Thesis for the degree of Doctor of Philosophy

Statistical analysis of gene expression data

Erik Kristiansson

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

Statistical analysis of gene expression data
Erik Kristiansson
ISBN 978-91-7385-039-1

Doktorsavhandlingar vid Chalmers Tekniska Högskola
Ny serie Nr 2720
ISSN 0346-718X

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

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

Statistical analysis of gene expression data
Erik Kristiansson
Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University

Abstract

Microarray technology has become one of the most important tools for genome-wide mRNA measurements. The technique has been successfully applied to many areas in modern biology including cancer research, identification of drug targets, and categorization of genes involved in the cell cycle. Nevertheless, the analysis of microarray data is difficult due to the vast dimensionality and the high levels of noise. The need for solid statistical methods is therefore strong.

The main results are presented in six papers. The first three develop a statistical model for quality assessment and improved gene ranking called Weighted Analysis of Microarray Experiments (WAME). Here, the customary assumption of independent samples is shown to be invalid and individual variances for each array and correlations between pairs of arrays are introduced. Comparisons to other common methods suggest that the proposed model produces more accurate results. The first paper describes the model for simple experimental setups for two-channel arrays. This model is then generalized to more complex designs in paper two and to one-channel microarrays in paper three.

Transcription factors govern gene expression in the cell by binding to short sequences called *cis*-regulatory elements. These sequences are located in the promoters, which are regions of DNA upstream of the genes. In paper four, we show that the lengths of these promoters are related to gene function. In particular, the promoters for stress responsive genes are in general longer than those of other genes. This is used in a novel method for identifying relevant *cis*-regulatory elements from a list of differentially expressed genes.

Papers five and six present microarray based studies from molecular biology and environmental toxicology respectively. In paper five, microarrays are used to identify *Saccharomyces cerevisiae* genes with changed mRNA levels under arsenic stress. In paper six, biomarkers for estrogen exposure in fish are found using both an in-house microarray experiment and a meta-analysis of several public gene expression datasets.

Keywords: gene expression, DNA microarrays, linear models, empirical Bayes, quality control, gene regulation, categorical data analysis, logistic regression, heavy metal stress, ecotoxicology

Acknowledgments

I would like to express my greatest appreciation to my adviser Olle Nerman for his guidance and encouragement during my years as a Ph.D. student. My thanks also go to my co-supervisor Graham Kemp for his help and support.

I would also like to thank Anders Sjögren, Henrik Nilsson, Michael Thorsen and Lina Gunnarsson for many years of nice collaborations. This thesis wouldn't look the same without you!

Other persons that has affected my research in a positive way are Alexandra Jauhiainen, Martin Ryberg, Joakim Larsson, Janeli Sarv, Gabriella Arne, Mats Rudemo, Markus Tamás, Ola Nilsson, Olga Kourtchenko, Sven Nelander, Selpi, John Gustafsson, Magnus Ekdahl, Staffan Nilsson and Petter Mostad. Thank you all!

Furthermore, I would like to thank Lech Maligranda and Reinhold Näslund who taught me during my undergraduate studies at Luleå University of Technology. Without your inspiration, this thesis would not exist.

I am also grateful to the Swedish National Research School in Bioinformatics and the coordinator Anders Blomberg for funding this work and keeping the school active.

My thanks also go to Oscar Hammar, Anna Larsson and the rest of biokorridoren, the lunch gang, the other lunch gang, *e*-team and c93 (*imperium nihil est*).

Finally, I would like to thank my family. They have always encouraged and supported me in my studies.

Erik Kristiansson
Göteborg
October 2007

Table of contents

List of included papers

This thesis contains the following papers:

1. Erik Kristiansson*, Anders Sjögren*, Mats Rudemo and Olle Nerman (2005). Weighted analysis of paired microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, **4**(1), Article 30.

2. Erik Kristiansson, Anders Sjögren, Mats Rudemo and Olle Nerman (2006). Quality optimