ is observed is identical to the MLE of $\boldsymbol{\delta}$ when $\mathbf{Z}$ is observed. The former is defined by

$$\hat{\boldsymbol{\delta}} = C\hat{\boldsymbol{\gamma}} = C\operatorname*{argmin}_{\boldsymbol{\gamma}} \|\,\mathbf{X} - D\,\boldsymbol{\gamma}\,\|_{\Sigma}^{2}\ .$$

Define $\mathcal{G}_0 = \{\boldsymbol{\gamma} : D\,\boldsymbol{\gamma} \in \mathcal{V}_0\}$ and $\mathcal{G}_1 = \{\boldsymbol{\gamma} : D\,\boldsymbol{\gamma} \in \mathcal{V}_0^{\perp}\}$ and note that for any $\boldsymbol{\gamma}$ there exist $\boldsymbol{\gamma}_0 \in \mathcal{G}_0$ and $\boldsymbol{\gamma}_1 \in \mathcal{G}_1$ such that $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1$. Thus,

$$\hat{\boldsymbol{\delta}} = C\operatorname*{argmin}_{\boldsymbol{\gamma}_0+\boldsymbol{\gamma}_1:\boldsymbol{\gamma}_0\in\mathcal{G}_0,\boldsymbol{\gamma}_1\in\mathcal{G}_1} \|\,\mathbf{X} - D(\boldsymbol{\gamma}_0+\boldsymbol{\gamma}_1)\|_{\Sigma}^{2}\ .$$

Now, since $\mathcal{P}_{\mathcal{V}_0^{\perp}} + \mathcal{P}_{\mathcal{V}_0} = I$,

$$\begin{aligned}
\hat{\boldsymbol{\delta}} &= C\operatorname*{argmin}_{\boldsymbol{\gamma}_0+\boldsymbol{\gamma}_1:\boldsymbol{\gamma}_0\in\mathcal{G}_0,\boldsymbol{\gamma}_1\in\mathcal{G}_1} \left(\|\,\mathcal{P}_{\mathcal{V}_0}(\mathbf{X} - D(\boldsymbol{\gamma}_0+\boldsymbol{\gamma}_1)) + \mathcal{P}_{\mathcal{V}_0^{\perp}}(\mathbf{X} - D(\boldsymbol{\gamma}_0+\boldsymbol{\gamma}_1))\|_{\Sigma}^{2}\right) \\
&= C\operatorname*{argmin}_{\boldsymbol{\gamma}_0+\boldsymbol{\gamma}_1:\boldsymbol{\gamma}_0\in\mathcal{G}_0,\boldsymbol{\gamma}_1\in\mathcal{G}_1} \left(\|\,\mathcal{P}_{\mathcal{V}_0}(\mathbf{X} - D\,\boldsymbol{\gamma}_0)\|_{\Sigma}^{2} + \|\,\mathcal{P}_{\mathcal{V}_0^{\perp}}(\mathbf{X} - D\,\boldsymbol{\gamma}_1)\|_{\Sigma_z}^{2}\right)\ ,
\end{aligned}$$

where the second equality follows from the generalised Theorem of Pythagoras [19], the Lemma, and the fact that $\mathcal{P}_{\mathcal{V}_0^{\perp}}\,D\,\boldsymbol{\gamma}_0 = 0$ and $\mathcal{P}_{\mathcal{V}_0}\,D\,\boldsymbol{\gamma}_1 = 0$. Now since $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ minimise the expression independently of each other and since $C\,\boldsymbol{\gamma}_0 = 0$ by construction,

$$\begin{aligned}
\hat{\boldsymbol{\delta}} &= C\left(\operatorname*{argmin}_{\boldsymbol{\gamma}_0\in\mathcal{G}_0} \|\,\mathcal{P}_{\mathcal{V}_0}(\mathbf{X} - D\,\boldsymbol{\gamma}_0)\|_{\Sigma}^{2} + \operatorname*{argmin}_{\boldsymbol{\gamma}_1\in\mathcal{G}_1} \|\,\mathbf{Z} - D_z\,\boldsymbol{\gamma}_1\,\|_{\Sigma_z}^{2}\right) \\
&= C\operatorname*{argmin}_{\boldsymbol{\gamma}_1\in\mathcal{G}_1} \|\,\mathbf{Z} - D_z\,\boldsymbol{\gamma}_1\,\|_{\Sigma_z}^{2}\ .
\end{aligned}$$

For all $\boldsymbol{\gamma}_0 \in \mathcal{G}_0$, $C\,\boldsymbol{\gamma}_0 = 0$ and $D_z\,\boldsymbol{\gamma}_0 = 0$, so the area of minimisation can be extended,

$$\hat{\boldsymbol{\delta}} = C\operatorname*{argmin}_{\boldsymbol{\gamma}} \|\,\mathbf{Z} - D_z\,\boldsymbol{\gamma}\,\|_{\Sigma_z}^{2}\ ,$$

which is the MLE of $\boldsymbol{\delta}$ based on $\mathbf{Z}$ by definition. □

Remark 1 Using the invertible map between any two choices of $\mathbf{Y}$, $\mathbf{Y}$ and $\mathbf{Y}'$, as defined in Step 1 above, the respective maximum likelihood estimates of α, Σ_y and $\Sigma_{y'}$ can be shown to be identical based on $\mathbf{Y}$ or $\mathbf{Y}'$. In this sense, the choice of $\tilde{\boldsymbol{\mu}}^0$ is thus truly irrelevant.

Remark 2 Sometimes, additional linear combinations of $\boldsymbol{\gamma}$ can be assumed to be zero for most genes, $C^* \boldsymbol{\gamma} = 0$ for some matrix C^* with rowspace being a superspace of the rowspace of C. Let P^* be any projection matrix on the corresponding space $\mathcal{V}_* = \{\boldsymbol{\mu} : \boldsymbol{\mu} = D\,\boldsymbol{\gamma},\ C^*\,\boldsymbol{\gamma} = \mathbf{0},\ \boldsymbol{\gamma} \in \mathbb{R}^q\}$ and let $\mathbf{Y}^* = \mathbf{X} - P^*\,\mathbf{X}$. It is straight forward to show that a variant of the Proposition still holds, so given the covariance structure matrices the inference results concerning $C\,\boldsymbol{\gamma}$ will be identical, based on $\mathbf{Y}$ or $\mathbf{Y}^*$ respectively. However, the estimates of the covariance structure matrices for $\mathbf{Y}$ and $\mathbf{Y}^*$ might not be coherent and the results are expected to differ slightly.

The estimator of power

Consider a certain experimental design, a level 1-λ test and a differential expression δ. Let a realisation of the experiment be given, which e.g. results in certain quality deviations between arrays. The conditional power is defined as the probability of identifying a random gene in the current experiment, i.e. conditional on e.g. the quality deviations, when the gene has differential expression δ. The power is then defined as the average conditional power over all possible realisations of the experimental design. The power is thus related to an unperformed experiment while the conditional power is related to a specific performed experiment. Here, the test is assumed to be valid conditional on the experiment, i.e. to have conditional power λ when $\boldsymbol{\delta} = \mathbf{0}$.

In Evaluation of power, the aim is to estimate the power for a hypothetical experiment where the distribution of the data under the null hypothesis is obtained by resampling of real data. For a given resample, a constant differential expression is added to randomly selected genes and the statistics t_g are computed. The estimate $\hat{t}_c$ of the conditional critical value is computed so that a proportion λ of the unregulated genes satisfy $|t_g| \geq \hat{t}_c$. The conditional power is then estimated by the proportion of regulated genes satisfying $|t_g| \geq \hat{t}_c$. The power is finally estimated by averaging the estimated conditional power over the resamples.

Authors contributions

ON provided initial ideas related to the generalisation. AS formulated the generalisation, performed the proofs in Methods, designed and programmed the analyses, simulations and plots. EK and ON helped refining the generalisation and Methods. AS, EK and MR drafted the manuscript. All authors continuously provided feedback on various parts of the work leading to the manuscript and approved the final version of the manuscript.

Acknowledgements

Lina Gunnarsson is acknowledged for providing feedback on the manuscript. AS and EK acknowledge the National Research School in Genomics and Bioinformatics for funding.

References

1. Auer H, Lyianarachchi S, Newsom D, Klisovic MI, Marcucci G, , Kornacker K: **Chipping away at the chip bias: RNA degradation in microarray analysis**. *Nature Genetics* 2003, **35**:292–293.
2. Kristiansson E, Sjögren A, Rudemo M, Nerman O: **Weighted Analysis of Paired Microarray Experiments**. *Statistical Applications in Genetics and Molecular Biology* 2005, **4**:Article 30.
3. Kristiansson E, Sjögren A, Rudemo M, Nerman O: **Quality Optimised Analysis of General Paired Microarray Experiments**. *Statistical Applications in Genetics and Molecular Biology* 2006, **5**:Article 10.
4. Arnold S: *The Theory of Linear Models and Multivariate Analysis*. John Wiley & Sons 1980.
5. Lönnstedt I, Speed T: **Replicated microarray data**. *Statistica Sinica* 2002, **12**:31–46.
6. Smyth G: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments**. *Statistical Applications in Genetics and Molecular Biology* 2004, **3**.
7. Ritchie M, Diyagama D, Neilson J, van Laar R, Dobrovic A, Holloway A, Smyth G: **Empirical array quality weights in the analysis of microarray data**. *BMC Bioinformatics* 2006, **7**:261.
8. R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2006, [http://www.R-project.org]. [ISBN 3-900051-07-0].
9. Edgar R, Domrachev M, Lash A: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Research* 2002, **30**:207–210.
10. Barth A, Merk S, Arnoldi E, Zwermann L, Kloos P, Gebauer M, Steinmeyer K, Bleich M, Kaab S, Hinterseer M, Kartmann H, Kreuzer E, Dugas M, Steinbeck G, Nabauer M: **Reprogramming of the Human Atrial Transcriptome in Permanent Atrial Fibrillation**. *Circulation Research* 2005, **96**(9):1022–1029.
11. Spira A, Beane J, Pinto-Plata V, Kadar A, Liu G, Shah V, Celli B, Brody J: **Gene expression profiling of human lung tissue from smokers with severe emphysema**. *American Journal of Respiratory Cell and Molecular Biology* 2004, **31**(6):601–610.
12. Irizarry R, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, Scherf U, Speed T: **Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data**. *Biostatistics* 2004, **4**(2):249–264.
13. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Friedrich L, Li C, Maechler M, Rossini A, Sawitzki G, Smith C, Smyth G, Tierney L, Yang J, Zhang J: **Bioconductor: Open software development for computational biology and bioinformatics**. *Genome Biology* 2004, **5**:R80, [http://genomebiology.com/2004/5/10/R80].

14. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM: **Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer**. *Science* 2005, **310**:644–648.

15. van Wieringen W, van de Wiel M, van der Vaart A: **A Test for Partial Differential Expression**. Tech. Rep. WS2006-4, Department of Mathematics, Vrije Universiteit 2006.

16. Tibshirani R, Hastie T: **Outlier sums for differential gene expression analysis**. *Biostatistics* 2007, **8**:2–8.

17. Klebanov L, Yakovlev A: **Treating Expression Levels of Different Genes as a Sample in Microarray Data Analysis: Is it Worth a Risk?** *Statistical Applications in Genetics and Molecular Biology* 2006, **5**:Article 9.

18. Pawitan Y, Calza S, Ploner A: **Estimation of the false discovery proportion under general dependence**. *Bioinformatics* 2006, **22**(24):3025–3031.

19. Anton H: *Elementary Linear Algebra*. Wiley, 6 edition 1991.

Figures

Figure 1 - Density plots

Distribution of transformed expression values, **Y**, for the different arrays, in the two datasets. Colour-coding according to sample variance is used for increased clarity (blue for low variance, red for high variance). Differences in variability can be noted for both datasets.

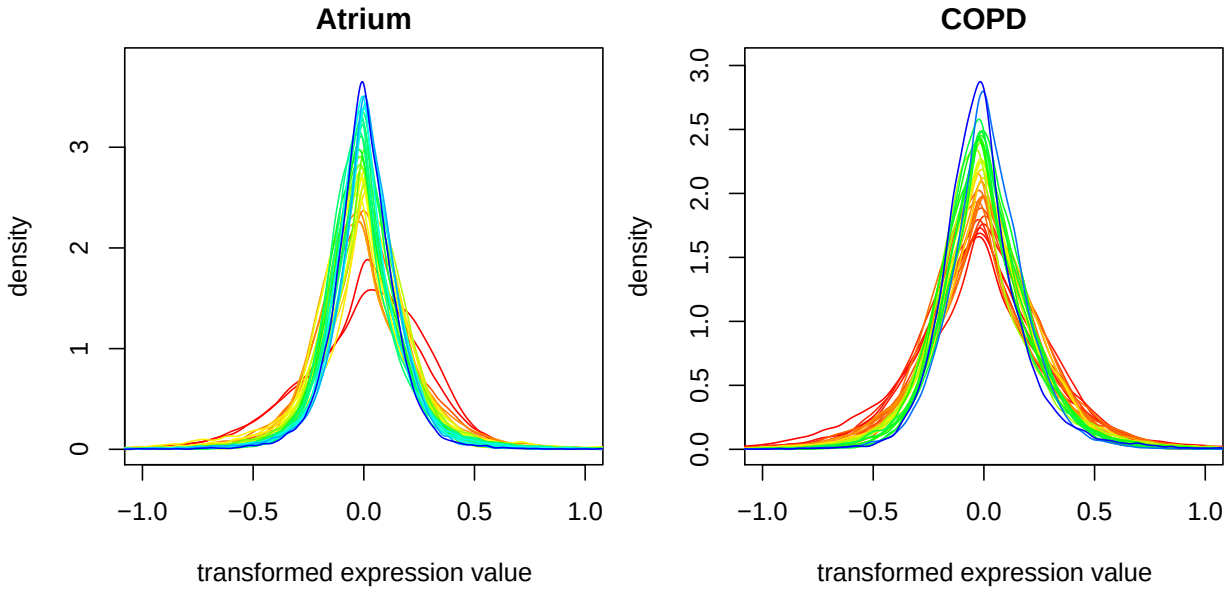


Figure 2 - Pairwise plots

Transformed expression values, $\mathbf{Y}_g$, for selected pairs of arrays within the same group. Different pairs within the same group have distinctly different correlations. Upper triangle contains scatterplots. Lower triangle contains heatmaps of the corresponding two-dimensional kernel density estimates, where the majority of the genes are in the red portion of the plot, revealing important trends inside the black clouds. Off-diagonal numbers show estimated correlations from WAME. Diagonal boxes contain sample names and weights as well as estimated variances from WAME.

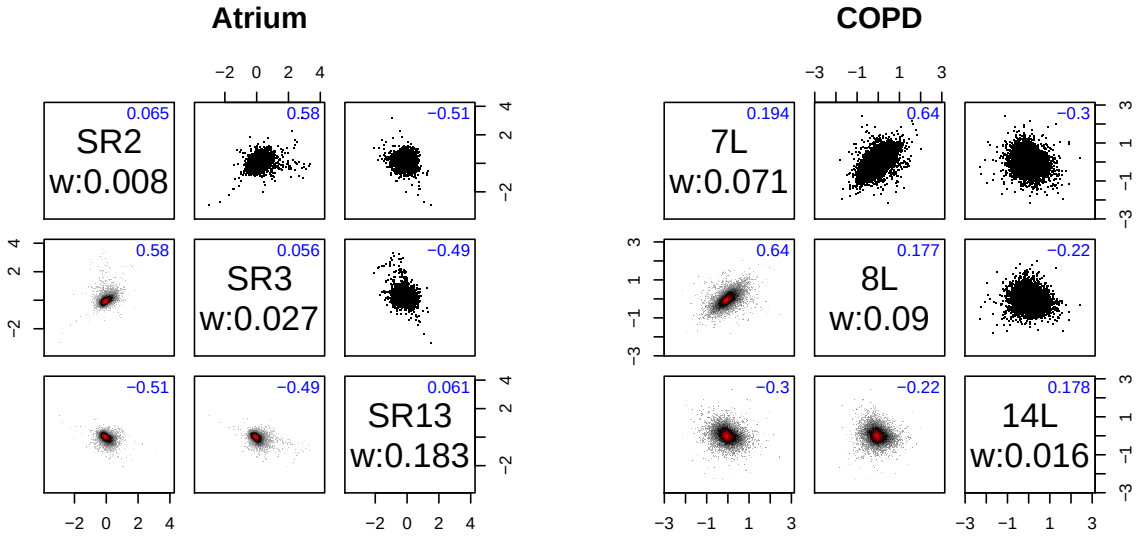


Figure 3 - Observed probability plots

Empirical distribution of p-values compared to the distribution expected for non-differentially expressed genes. The OLM curve is largely hidden by the LIMMA curve.

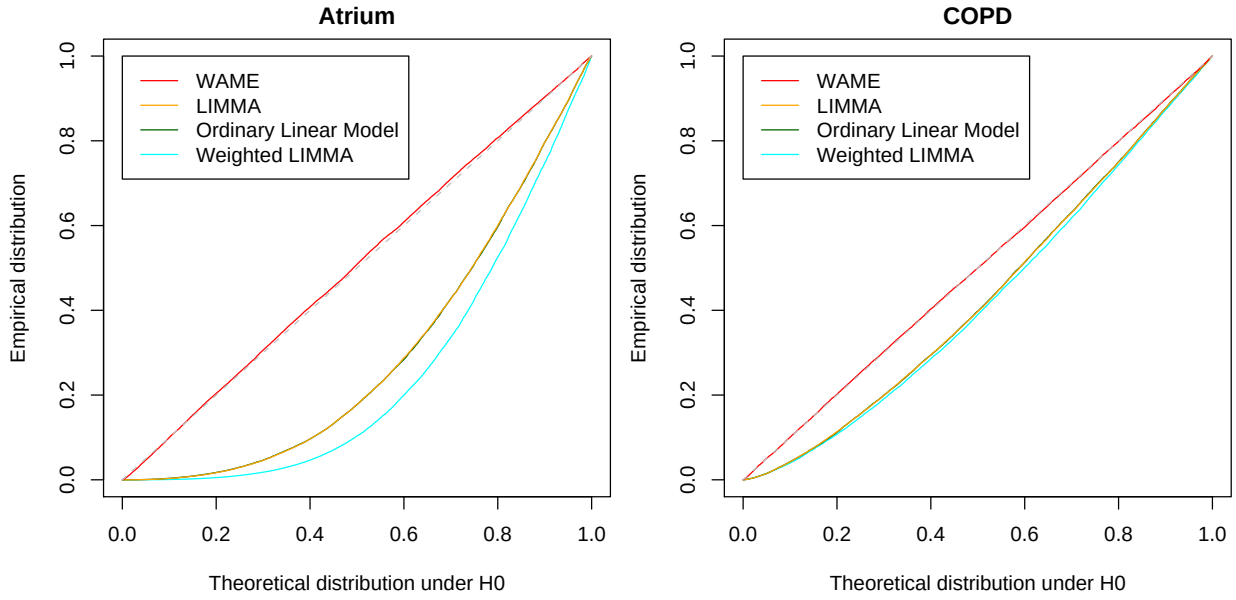


Figure 4 - Probability plots

Empirical distributions of p-values for LIMMA, weighted LIMMA, OLM and WAME from tests on 100 resamples from the COPD dataset. Average empirical distribution indicated. Since no signal is added, the curves should ideally follow the diagonal.

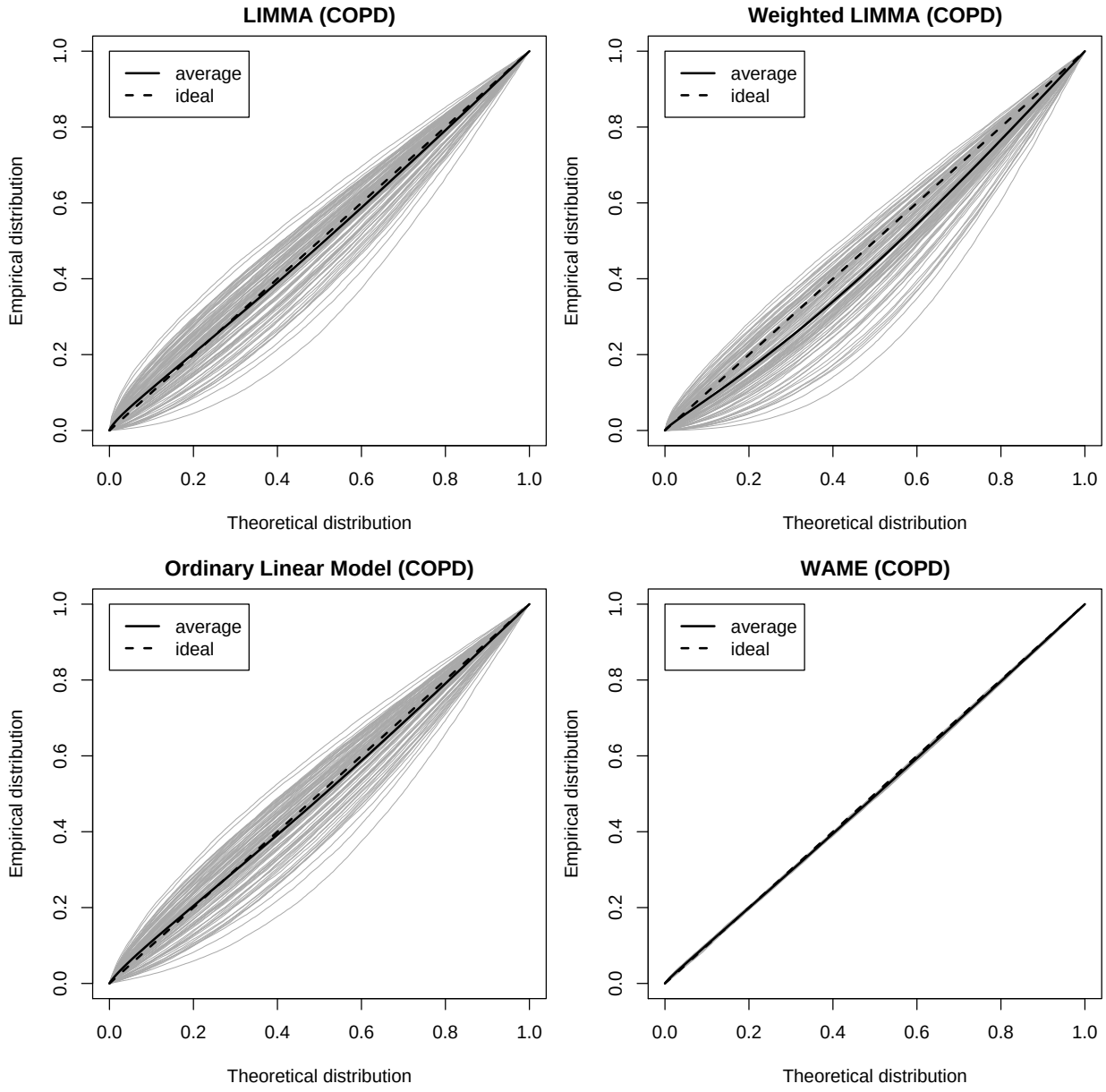


Figure 5 - Average empirical p-value distribution for WAME under regulation

Average empirical p-value distribution of the unregulated genes, calculated using WAME, when 0%, 0.1%, 1%, 5% and 10% of the genes have a $\log_2$ differential expression of 1, i.e. a two-fold change. When genes are regulated the estimate of Σ is biased, leading to conservative, non-diagonal curves.

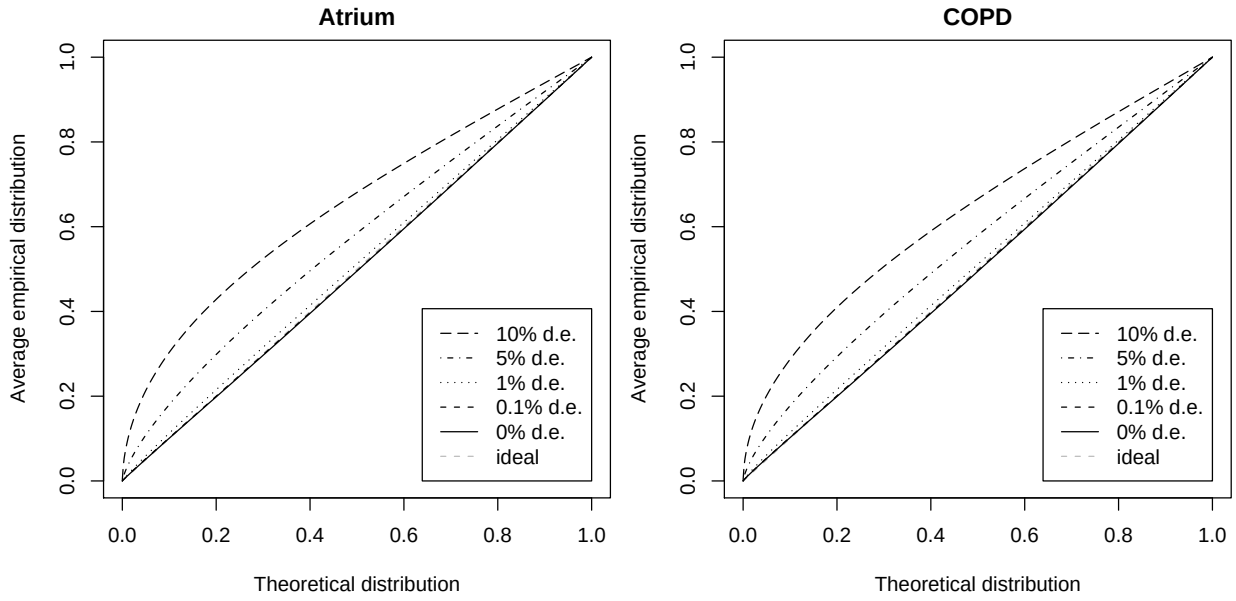
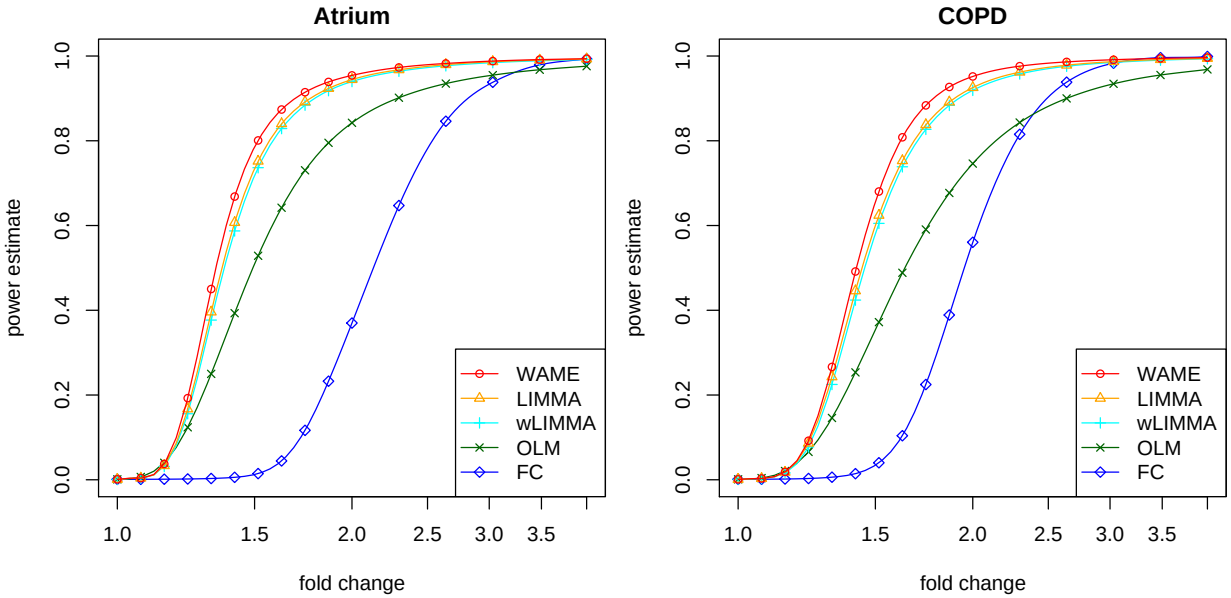


Figure 6 - Estimated power

Estimated power in the simulated data for level 0.1% tests, based on resamples from the respective larger group in the Atrium and COPD datasets. Power is estimated at the marked points and spline interpolation is used in between.



Tables

Table 1 - WAME weights

Weights in percent from estimate of differential expression using WAME.

Atrium

Sinus rythm							Atrial fibrillation			
3.0	-0.8	-2.7	-1.9	-4.6	-0.7	14.9	8.5	21.0	12.2	10.7
-9.4	1.9	-5.1	0.3	-5.2	-18.3	7.5	16.6	2.1	11.8	5.3
-10.6	-8.9	-9.9	-19.8	-9.4	-20.4		6.5	5.2		

COPD

No/mild emphysema				Severe emphysema					
-18.0	-6.7	-3.9	-8.9	11.8	2.6	12.0	4.0	12.6	7.6
-10.6	-7.3	-8.0	-5.6	7.1	9.0	6.7	0.9	6.2	5.5
-8.3	-3.6	-14.9	-4.3	-0.3	1.6	3.2	7.6	4.3	-2.5

Table 2 - Concordance of top lists

Number of mutually included genes in the top-100 lists as determined by the different methods.

Atrium	WAME	LIMMA	wLIMMA	OLM	FC
WAME	100	47	45	44	15
LIMMA	47	100	80	88	26
wLIMMA	45	80	100	76	21
OLM	44	88	76	100	21
FC	15	26	21	21	100

COPD	WAME	LIMMA	wLIMMA	OLM	FC
WAME	100	46	47	41	22
LIMMA	46	100	77	78	35
wLIMMA	47	77	100	66	32
OLM	41	78	66	100	25
FC	22	35	32	25	100

Paper IV

Separation of human adipocytes by size: hypertrophic fat cells display distinct gene expression

Margareta Jernås,* Jenny Palming,* Kajsa Sjöholm,* Eva Jennische,† Per-Arne Svensson,* Britt G. Gabrielsson,* Max Levin,‡ Anders Sjögren,§ Mats Rudemo,§ Theodore C. Lystig,* Björn Carlsson,* Lena M. S. Carlsson,* and Malin Lönn*,[1]

*Research Centre for Endocrinology and Metabolism, Division of Body Composition and Metabolism, Department of Internal Medicine, †Institute of Anatomy and Cell Biology, and ‡Cardiovascular Institute and Wallenberg Laboratory, Sahlgrenska Academy at Göteborg University, Göteborg, Sweden; and §Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden

ABSTRACT **Enlarged adipocytes are associated with insulin resistance and are an independent predictor of type 2 diabetes. To understand the molecular link between these diseases and adipocyte hypertrophy, we developed a technique to separate human adipocytes from an adipose tissue sample into populations of small cells (mean 57.6±3.54 μm) and large cells (mean 100.1±3.94 μm). Microarray analysis of the cell populations separated from adipose tissue from three subjects identified 14 genes, of which five immune-related, with more than fourfold higher expression in large cells than small cells. Two of these genes were serum amyloid A (SAA) and transmembrane 4 L six family member 1 (TM4SF1). Real-time RT-PCR analysis of SAA and TM4SF1 expression in adipocytes from seven subjects revealed 19-fold and 22-fold higher expression in the large cells, respectively, and a correlation between adipocyte size and both SAA and TM4SF1 expression. The results were verified using immunohistochemistry. In comparison with 17 other human tissues and cell types by microarray, large adipocytes displayed by far the highest SAA and TM4SF1 expression. Thus, we have identified genes with markedly higher expression in large, compared with small, human adipocytes. These genes may link hypertrophic obesity to insulin resistance/type 2 diabetes.—Jernås, M., Palming, J., Sjöholm, K., Jennische, E., Svensson, P.-A., Gabrielsson, B. G., Levin, M., Sjögren, A., Rudemo, M., Lystig, T. C., Carlsson, B., Carlsson, L. M. S., Lönn, M. Separation of human adipocytes by size: hypertrophic fat cells display distinct gene expression. *FASEB J.* 20, E832–E839 (2006)**

Key Words: cell size · serum amyloid a · transmembrane 4 L six family member 1 · leptin · insulin resistance

Obesity-related disease is the leading cause of death in the industrialized world. Abdominal obesity in particular increases the risk of several metabolic disorders, including type 2 diabetes and cardiovascular disease (1). Obesity and type 2 diabetes both have features of acute-phase activation and low-grade inflammation (2, 3). Adipocytes produce a number of cytokines and other bioactive molecules, together termed adipokines (4). Some act predominantly in an autocrine or paracrine manner, while others are released into the systemic circulation and act as signaling molecules in other tissues. Therefore, the production and secretion of bioactive molecules by adipocytes may underlie many components of obesity-related disease.

The risk of metabolic complication is increased not only by the amount and location of adipose tissue, but also by the size of the fat cells. Human fat cells can change ~20-fold in diameter and several thousand-fold in volume. Enlargement of subcutaneous (s.c.) abdominal adipocytes is associated with insulin resistance and is an independent predictor of type 2 diabetes (5). Lipid mobilization and glucose metabolism are increased in enlarged adipocytes (6). In contrast, the stimulating effect of insulin on the rate of glucose metabolism is inversely related to the size of the fat cell (7, 8). Cytokine release within adipose tissue also appears to be correlated with adipocyte size (9–12), and hypertrophic adipocytes may contribute to lipotoxicity (13).

In most studies of the impact of adipocyte size, fat cells or biopsies with different mean adipocyte diameters were obtained from different tissue locations or even from different donors (6–10). Therefore, differences in environmental conditions or genetic factors that affect adipocyte gene expression and metabolism could not be excluded. Thus, it is not clear whether the functions of the fat cell vary with adipocyte size *per se*.

In 1972, a procedure was developed to separate human adipocytes of different sizes from the same adipose tissue sample by exploiting differences in the flotation rates of large and small fat cells. Analysis of small and large adipocytes obtained by this rather

[1] Correspondence: RCEM/Division of Body Composition and Metabolism, Department of Internal Medicine, Vita Stråket 15, SE 413 45 Göteborg, Sweden. E-mail: malin.lonn@medic.gu.se

doi: 10.1096/fj.05-5678fje

0892-6638/06/0020-0832 © FASEB

complicated procedure, which required vertical dialysis tubes several meters long, showed that triglyceride turnover increases with cell size (14). In studies of rat adipocytes from a single tissue sample, also separated by size using techniques based on cell buoyancy, various metabolic functions including leptin gene expression were dependent on cell size (15, 16).

The aim of the present study was to detect factors linking human adipocyte hypertrophy to insulin resistance/type 2 diabetes. We developed a simple and accurate procedure for separating human adipocytes from an adipose tissue sample into two populations: large cells and small cells. Using a computerized image-analysis technique that allows the assessment of 10-fold more cells than conventional methods (17), we characterized the size distribution of these populations in detail. Gene expression profiles were analyzed to identify genes differentially expressed in small and large adipocytes. Classification of the genes with markedly higher expression in large compared with small cells revealed that the majority were immune-related, with importance for cell structure, or with unknown function. These genes may connect hypertrophic obesity to metabolic disorders.

MATERIALS AND METHODS

Human adipose tissue samples

Subcutaneous abdominal adipose tissue from 12 subjects, 3 men and 9 women (2 postmenopausal), was obtained after an overnight fast. The subjects were 24–57 yr of age (mean 39.8±3.1 yr) and had BMIs of 23.0–28.7 kg/m^2 (mean 25.4±0.6 kg/m^2). Surgical biopsies were taken from 8 patients undergoing abdominal surgery for nonmalignant conditions and from 1 healthy volunteer. Needle aspirations were obtained from 3 healthy volunteers. One patient had type 2 diabetes. The study protocol was approved by the Regional Ethical Review Board in Göteborg, and all participants gave written informed consent.

Adipocyte isolation and separation

The tissue (4–52 g) was cut into small pieces and treated with 1.05 mg/ml collagenase (Type A, Roche, Mannheim, Germany) in minimum essential medium (Invitrogen, Carlsbad, CA) containing 5.5 mM glucose, 25 mM HEPES, 4% bovine albumin (Fraction V, Sigma, St. Louis, MO), and 0.15 μM adenosine, pH 7.4, for 60 min at 37°C as described (18). After filtration through a 250 μm nylon mesh, the adipocytes were washed three times and suspended in fresh medium (cells from ~1 g tissue/45 ml medium) in 50-ml Falcon tubes. After gentle agitation of the suspension, cells that resurfaced within 30 s were transferred to new tubes; this procedure was repeated once. These more buoyant cells were then filtered with a 70 μm nylon mesh and rinsed with fresh medium. Cells not passing through the mesh were resuspended in medium as the final preparation of large adipocytes. The denser cells that did not resurface within 30 s were filtered with a 50 μm nylon mesh. Cells that passed though the mesh were considered the final preparation of small adipocytes. The medium and the adipocyte suspensions were maintained at 37°C.

Adipocyte size

The mean adipocyte size and the size distribution of the cell populations, before and after separation, were determined by computerized image analysis (KS400 software, Carl Zeiss, Oberkochen, Germany) (17). In brief, the cell suspension was placed between a siliconized glass slide and a coverslip and transferred to the microscope stage. Nine random visual fields were photographed with a CCD camera (Axiocam, Carl Zeiss, Oberkochen, Germany). The surface of the relevant areas was measured automatically, and the diameter of the corresponding circles was calculated. Uniform microspheres 98.00 μm in diameter (Bangs Laboratories, Fishers, IN) were used as a reference. Because of technical problems, the adipocyte size of the cell populations in one of the 12 separations was determined by a conventional method (18).

RNA preparation

Total RNA was prepared with the phenol-chloroform extraction method of Chomczynski and Sacchi (19) and the RNeasy lipid tissue kit (Qiagen, Chatsworth, CA). After further purification with RNeasy clean-up columns, the RNA concentration was measured spectrophotometrically; the A_{260}/A_{280} ratio was 1.8–2.0. The quality of the RNA was verified by agarose gel electrophoresis before reverse transcription into cDNA.

Microarray analysis

cRNA was prepared and hybridized as recommended in the Affymetrix Gene Chip Expression Analysis manual. In brief, biotin-labeled target cRNA was prepared by *in vitro* transcription (Enzo Diagnostics, Farmingdale, NY) and hybridized to Human Genome U133A arrays (Affymetrix, Santa Clara, CA), composed of 22,283 probe sets representing ~14,000 expressed genes. Arrays were scanned with a confocal laser scanner (GeneArray scanner G2500A Hewlett Packard, Palo Alto, CA). Gene expression levels were calculated by the Robust Multiarray Average (RMA) method (20). To identify differences in gene expression between small and large adipocytes, Weighted Analysis of paired Microarray Experiments (WAME) was used (21), weighting samples according to precision in calculation of (geometric) signal means and *P* values for differential gene expression.

Tissue expression analysis

Gene expression in different human tissues and cell types was assessed with DNA microarray. Duplicate GeneChip HG-U133A expression profiles from 17 different tissues were downloaded from the SymAtlas dataset (22). The expression profile for each tissue was obtained by calculating the average signal value for each duplicate and gene, respectively. For comparison of gene expression in different tissues, the signal value for each gene was normalized by dividing the signal by the average signal of the entire array for each tissue. In addition, our own expression profiles, originating from small and large adipocytes, were included and normalized as outlined above.

RT-PCR analysis of gene expression

Reagents for real-time RT-PCR analysis of LDL receptor–related protein 10 (LRP10), TM4SF1 and leptin (Assays-on-Demand, TaqMan Reverse Transcriptase reagents, and Taq-

Man Universal PCR Master Mix) were from Applied Biosystems (Foster City, CA) and used according to the manufacturer's protocol. Since SAA1 and SAA2 have highly similar sequences and cannot be studied separately, probe and primer sequences for SAA (Cybergene, Huddinge, Sweden) were designed to span an exon-exon boundary to avoid genomic DNA amplification and to detect both isoforms (23) (sequences available on request). cDNA was synthesized from 500 ng of total RNA in a total reaction vol of 50 μl. cDNA corresponding to 10 ng of RNA per reaction was used for real-time PCR amplification. Specific products were amplified and detected with the ABI Prism 7900HT Sequence Detection System (Applied Biosystems) using default cycle parameters. A standard curve was plotted for each primer-probe set with a serial dilution of pooled adipocyte cDNA. Based on our previous report (24) and expression profiles in the present study, human LRP10 was used as reference to normalize the expression levels between samples. All standards and samples were analyzed in triplicate.

Immunohistochemistry

Adipocytes were fixed in 4% buffered formaldehyde, embedded in agar, dehydrated, embedded in paraffin, and cut into 5 μm-thick sections. The sections were incubated first with monoclonal antibodies against SAA (HyCult Biotechnology) or TM4SF1 (L6 tumor antigen, Chemicon International). As secondary reagent for SAA, alkaline phosphatase antialkaline phosphatase (Dako Cytomation, Glostrup, Denmark), followed by NBT/BCIP (Roche) as substrate was used. After counterstaining with Nuclear Fast Red, the sections were mounted in glycerol gelatin. For TM4SF1, peroxidase antiperoxidase (Dako Cytomation, Glostrup, Denmark), followed by diaminobenzidine as substrate was used before mounting the sections in glycerol gelatin.

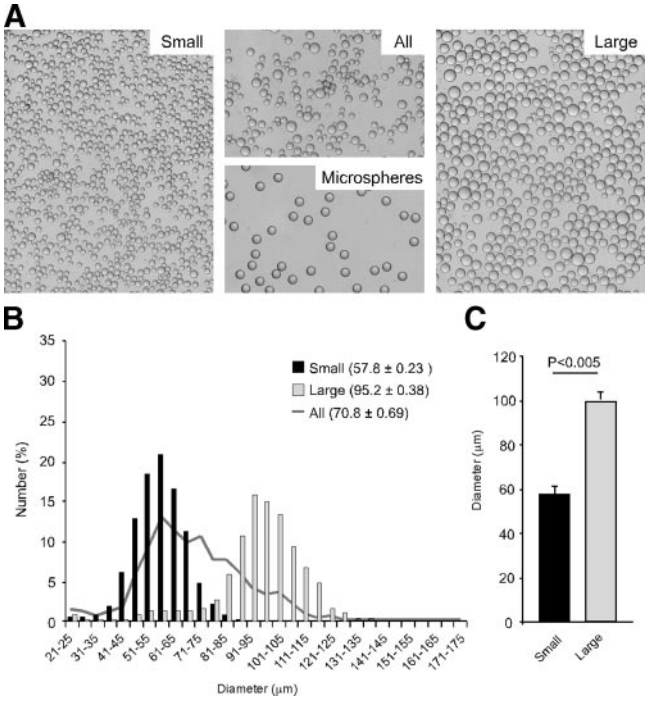


Figure 1. Separation of human adipocytes by size. *A)* Representative images of human adipocytes from one adipose tissue sample before separation (All) and after separation (Small and Large). Microspheres (98.00 μm in diameter) were used for reference in the computerized image analysis. *B)* Size distributions, determined by computerized image analysis, and mean diameters in random samples of all adipocytes (n=742), small adipocytes (n=1965), and large adipocytes (n=2382) from one adipose tissue sample. Patient characteristics; man, age: 41 yr, body mass index (BMI): 26.0 kg/m^2. *C)* Mean diameter of small and large adipocytes from 12 adipose tissue biopsies. Error bars indicate SEM.

Statistical analysis

Values are expressed as means ± SEM Differences in gene expression between cell populations were analyzed with the Wilcoxon signed-rank test. Differences in adipocyte size distributions were analyzed with the Kolmogorov-Smirnov two-sample test (25). Relationships between gene expression and adipocyte size were analyzed with the Spearman rank correlation test and plotted with a robust regression technique, an M-estimator with Huber's psi-function (26). Bonferroni correction of P values from the microarray analysis was used to control for multiple comparisons.

RESULTS

Separation of small and large adipocytes

The mean size and the size distribution of adipocytes were determined by computerized image analysis before and after separation of isolated adipocytes into populations of small and large cells (**Fig. 1*A*, *B***). In each population, 305–3660 cells (mean 1046±135) were analyzed. The mean diameters of small and large adipocytes from 12 adipose tissue samples were 57.6 ± 3.54 μm and 100.1 ± 3.94 μm, respectively (Fig. 1*C*). The mean diameter of the reference microspheres was 97.91 ± 0.089 μm (range 97.18–98.22 μm), n = 12. The cell size distributions of the small and large populations differed significantly (P<0.001).

Microarray analysis of gene expression in small and large adipocytes

To identify genes with increased expression in large adipocytes, we performed DNA microarray analyses of small and large cells from three biopsies. Fourteen genes were expressed at > 4-fold higher levels in the large cells with P values that after Bonferroni correction (multiplication with 22,283) were < 0.01 (Table 1). Classification by cellular or organism function based on Gene Ontology definitions (http://www.geneontology.org/) revealed that five of those genes were immune-related. The remaining nine genes were referred to structure (four), unknown function (three), growth (one) and transport (one) (Table 1). Differences in sample preparation or hybridization were excluded since there was no difference between small and large adipocytes in the expression of LRP10, CLN3, or COBRA1, previously identified as suitable reference genes for studies of human adipose tissue (24).

One immune-related gene, SAA, and one gene with unknown function, TM4SF1, were selected among the genes with > 4-fold higher expression in large vs. small

TABLE 1. *Genes detected with more than fourfold higher expression in large adipocytes than small adipocytes as analyzed by DNA microarray*

Gene symbol	Gene name	ID	Classification	Fold change	*P* value	Mean signal small	Mean signal large
SELE	selectin E	206211_at	defense	15.2	1.29E-09	11.5	174.4
SPARCL1	SPARC-like 1	200795_at	unknown	14.9	1.22E-09	33.5	498.8
TM4SF1	transmembrane 4 L six family member 1	209386_at	unknown	11.6	4.46E-11	50.6	586.3
TM4SF1	transmembrane 4 L six family member 1	209387_s_at		9.6	4.19E-09	22.5	216.1
TM4SF1	transmembrane 4 L six family member 1	215034_s_at		9.2	2.29E-10	26.7	245.5
DCN	decorin	211896_s_at	structure	9.4	8.18E-09	59.3	554.7
DCN	decorin	201893_x_at		6.9	3.70E-09	156.8	1076.5
DCN	decorin	211813_x_at		5.6	1.90E-09	118.4	659.1
IL8	interleukin 8	202859_x_at	defense	9.0	4.44E-09	186.8	1682.5
IL8	interleukin 8	211506_s_at		7.5	3.54E-07	55.5	415.2
PALLD	palladin	200897_s_at	structure	8.7	2.53E-09	24.6	213.8
SAA2	serum amyloid A2	208607_s_at	defense	7.8	3.82E-08	245.7	1918.5
SAA2	serum amyloid A2	214456_x_at		7.6	2.24E-07	529.1	4009.6
CLEC3B	C-type lectin domain family 3, member B	205200_at	growth	6.3	8.22E-09	55.8	353.5
C1QR1	complement component 1, q subcomponent, receptor 1	202878_s_at	defense	6.1	9.40E-10	30.0	183.9
COL1A1	collagen, type I, alpha 1	202310_s_at	structure	5.9	2.27E-08	26.2	155.6
CXCL2	chemokine (C-X-C motif) ligand 2	209774_x_at	defense	5.4	8.09E-09	184.4	998.2
COL1A2	collagen, type I, alpha 2	202403_s_at	structure	5.3	2.44E-07	57.1	300.9
FLJ14054	—	219054_at	unknown	4.2	4.65E-09	65.2	273.6
AQP1	aquaporin 1	209047_at	transport	4.2	5.90E-09	43.2	179.7

cells for further analysis (Table 1). SAA, a risk factor for cardiovascular disease and in focus in one of our previous studies (27, 23), was expressed at ~8-fold higher levels in large cells than in small cells. TM4SF1, a highly expressed surface protein of various carcinomas and possibly implicated in signal transduction events mediating cell proliferation and activation (28, 29) was 10-fold higher expressed in the large adipocytes (Table 1). Leptin, being 3-fold higher expressed in the large adipocytes (data not shown), and previously identified with higher expression in large compared with small rat adipocytes (15, 16), was also included in the following studies.

Expression of SAA, TM4SF1 and leptin determined by real-time RT-PCR

The up-regulation of SAA, TM4SF1 and leptin expression in large cells was confirmed by RT-PCR analysis of small (mean 59.3±4.47 μm) and large (mean 97.1±5.69 μm) adipocytes from seven different adipose tissue samples. In all cases, SAA, TM4SF1 and leptin were expressed at higher levels in large cells (P=0.018) (**Fig. 2*A, C, E***). The mean fold increase in expression was 18.7 ± 15.1 for SAA, 22.3 ± 6.4 for TM4SF1, and 3.9 ± 1.4 for leptin. In addition, adipocyte size correlated with the expression of SAA (P=0.015), TM4SF1 (P=0.012), and leptin (P=0.0009) (Fig. 2*B, D* and *F*).

Tissue expression analysis

SAA, TM4SF1 and leptin expression levels in large adipocytes were compared to the levels in other human tissues (22) and small adipocytes. The three genes were expressed at markedly higher levels in large adipocytes than in all other tissues/cell types (**Fig. 3*A, B C***).

Immunoreactivity

Expression of SAA and TM4SF1 in adipocytes was also demonstrated immunohistochemically. Although SAA immunoreactivity varied between cells of the same size, it was generally greater in large than in small adipocytes in the same sample (**Fig. 4**, *upper*). TM4SF1 immunoreactivity was demonstrated mainly in large but to some extent also in medium-sized adipocytes. Positive TM4SF1 signal appeared in a dot-like pattern in the cell membrane. Small adipocytes were completely without TM4SF1 immunoreactivity (Fig. 4, *lower*).

DISCUSSION

In this study, we developed a new technique to separate populations of small and large human adipocytes from a single adipose tissue sample. The two populations of cells obtained with our technique differed significantly in size, as determined with a computer-based image-analysis method that allows rapid analysis of 10-fold more cells than conventional methods (17). DNA microarray analysis of the two populations showed that several genes were expressed at markedly higher levels in the large cells, demonstrating that hypertrophy *per se* can significantly alter gene expression and thereby presumably adipocyte function.

Previous studies of the metabolic activity of small and large adipocytes have indicated that adipocyte size influences several metabolic functions (6–10). How-

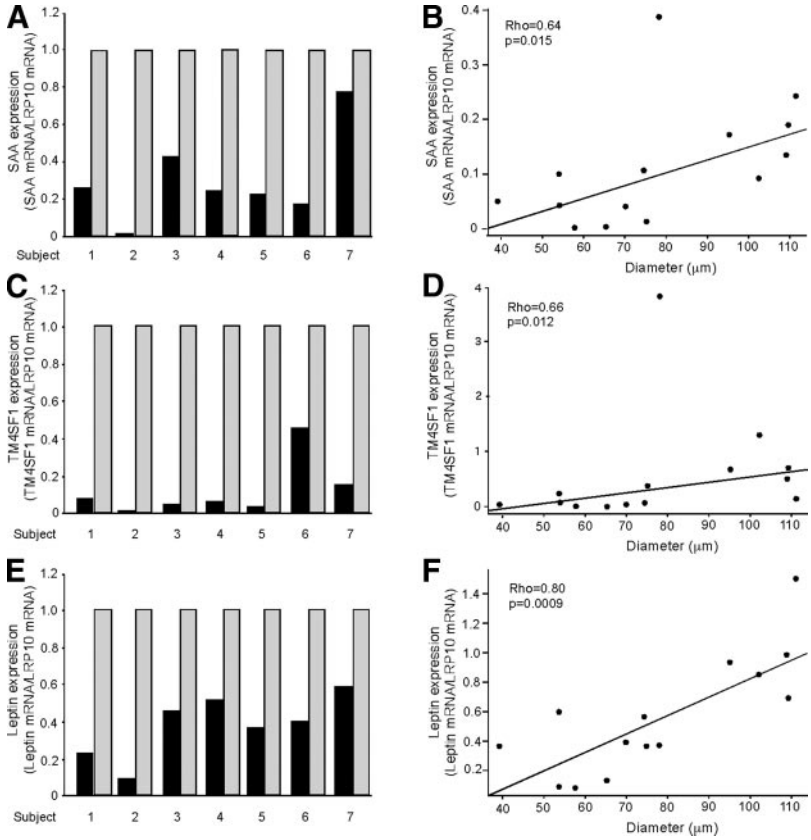


Figure 2. Expression of serum amyloid A (SAA), transmembrane 4 L six family member 1 (TM4SF1), and leptin in small adipocytes (solid bars) and large adipocytes (shaded bars) isolated from adipose tissue samples from seven subjects, measured by RT-PCR. The expression of SAA, TM4SF1, and leptin in the large cells from each sample was set to 1.0. In all cases, SAA (*A*), TM4SF1 (*C*), and leptin (*E*) were expressed at higher levels in the large cells. Robust regression and rank correlation for SAA expression and adipocyte size (*B*), for TM4SF1 expression and adipocyte size (*D*), and for leptin expression and adipocyte size (*F*). Rho; Spearman's correlation coefficient.

ever, the cells of different sizes were obtained from different adipose tissue locations or from different donors. Thus, it was not possible to exclude environmental influences, such as nutritional/hormonal conditions, or genetic factors that might have affected gene expression and thus adipocyte metabolism. Our technique avoids these problems and will facilitate metabolic studies of fat cells of different sizes. For example, a positive correlation between human adipocyte size and leptin expression/secretion has previously been suggested after analysis of adipocytes from two depots: omental fat cells and twice as large s.c. fat cells (9). In the present study of human adipocytes from a single adipose tissue sample, the previous findings were confirmed since leptin was indeed expressed at higher levels in the large cells in all cases. Moreover, we identified several genes that, compared with leptin, showed a more pronounced differential expression in large vs. small adipocytes.

Among the fourteen genes with markedly higher expression in large compared with small adipocytes, five were classified as immune-related; E-selectin, interleukin (IL)-8, SAA, C1q receptor 1, and CXCL2 also known as MIP-2 or macrophage inflammatory protein-2. Components of the metabolic syndrome, such as obesity and type 2 diabetes, are associated with a systemic increase in inflammatory markers (30–32). The acute-phase proteins SAA and C-reactive protein have attracted particular attention because they are independent risk factors for coronary artery disease (27, 33, 34). We (23), and others (35), have shown that adipose tissue is a major site of SAA production and is likely to be a major source of circulating SAA in obese patients. Moreover, serum SAA concentrations were correlated to fasting insulin levels indicating a link to insulin resistance (23). The current study extends previous findings by demonstrating that SAA is expressed at the highest concentration by the large adipocytes. SAA has been implicated in inflammation, insulin resistance and impairment of reverse cholesterol transport. Our data may therefore suggest that adipocyte-derived SAA, likely having both local effects and endocrine functions, is a potential mediator of the link between hypertrophic adipocytes and type 2 diabetes.

Further support for an association between obesity and inflammation comes from studies showing macrophage infiltration of adipose tissue (36). Moreover, the proportion of cells expressing CD68, a macrophage marker, increases with increasing average adipocyte area in human s.c. adipose tissue (37). This relationship between adipocyte size and macrophage accumulation suggests that hypertrophic adipocytes secrete factors that attract monocytes. Again, SAA derived from hypertrophic adipocytes may be involved in this process because SAA activates the chemotactic formyl peptide receptor like-1, which results in migration of blood monocytes and neutrophils (38). IL-8, also found to be expressed at high levels in hypertrophic adipocytes, may act as an other potential monocyte recruiting factor in adipose tissue (39). Accumulation of macrophages in adipose tissue is likely to further increase the levels of inflammatory cytokines in adipose tissue and thereby increase insulin resistance.

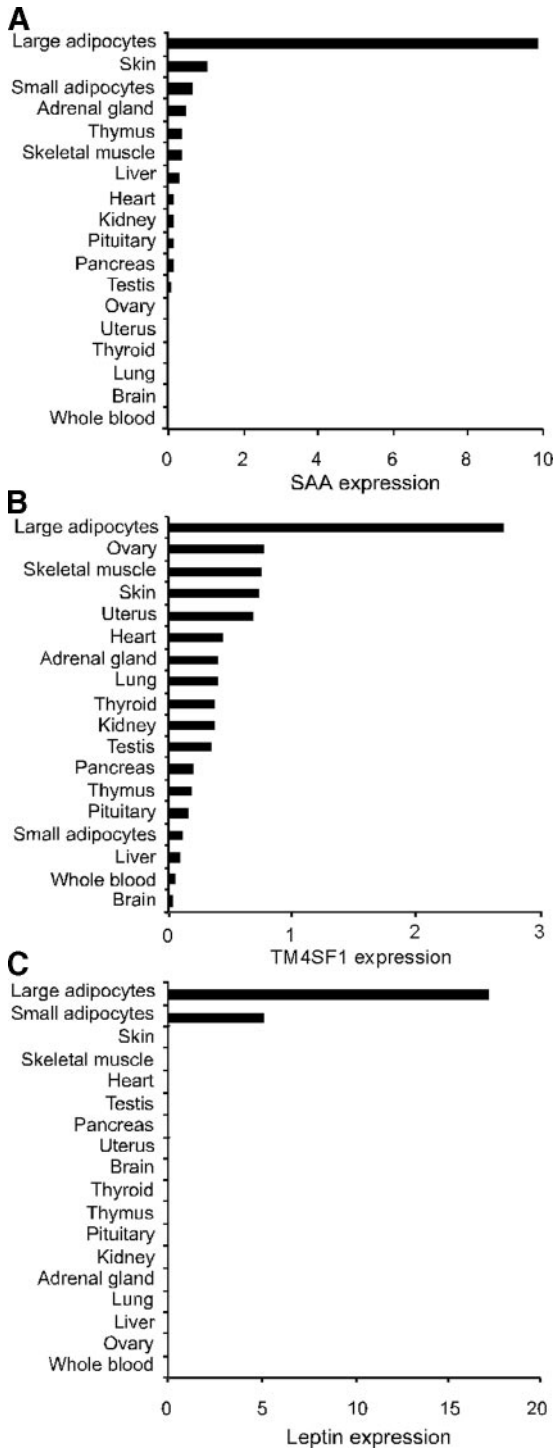


Figure 3. Expression of SAA (*A*), TM4SF1 (*B*), and leptin (*C*) in large adipocytes and other human tissues/cell types, measured by microarray analysis. The relative expression levels of SAA, TM4SF1, and leptin in each tissue were determined as described in Materials and Methods.

SAA in the circulation is primarily associated with HDL and may modulate its role in cholesterol transport (40). The presence of SAA on HDL has been reported both to promote (41), and to reduce (42), cholesterol efflux to HDL suggesting that SAA may alter cholesterol removal from cells. Moreover, SAA reduces cellular uptake of cholesterol from HDL by inhibiting the HDL receptor, SR-BI (scavenger receptor class B type I) (43). Such an effect would be expected to impair reverse cholesterol transport—the transfer of cholesterol from peripheral tissues (vascular wall) to the liver for excretion. Non-HDL-bound SAA has been reported to promote cholesterol efflux from cells (44).

Recent studies suggest an intricate relationship between adipocyte size, cholesterol concentration and insulin sensitivity (45). Cholesterol accumulates within the adipocyte lipid droplet proportionally to the triglyceride content. However, within an adipose tissue sample, large adipocytes have reduced membrane cholesterol concentrations compared with small fat cells, demonstrating that a changed cholesterol distribution is characteristic of adipocyte hypertrophy. By reducing adipocyte plasma membrane cholesterol it is possible to reproduce part of the defects seen in hypertrophic adipocytes such as insulin resistance (45). SAA may impair adipocyte uptake of HDL cholesterol by the same mechanisms described above for reverse cholesterol transport. Furthermore, locally produced lipid-

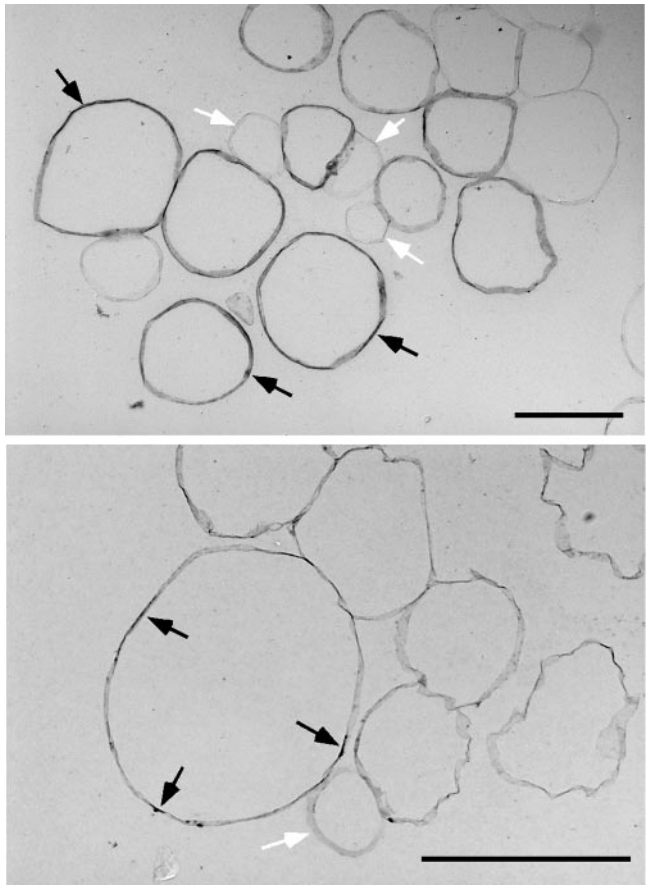


Figure 4. SAA immunoreactivity of human adipocytes in a paraffin section. Positive signal appears dark gray. Black arrows indicate large adipocytes with strong SAA signal; white arrows indicate small adipocytes with weak SAA immunoreactivity that are barely visible. Scale Bar = 100 μm. (*top*). TM4SF1 immunoreactivity of human adipocytes in a paraffin section. Positive signal appears dark gray and in a dotlike pattern as indicated by black arrows. White arrow indicates small adipocyte completely without immunoreactivity. Scale Bar = 100 μm. (*bottom*).

free SAA may reduce adipocyte levels of cholesterol (44). Our findings may therefore suggest that the insulin resistance in hypertrophic adipocytes could be related to the high expression of SAA in these cells. SAA has also been shown to bind to cell-surface receptors, including Tanis/SelS, that have been implicated in regulation of insulin sensitivity (46). Thus, SAA may influence insulin sensitivity in adipose tissue by several mechanisms, including interactions with receptors, recruitment of inflammatory cells and local impairment of cholesterol metabolism.

The significance of the increased expression of TM4SF1 in large adipocytes remains to be elucidated. TM4SF1 shares topological features with members of the tetraspanin superfamily (29). However, recent analysis suggests that TM4SF1 belongs to a new transmembrane-4 superfamily comprising TM4SF1, TM4SF5, IL-TMP and L6D (28). The biological properties of this superfamily remain largely unknown while the activities of tetraspanin proteins often are attributed to their association with integrins and specific surface proteins (29). We have previously identified TM4SF1 as one of the genes with higher expression in visceral compared to s.c. adipose tissue in obese men (47). Moreover, genomic scans have identified loci linked to fasting plasma insulin levels in Pima Indians (48) and BMI in white, black, Asian and Mexican-American ethnic groups (49) in the chromosomal region harboring the TM4SF1 gene. We have analyzed a publicly available DNA microarray dataset originating from isolated abdominal s.c. adipocytes from obese and nonobese Pima Indians (50). In comparison with the lean group, the obese group had higher fasting glucose and fasting insulin levels and higher 2-h insulin levels after an oral glucose tolerance test. Interestingly, the TM4SF1 expression was significantly higher in adipocytes from the obese subjects compared with adipocytes from the nonobese subjects (50). Furthermore, in a recent study, TM4SF1, together with several genes involved in inflammatory processes in the brain, was found to be upregulated in C/EBPβ-overexpressing neuronal cells (51). Thus, a potential role of TM4SF1 in immunological processes in human adipose tissue, and in subsequent development of metabolic disturbances, may be speculated.

To summarize, we have developed a technique to separate human adipocytes, from a single adipose tissue sample, by size. The resulting populations of small and large adipocytes have significantly different cell size distributions. Gene expression profiling of the small and large adipocytes identified genes, many of them immune-related, with markedly higher mRNA expression in the large cells. Two of those genes, SAA (an acute-phase protein implicated in inflammation, insulin resistance and impairment of reverse cholesterol transport) and TM4SF1 (a membrane protein with unknown function), were ~20-fold higher expressed in the large cells, a difference reflected also at the protein level. Moreover, in comparison with several other human tissues, large adipocytes displayed by far the highest SAA and TM4SF1 expression. In the light of previous studies reporting that adipocyte hypertrophy is associated with insulin resistance and is an independent predictor of type 2 diabetes, the findings in the current work provide novel insights into the molecular connection between hypertrophic obesity and insulin resistance/type 2 diabetes. FJ

We thank Lisa Södergren (Program in Medical Laboratory Science) for excellent assistance with the separation technique and members of Research Centre for Endocrinology and Metabolism II for excellent technical assistance. This work was supported by grants from the Swedish Research Council (Project No. 521-2002-6480, 529-2002-6671, 521-2002-6356, 521-2005-6736, 11285), the Swedish Diabetes Foundation, The National Board of Health and Welfare, Wilhelm and Martina Lundgren Foundation, Adlerbertska Research Foundation, Åke Wiberg Foundation, Magn. Bergvall Foundation, Fredrik and Ingrid Thuring Foundation, Tore Nilson Foundation, the Swedish federal government under the LUA/ALF agreement and Swegene.

REFERENCES

1. Kopelman, P. G. (2000) Obesity as a medical problem. *Nature* **404,** 635–643
2. Yudkin, J. S. (2003) Adipose tissue, insulin action and vascular disease: inflammatory signals. *Int. J. Obes. Relat. Metab. Disord.* **27,** S25–S28
3. Pickup, J. C. (2004) Inflammation and activated innate immunity in the pathogenesis of type 2 diabetes. *Diabetes Care* **27,** 813–823
4. Trayhurn, P., and Wood, I. S. (2004) Adipokines: inflammation and the pleiotropic role of white adipose tissue. *Br. J. Nutr.* **92,** 347–355
5. Weyer, C., Foley, J. E., Bogardus, C., Tataranni, P. A., and Pratley, R. E. (2000) Enlarged subcutaneous abdominal adipocyte size, but not obesity itself, predicts type II diabetes independent of insulin resistance. *Diabetologia* **43,** 1498–1506
6. Smith, U. (1972) Studies of human adipose tissue in culture. I. Incorporation of glucose and release of glycerol. *Anat. Rec.* **172,** 597–602
7. Salans, L. B., Knittle, J. L., and Hirsch, J. (1968) The role of adipose cell size and adipose tissue insulin sensitivity in the carbohydrate intolerance of human obesity. *J. Clin. Invest.* **47,** 153–165
8. Salans, L. B., and Dougherty, J. W. (1971) The effect of insulin upon glucose metabolism by adipose cells of different size. *J. Clin. Invest.* **50,** 1399–1410
9. Van Harmelen, V., Reynisdottir, S., Eriksson, P., Thörne, A., Hoffstedt, J., Lönnqvist, F., and Arner, P. (1998) Leptin secretion from subcutaneous and visceral adipose tissue in women. *Diabetes* **47,** 913–917
10. Zhang, Y., Guo, K. Y., Diaz, P. A., Heo, M., and Leibel, R. L. (2002) Determinants of leptin gene expression in fat depots of lean mice. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **282,** R226–R234
11. Sopasakis, V. R., Sandqvist, M., Gustafson, B., Hammarstedt, A., Schmelz, M., Yang, X., Jansson, P. A., and Smith, U. (2004) High local concentrations and effects on differentiation implicate interleukin-6 as a paracrine regulator. *Obes. Res.* **12,** 454–460
12. Winkler, G., Kiss, S., Keszthelyi, L., Sapi, Z., Ory, I., Salamon, F., Kovacs, M., Vargha, P., Szekeres, O., Speer, et al. (2003) Expression of tumor necrosis factor (TNF)-alpha protein in the subcutaneous and visceral adipose tissue in correlation with adipocyte cell volume, serum TNF-alpha, soluble serum TNF-receptor-2 concentrations and C-peptide level. *Eur. J. Endocrinol.* **149,** 129–135
13. Lelliott, C., and Vidal-Puig, A. J. (2004) Lipotoxicity, an imbalance between lipogenesis de novo and fatty acid oxidation. *Int. J. Obes. Relat. Metab. Disord.* **28,** S22–S28

14. Björntorp, P., and Sjöström, L. (1972) The composition and metabolism in vitro of adipose tissue fat cells of different sizes. *Eur. J. Clin. Invest.* **2,** 78–84
15. Farnier, C., Krief, S., Blache, M., Diot-Dupuy, F., Mory, G., Ferre, P., and Bazin, R. (2003) Adipocyte functions are modulated by cell size change: potential involvement of an integrin/ERK signalling pathway. *Int. J. Obes. Relat. Metab. Disord.* **27,** 1178–1186
16. Guo, K. Y., Halo, P., Leibel, R. L., and Zhang, Y. (2004) Effects of obesity on the relationship of leptin mRNA expression and adipocyte size in anatomically distinct fat depots in mice. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **287,** R112–R119
17. Björnheden, T., Jakubowicz, B., Levin, M., Odén, B., Edén, S., Sjöström, L., and Lönn, M. (2004) Computerized determination of adipocyte size. *Obes. Res.* **12,** 95–105
18. Smith, U., Sjöström, L., and Björntorp, P. (1972) Comparison of two methods for determining human adipose cell size. *J. Lipid Res.* **13,** 822–824
19. Chomczynski, P., and Sacchi, N. (1987) Single-step method for RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162,** 156–159
20. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003) Exploration, normalization, and summaries of high-density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264
21. Kristiansson, E., Sjögren, A., Rudemo, M., and Nerman, O. (2005) Weighted analysis of paired microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **4,** No. 1, Article 30 http://www.bepress.com/sagmb/vol4/iss1/art30
22. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 6062–6067
23. Sjöholm, K., Palming, J., Olofsson, L. E., Gummesson, A., Svensson, P. A., Lystig, T. C., Jennische, E., Brandberg, J., Torgerson, J. S., Carlsson, B., et al. (2005) A microarray search for genes predominantly expressed in human omental adipocytes: adipose tissue as a major production site of serum amyloid A. *J. Clin. Endocrinol. Metab.* **90,** 2233–2239
24. Gabrielsson, B. G., Olofsson, L. E., Sjögren, A., Jernås, M., Elander, A., Lönn, M., Rudemo, M., and Carlsson, L. M. (2005) Evaluation of reference genes for studies of gene expression in human adipose tissue. *Obes. Res.* **13,** 649–652
25. Fisher, L. D., and van Belle, G. (1993) The Kolmogorov-Smirnov two-sample test. In *Biostatistics: a Methodology for the Health Sciences*, pp. 319–321, John Wiley & Sons, Inc., New York
26. Draper, N. R., and Smith, H. (1998) *Applied Regression Analysis*, 3rd ed., p. 570, John Wiley & Sons, Inc., New York
27. Johnson, B. D., Kip, K. E., Marroquin, O. C., Ridker, P. M., Kelsey, S. F., Shaw, L. J., Pepine, C. J., Sharaf, B., Bairey Merz, C. N., Sopko, G., *et al.* (2004) Serum amyloid A as a predictor of coronary artery disease and cardiovascular outcome in women: the National Heart, Lung, and Blood Institute-Sponsored Women's Ischemia Syndrome Evaluation (WISE). *Circulation* **109,** 726–732
28. Wright, M. D., Ni, J., and Rudy, G. B. (2000) The L6 membrane proteins—a new four-transmembrane superfamily. *Protein Sci.* **9,** 1594–1600
29. Maecker, H. T., Todd, S. C., and Levy, S. (1997) The tetraspanin superfamily: molecular facilitators. *FASEB J.* **11,** 428–442
30. Tataranni, P. A., and Ortega, E. (2005) A burning question: does an adipokine-induced activation of the immune system mediate the effect of overnutrition on type 2 diabetes? *Diabetes* **54,** 917–927
31. Wellen, K. E., and Hotamisligil, G. S. (2005) Inflammation, stress, and diabetes. *J. Clin. Invest.* **115,** 1111–1119
32. Leinonen, E., Hurt-Camejo, E., Wiklund, O., Mattson Hultén, L., Hiukka, A., and Taskinen, M.-R. (2003) Insulin resistance and adiposity correlate with acute-phase reaction and soluble cell adhesion molecules in type 2 diabetes. *Atherosclerosis* **166,** 387–394
33. Fyfe, A. I., Rothenberg, L. S., DeBeer, F. C., Cantor, R. M., Rotter, J. I., and Lusis, A. J. (1997) Association between serum amyloid A proteins and coronary artery disease: evidence from two distinct arteriosclerotic processes. *Circulation* **96,** 2914–2919
34. Arroyo-Espliguero, R., Avanzas, P., Cosin-Sales, J., Aldama, G., Pizzi, C., and Kaski, J. C. (2004) C-reactive protein elevation and disease activity in patients with coronary artery disease. *Eur. Heart J* **25,** 401–408
35. Poitou, C., Viguerie, N., Cancello, R., De Matteis, R., Cinti, S., Stich, V., Coussieu, C., Gauthier, E., Courtine, M., Zucker, et al. (2005) Serum amyloid A: production by human white adipocyte and regulation by obesity and nutrition. *Diabetologia* **48,** 519–528
36. Wellen, K. E., and Hotamisligil, G. S. (2003) Obesity-induced inflammatory changes in adipose tissue. *J. Clin. Invest.* **112,** 1785–1788
37. Weisberg, S. P., McCann, D., Desai, M., Rosenbaum, M., Leibel, R. L., and Ferrante, A. W. Jr. (2003) Obesity is associated with macrophage accumulation in adipose tissue. *J. Clin. Invest.* **112,** 1796–1808
38. Su, S. B., Gong, W., Gao, J. L., Shen, W., Murphy, P. M., Oppenheim, J. J., and Wang, J. M. (1999) A seven-transmembrane, G protein-coupled receptor, FPRL1, mediates the chemotactic activity of serum amyloid A for human phagocytic cells. *J. Exp. Med.* **189,** 395–402
39. Boisvert, W. A. (2004) Modulation of atherogenesis by chemokines. *Trends Cardiovasc. Med.* **14,** 161–165
40. Chait, A., Han, C. Y., Oram, J. F., and Heinecke, J. W. (2005) Thematic review series: The immune system and atherogenesis. Lipoprotein-associated inflammatory proteins: markers or mediators of cardiovascular disease? *J. Lipid Res.* **46,** 389–403
41. van der Westhuyzen, D. R., Cai, L., de Beer, M. C., and de Beer, F. C. (2005) Serum amyloid A promotes cholesterol efflux mediated by SR-BI. *J. Biol. Chem.* **280,** 35890–35895
42. Artl, A., Marsche, G., Lestavel, S., Sattler, W., and Malle, E. (2000) Role of serum amyloid A during metabolism of acute-phase HDL by macrophages. *Arterioscler. Thromb. Vasc. Biol.* **20,** 763–772
43. Cai, L., de Beer, M. C., de Beer, F. C., and van der Westhuyzen, D. R. (2005) Serum amyloid A is a ligand for scavenger receptor class B type I and inhibits high-density lipoprotein binding and selective lipid uptake. *J. Biol. Chem.* **280,** 2954–2961
44. Stonik, J. A., Remaley, A. T., Demosky, S. J., Neufeld, E. B., Bocharov, A., and Brewer, H. B. (2004) Serum amyloid A promotes ABCA1-dependent and ABCA1-independent lipid efflux from cells. *Biochem. Biophys. Res. Commun.* **321,** 936–941
45. Le Lay, S., Krief, S., Farnier, C., Lefrere, I., Le Liepvre, X., Bazin, R., Ferre, P., and Dugail, I. (2001) Cholesterol, a cell size-dependent signal that regulates glucose metabolism and gene expression in adipocytes. *J. Biol. Chem.* **276,** 16904–16910
46. Karlsson, H. K., Tsuchida, H., Lake, S., Koistinen, H. A., and Krook, A. (2004) Relationship between serum amyloid A level and Tanis/SelS mRNA expression in skeletal muscle and adipose tissue from healthy and type 2 diabetic subjects. *Diabetes* **53,** 1424–1428
47. Gabrielsson, B. G., Johansson, J. M., Lönn, M., Jernås, M., Olbers, T., Peltonen, M., Larsson, I., Lönn, L., Sjöström, L., Carlsson, B., and Carlsson, L. M. (2003) High expression of complement components in omental adipose tissue in obese men. *Obes. Res.* **11,** 699–708
48. Pratley, R. E., Thompson, D. B., Prochazka, M., Baier, L., Mott, D., Ravussin, E., Sakul, H., Ehm, M. G., Burns, D. K., Foroud, T., Garvey, W. T., Hanson, R. L., Knowler, W. C., Bennett, P. H., and Bogardus, C. (1998) An autosomal genomic scan for loci linked to prediabetic phenotypes in Pima Indians. *J. Clin. Invest.* **101,** 1757–1764
49. Wu, X., Cooper, R. S., Borecki, I., Hanis, C., Bray, M., Lewis, C. E., Zhu, X., Kan, D., Luke, A., and Curb, D. (2002) A combined analysis of genomewide linkage scans for body mass index from the National Heart, Lung, and Blood Institute Family Blood Pressure Program. *Am. J. Hum. Genet.* **70,** 1247–1255
50. Lee, Y. H., Nair, S., Rousseau, E., Allison, D. B., Page, G. P., Tataranni, P. A., Bogardus, C., and Permana, P. A. (2005) Microarray profiling of isolated abdominal subcutaneous adipocytes from obese vs non-obese Pima Indians: increased expression of inflammation-related genes. *Diabetologia* **48,** 1776–1783
51. Cortes-Canteli, M., Wagner, M., Ansorge, W., and Perez-Castillo, A. (2004) Microarray analysis supports a role for ccaat/enhancer-binding protein-beta in brain injury. *J. Biol. Chem.* **279,** 14409–14417

Received for publication January 12, 2006.
Accepted for publication March 14, 2006.

Separation of human adipocytes by size: hypertrophic fat cells display distinct gene expression

Margareta Jernås,* Jenny Palming,* Kajsa Sjöholm,* Eva Jennische,† Per-Arne Svensson,* Britt G. Gabrielsson,* Max Levin,‡ Anders Sjögren,§ Mats Rudemo,§ Theodore C. Lystig,* Björn Carlsson,* Lena M. S. Carlsson,* and Malin Lönn*,[1]

*Research Centre for Endocrinology and Metabolism, Division of Body Composition and Metabolism, Department of Internal Medicine, †Institute of Anatomy and Cell Biology, and ‡Cardiovascular Institute and Wallenberg Laboratory, Sahlgrenska Academy at Göteborg University, Göteborg, Sweden; and §Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden

To read the full text of this article, go to http://www.fasebj.org/cgi/doi/10.1096/fj.05-5678fje

SPECIFIC AIMS

Enlargement of subcutaneous (s.c.) abdominal adipocytes is associated with insulin resistance and is an independent predictor of type 2 diabetes. The aim of the present study was to detect factors linking human adipocyte hypertrophy to insulin resistance/type 2 diabetes.

PRINCIPAL FINDINGS

1. Isolated human adipocytes, from a single adipose tissue sample, can be separated into populations of small cells and large cells

Most previous studies of the impact of adipocyte size, have studied fat cells or biopsies with different mean adipocyte diameters obtained from different tissue locations or even from different donors. Therefore, differences in environmental conditions or genetic factors that affect adipocyte gene expression and metabolism could not be excluded. Thus, it is not clear whether the functions of the fat cell vary with adipocyte size *per se*.

In the present study, a technique for separation of human adipocytes by size was developed. The technique, based on cell buoyancy and mesh filtration, separated isolated adipocytes from an adipose tissue sample into populations of small cells (mean 57.6±3.54 μm) and large cells (mean 100.1±3.94 μm). The mean size and the size distribution of the small and large populations, determined by computerized image analysis, differed significantly ($P<0.005$ and $P<0.001$, respectively) (**Fig. 1**).

2. Microarray analysis of the cell populations identified genes with markedly higher mRNA expression in large cells than small cells

Gene expression profiling (Affymetrix GeneChip HG-U133A arrays composed of 22,283 probe sets) of the cell populations separated from adipose tissue from three subjects identified 14 genes with more than 4-fold higher expression in large cells than small cells ($P<0.01$) (**Table 1**). Classification by cellular or organism function based on Gene Ontology definitions revealed that five of those genes were immune-related. The remaining nine genes were referred to structure (four), unknown function (three), growth (one) and transport (one) (Table 1). Differences in sample preparation or hybridization were excluded since there was no difference between small and large adipocytes in the expression of LRP10, CLN3, or COBRA1, suitable reference genes for studies of human adipose tissue.

3. SAA and TM4SF1 were ~20-fold higher expressed in the large adipocytes as determined by real-time RT-PCR, and adipocyte size correlated with the expression of both SAA and TM4SF1

One immune-related gene, serum amyloid A (SAA), and one gene with unknown function, transmembrane 4 L six family member 1 (TM4SF1), were selected among the genes with >4-fold higher expression in large vs. small cells for further analysis (Table 1). Leptin, being 3-fold higher expressed in the large adipocytes (data not shown), and previously suggested to be higher expressed in large compared with small adipocytes, was also included in the following studies.

The up-regulation of SAA, TM4SF1 and leptin expression in large cells was confirmed by real-time RT-PCR chain reaction analysis of small (mean 59.3±4.47 μm) and large (mean 97.1±5.69 μm) adipocytes from seven different adipose tissue samples. In all cases, SAA, TM4SF1 and leptin were expressed at

[1] Correspondence: RCEM/Division of Body Composition and Metabolism, Department of Internal Medicine, Vita Stråket 15, SE 413 45 Göteborg, Sweden. E-mail: malin.lonn@medic.gu.se
doi: 10.1096/fj.05-5678fje

0892-6638/06/0020-1540 © FASEB

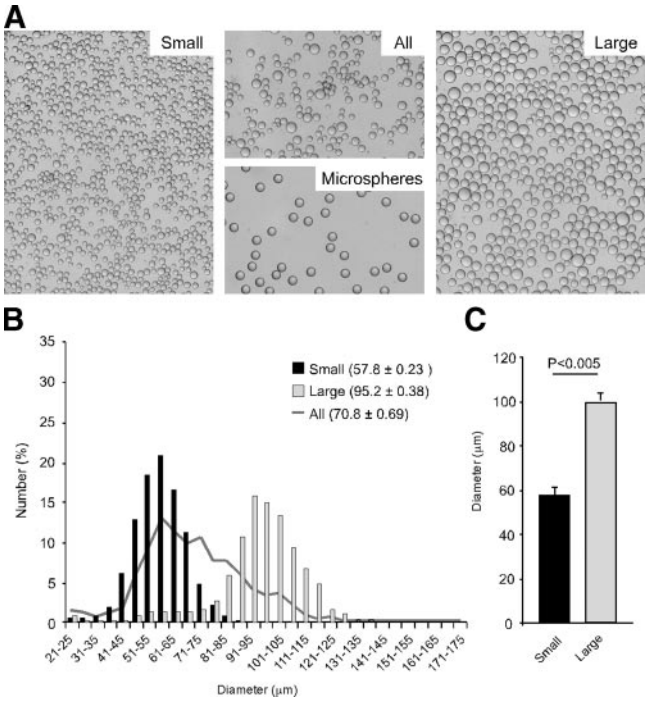


Figure 1. Separation of human adipocytes by size. *A)* Representative images of human adipocytes from one adipose tissue sample before separation (All) and after separation (Small and Large). Microspheres (98.00 μm in diameter) were used for reference in the computerized image analysis. *B)* Size distributions, determined by computerized image analysis, and mean diameters in random samples of all adipocytes (n=742), small adipocytes (n=1965), and large adipocytes (n=2382) from one adipose tissue sample. Patient characteristics; man, age: 41 yr, body mass index (BMI): 26.0 kg/m^2. *C)* Mean diameter of small and large adipocytes from 12 adipose tissue biopsies. Error bars indicate SEM.

higher levels in large cells (P=0.018). The mean fold increase in expression was 18.7 ± 15.1 for SAA, 22.3 ± 6.4 for TM4SF1, and 3.9 ± 1.4 for leptin. In addition, adipocyte size correlated with the expression of SAA (P=0.015), TM4SF1 (P=0.012), and leptin (P=0.0009).

4. In comparison with 17 other human tissues/cell types by microarray, large adipocytes displayed by far the highest SAA and TM4SF1 expression

GeneChip HG-U133A expression profiles from 17 different tissues were downloaded from the SymAtlas dataset (http://symatlas.gnf.org/Symatlas/). For comparison of gene expression in different tissues, the signal value for each gene was normalized by dividing the signal by the average signal of the entire array for each tissue. In addition, our own expression profiles, originating from small and large adipocytes, were included and normalized as outlined above. SAA, TM4SF1 and leptin expression levels in large adipocytes were compared to the levels in other human tissues and small adipocytes. The three genes were expressed at markedly higher levels in large adipocytes than in all other tissues/cell types.

5. The higher mRNA expression of SAA and TM4SF1 in large compared to small adipocytes was reflected also at the protein level

Expression of SAA and TM4SF1 in adipocytes was also demonstrated immunohistochemically. Although SAA immunoreactivity varied between cells of the same size, it was generally greater in large than in small adipocytes in the same sample. TM4SF1 immunoreactivity was demonstrated mainly in large but to some extent also in medium-sized adipocytes. Positive TM4SF1 signal appeared in a dot-like pattern in the cell membrane. Small adipocytes were completely without TM4SF1 immunoreactivity.

CONCLUSIONS AND SIGNIFICANCE

In this study, we developed a new technique to separate populations of small and large human adipocytes from a single adipose tissue sample. The two populations of cells obtained with our technique differed significantly in size, as determined with a computer-based image-analysis method that allows rapid analysis of 10-fold more cells than conventional methods. DNA microarray analysis of the two populations showed that several genes were expressed at markedly higher levels in the large cells, demonstrating that hypertrophy per se can significantly alter gene expression and thereby presumably adipocyte function (**Fig. 2**).

Previous studies of the metabolic activity of small and large adipocytes have indicated that adipocyte size influences various adipocyte metabolic functions. However, the cells of different sizes were obtained from different adipose tissue locations or from different donors. Thus, it was not possible to exclude environmental influences, such as nutritional/hormonal conditions, or genetic factors that might have affected gene expression and thus adipocyte metabolism. Our technique avoids these problems and will facilitate metabolic studies of fat cells of different sizes. For example, a positive correlation between human adipocyte size and leptin expression/secretion has previously been suggested. In the present study of human adipocytes from a single adipose tissue sample, the previous findings were confirmed since leptin was indeed expressed at higher levels in the large cells in all cases. Moreover, we identified several genes that, compared with leptin, showed a more pronounced differential expression in large vs. small adipocytes (Fig. 2).

Among the fourteen genes with markedly higher expression in large compared with small adipocytes, five were classified as immune-related; E-selectin, interleukin-8, SAA, C1q receptor 1, and CXCL2 also known as MIP-2 or macrophage inflammatory protein-2. Components of the metabolic syndrome, such as obesity and type 2 diabetes, are associated with a systemic increase in inflammatory markers. The acute-phase proteins SAA and C-reactive protein have attracted particular attention because they are independent risk factors for

TABLE 1. *Genes detected with more than fourfold higher expression in large adipocytes than small adipocytes as analyzed by DNA microarray*

Gene symbol	Gene name	ID	Classification	Fold change	*P* value	Mean signal small	Mean signal large
SELE	selectin E	206211_at	defense	15.2	1.29E-09	11.5	174.4
SPARCL1	SPARC-like 1	200795_at	unknown	14.9	1.22E-09	33.5	498.8
TM4SF1	transmembrane 4 L six family member 1	209386_at	unknown	11.6	4.46E-11	50.6	586.3
TM4SF1	transmembrane 4 L six family member 1	209387_s_at		9.6	4.19E-09	22.5	216.1
TM4SF1	transmembrane 4 L six family member 1	215034_s_at		9.2	2.29E-10	26.7	245.5
DCN	decorin	211896_s_at	structure	9.4	8.18E-09	59.3	554.7
DCN	decorin	201893_x_at		6.9	3.70E-09	156.8	1076.5
DCN	decorin	211813_x_at		5.6	1.90E-09	118.4	659.1
IL8	interleukin 8	202859_x_at	defense	9.0	4.44E-09	186.8	1682.5
IL8	interleukin 8	211506_s_at		7.5	3.54E-07	55.5	415.2
PALLD	palladin	200897_s_at	structure	8.7	2.53E-09	24.6	213.8
SAA2	serum amyloid A2	208607_s_at	defense	7.8	3.82E-08	245.7	1918.5
SAA2	serum amyloid A2	214456_x_at		7.6	2.24E-07	529.1	4009.6
CLEC3B	C-type lectin domain family 3, member B	205200_at	growth	6.3	8.22E-09	55.8	353.5
C1QR1	complement component 1, q subcomponent, receptor 1	202878_s_at	defense	6.1	9.40E-10	30.0	183.9
COL1A1	collagen, type I, alpha 1	202310_s_at	structure	5.9	2.27E-08	26.2	155.6
CXCL2	chemokine (C-X-C motif) ligand 2	209774_x_at	defense	5.4	8.09E-09	184.4	998.2
COL1A2	collagen, type I, alpha 2	202403_s_at	structure	5.3	2.44E-07	57.1	300.9
FLJ14054	—	219054_at	unknown	4.2	4.65E-09	65.2	273.6
AQP1	aquaporin 1	209047_at	transport	4.2	5.90E-09	43.2	179.7

coronary artery disease. We, and others, have previously shown that adipose tissue is a major site of SAA production and is likely to be a major source of circulating SAA in obese patients. The current study extends these findings by demonstrating that SAA is expressed at the highest concentration by the large adipocytes. SAA has been implicated in inflammation, insulin resistance and impairment of reverse cholesterol transport. Our data may therefore suggest that adipocyte-derived SAA, likely having both local effects and endocrine functions, is a potential mediator of the link between hypertrophic adipocytes and type 2 diabetes (Fig. 2).

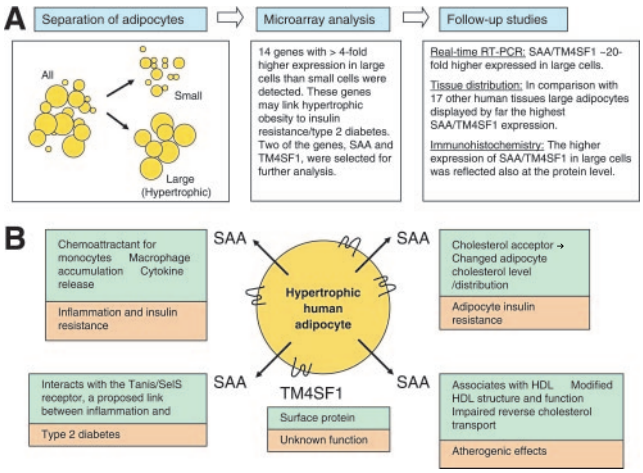


Figure 2. Schematic diagram. Flow chart and summary of the principal findings (*A*). Potential mechanisms by which serum amyloid A (SAA) and transmembrane 4 L six family member 1 (TM4SF1), expressed at the highest levels by large adipocytes, may mediate a link between hypertrophic obesity and insulin resistance/type 2 diabetes (*B*).

To summarize, we have developed a technique to separate human adipocytes, from a single adipose tissue sample, by size. The resulting populations of small and large adipocytes have significantly different cell size distributions. Gene expression profiling of the small and large adipocytes identified genes, many of them immune-related, with markedly higher mRNA expression in the large cells. Two of those genes, SAA (an acute-phase protein implicated in inflammation, insulin resistance and impairment of reverse cholesterol transport) and TM4SF1 (a membrane protein with unknown function), were ~20-fold higher expressed in the large cells, a difference reflected also at the protein level. Moreover, in comparison with several other human tissues, large adipocytes displayed by far the highest SAA and TM4SF1 expression. In the light of previous studies reporting that adipocyte hypertrophy is associated with insulin resistance and is an independent predictor of type 2 diabetes, the findings in the current work provide novel insights into the molecular connection between hypertrophic obesity and insulin resistance/type 2 diabetes. FJ

Figures

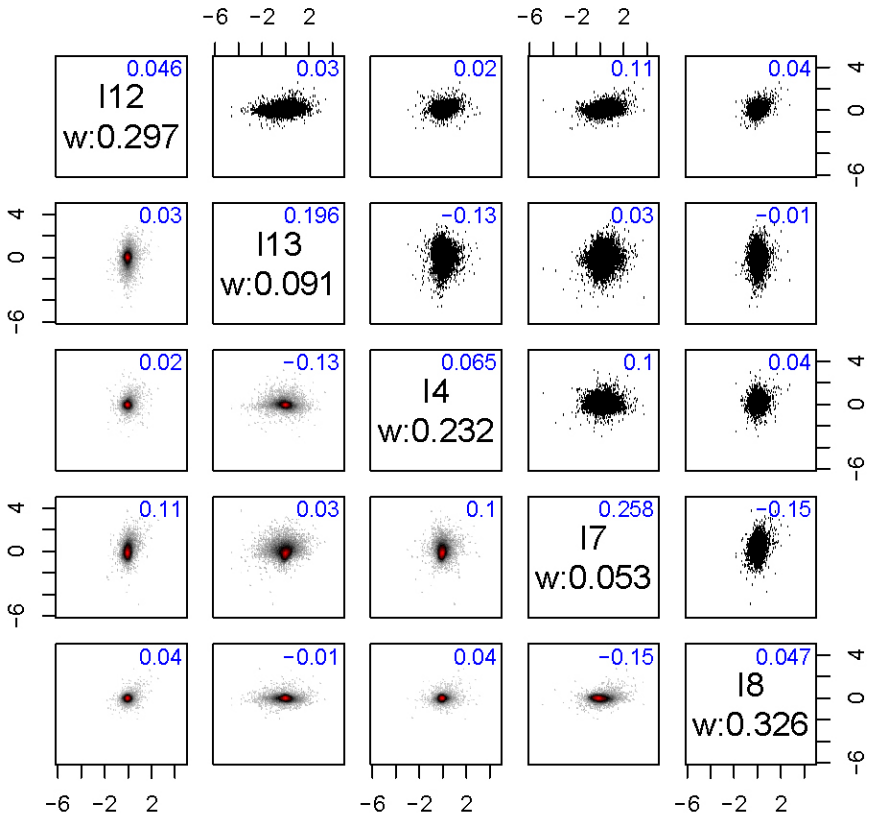


Figure 1: Figure 3 from Paper I, with estimates from WAME added. Pairwise plots of the $\log_2$-ratios of the patients in the ishemic group of the Cardiac dataset. The plots to the lower-left show two-dimensional kernel density estimates of the distribution of $\log_2$-ratios in each pair of patients. This provides information in the central areas where the corresponding scatterplots are solid black (cf. Figure 6 in (Huber et al., 2003)). The colour-scale is, in increasing level of density: white, grey, black and red. Off-diagonal numbers show estimated correlations from WAME. Diagonal boxes contain sample name and weights as well as estimated variances from WAME.

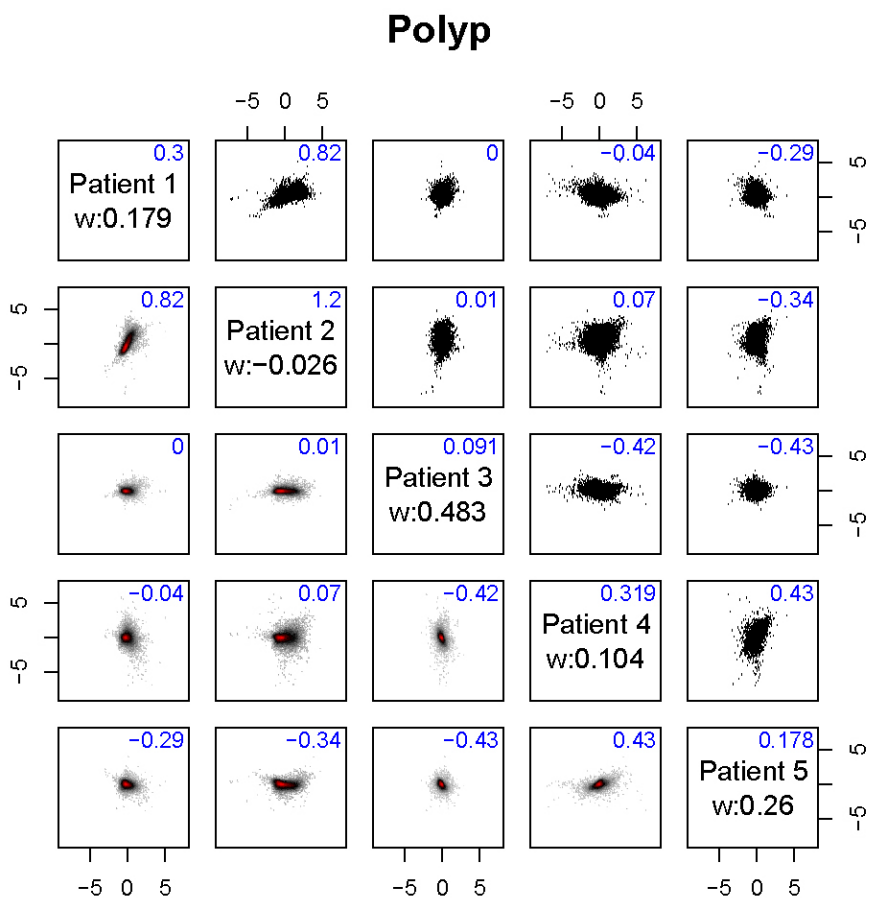


Figure 2: Figure 6 from Paper I, with estimates from WAME added. Pairwise plots of the $\log_2$-ratios of the patients in the Polyp dataset. The plots to the lower-left show two-dimensional kernel density estimates of the distribution of $\log_2$-ratios in each pair of patients. This provides information in the central areas where the corresponding scatterplots are solid black (cf. Figure 6 in Huber et al. (2003)). The colour-scale is, in increasing level of density: white, grey, black and red. Off-diagonal numbers show estimated correlations from WAME. Diagonal boxes contain sample name and weights as well as estimated variances from WAME.

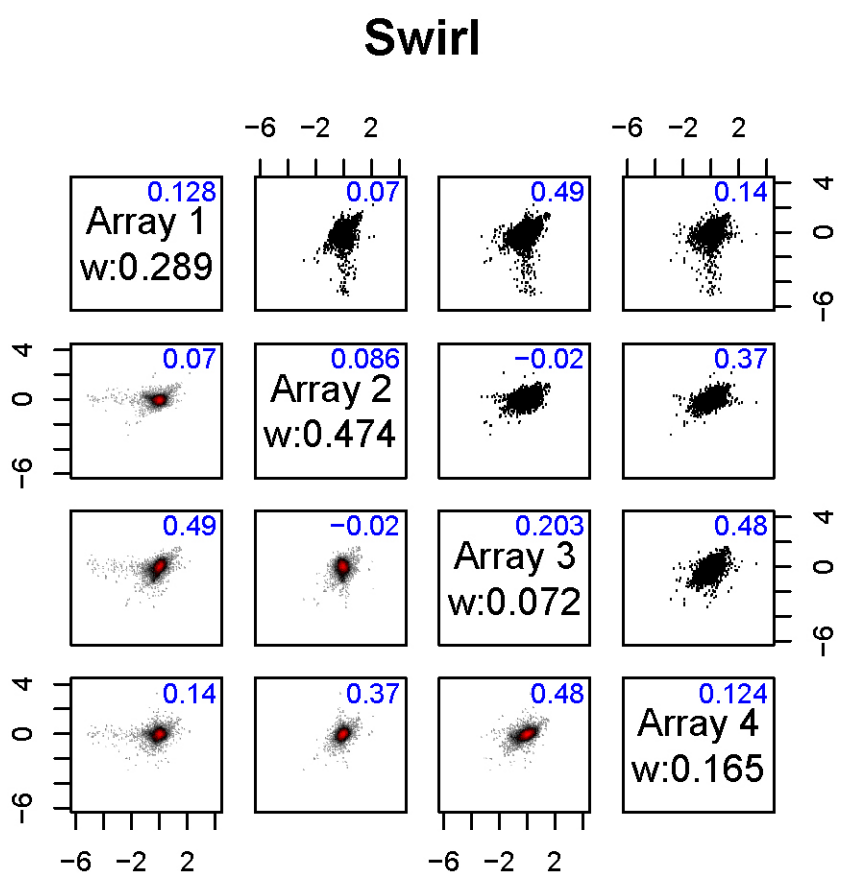
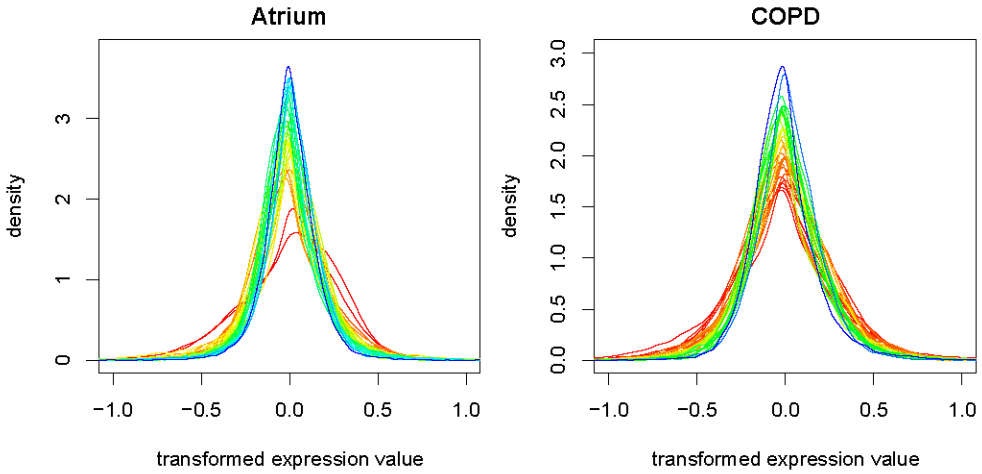


Figure 3: Figure 8 from Paper I, with estimates from WAME added. Pairwise plots of the $\log_2$-ratios of the arrays in the Swirl dataset. The plots to the lower-left show two-dimensional kernel density estimates of the distribution of $\log_2$-ratios in each pair of patients. This provides information in the central areas where the corresponding scatterplots are solid black (cf. Figure 6 in Huber et al. (2003)). The colour-scale is, in increasing level of density: white, grey, black and red. Off-diagonal numbers show estimated correlations from WAME. Diagonal boxes contain sample name and weights as well as estimated variances from WAME.



expression values, $\mathbf{Y}$, for the different arrays, in the two datasets. Colour-coding according to sample variance is used for increased clarity (blue for low variance, red for high variance). Differences in variability can be noted in both datasets.

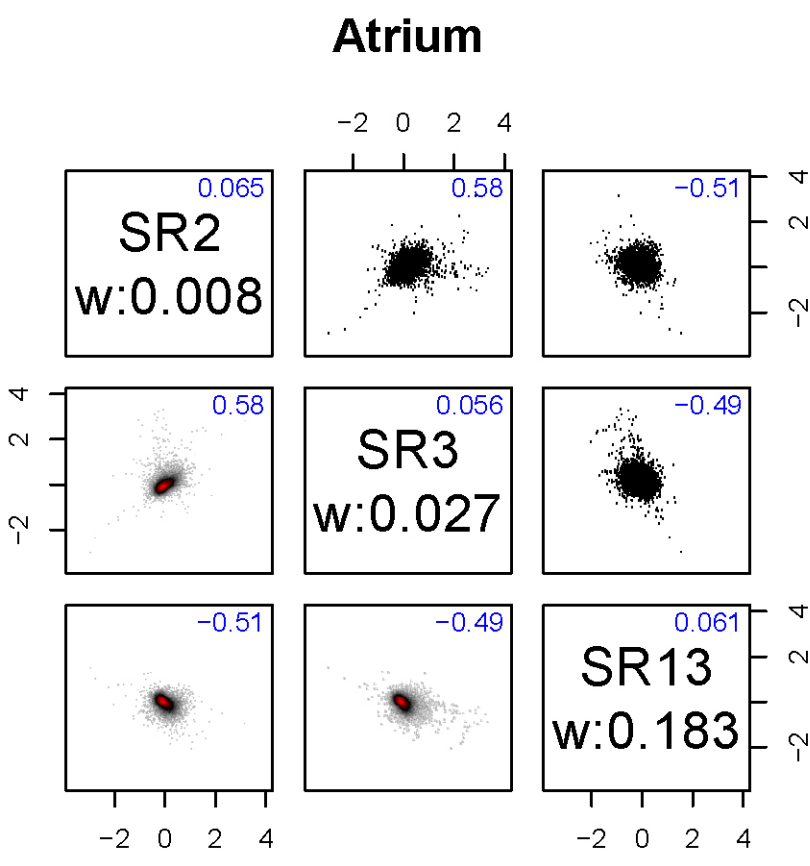


Figure 5: The Atrium part of Figure 2 in Paper III. Pairwise plots of transformed expression values, $\mathbf{Y}_g$, for selected pairs of arrays within the same group. Different pairs within the same group have distinctly different correlations. Upper triangle contains scatterplots. Lower triangle contains heatmaps of the corresponding two-dimensional kernel density estimates, where the majority of the genes are in the red portion of the plot, revealing important trends inside the black clouds. Off-diagonal numbers show estimated correlations from WAME. Diagonal boxes contain sample names and weights as well as estimated variances from WAME.

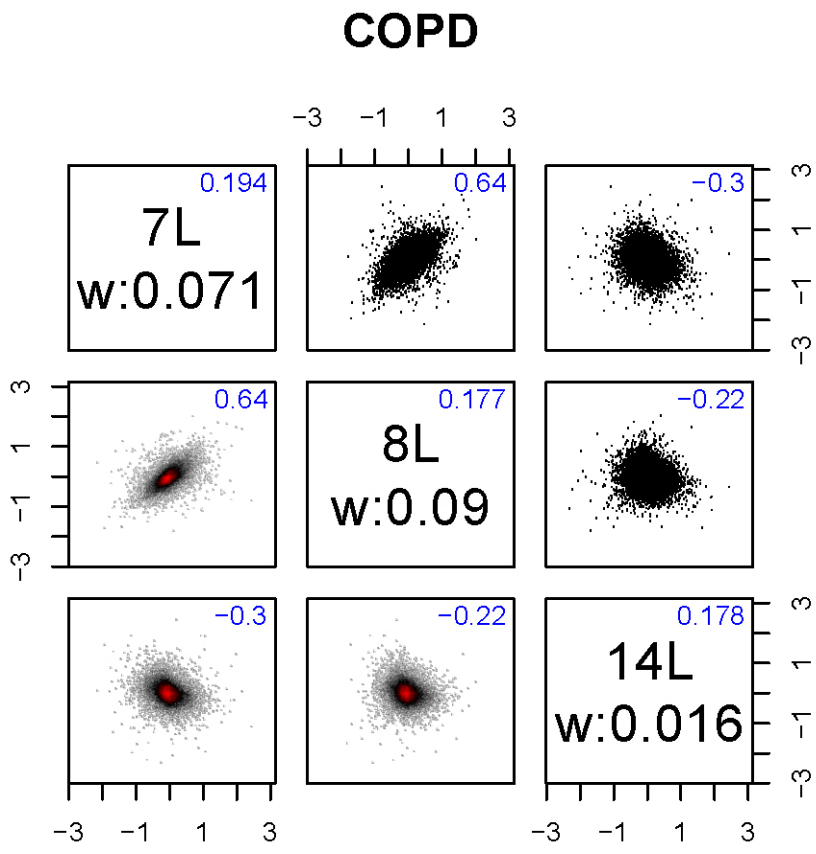


Figure 6: The COPD part of Figure 2 in Paper III. Pairwise plots of transformed expression values, $\mathbf{Y}_g$, for selected pairs of arrays within the same group. Different pairs within the same group have distinctly different correlations. Upper triangle contains scatterplots. Lower triangle contains heatmaps of the corresponding two-dimensional kernel density estimates, where the majority of the genes are in the red portion of the plot, revealing important trends inside the black clouds. Off-diagonal numbers show estimated correlations from WAME. Diagonal boxes contain sample names and weights as well as estimated variances from WAME.

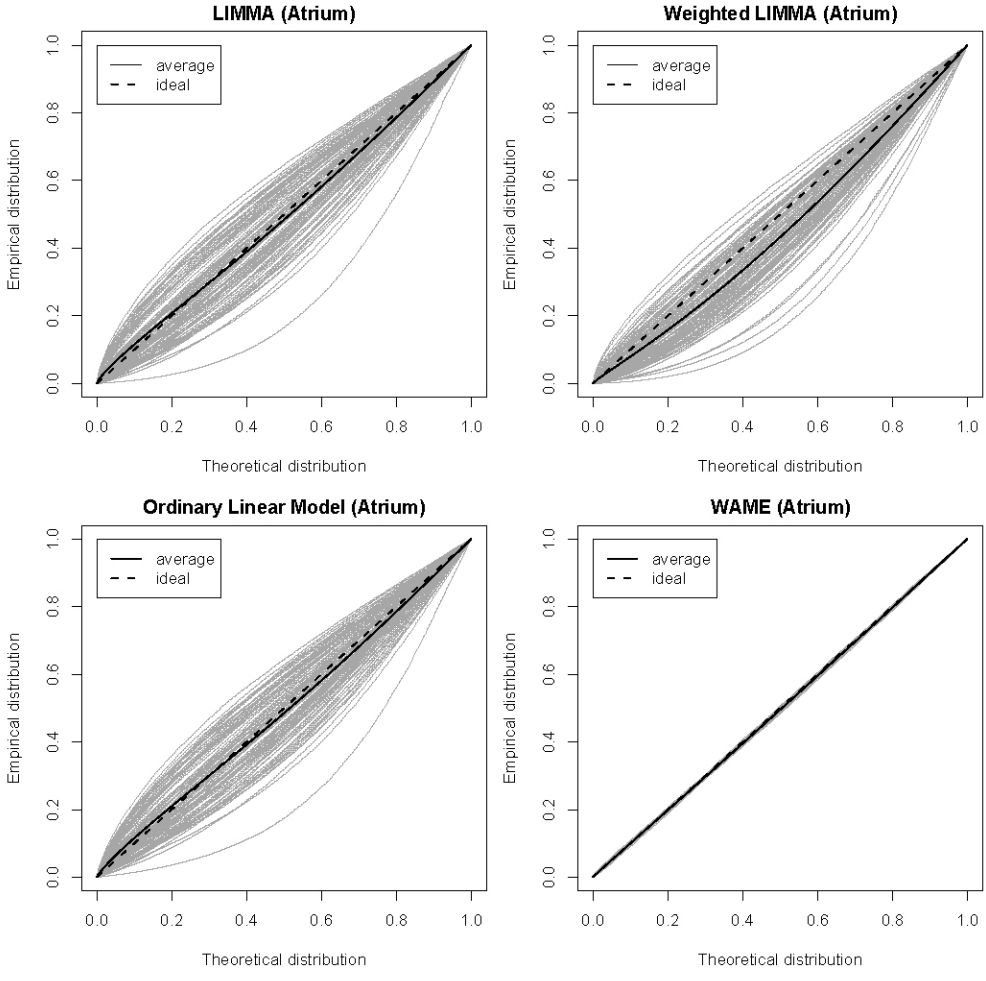


Figure 7: Empirical distributions of p-values for LIMMA, weighted LIMMA, OLM and WAME from tests on 100 resamples from the Atrium dataset. Average empirical distribution indicated. Since no signal is added, the curves should ideally follow the diagonal. The results are very similar to those from the COPD dataset (see Figure 4 in Paper III).

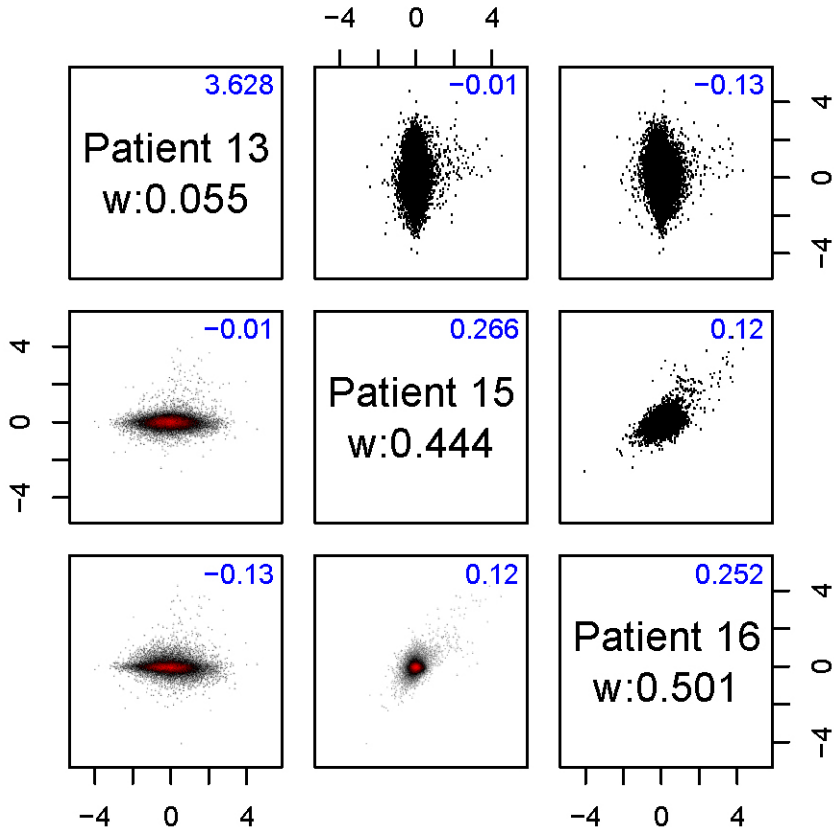


Figure 8: Pair-wise plots of the $\log_2$-ratios between large and small adipocytes for the three patients in Paper IV. Upper triangle contains scatterplots. Lower triangle contains heatmaps of the corresponding two-dimensional kernel density estimates, where the majority of the genes are in the red portion of the plot, revealing important trends inside the black clouds. Off-diagonal numbers show estimated correlations from WAME. Diagonal boxes contain sample name and weights as well as estimated variances from WAME. Note the high variance for Patient 13.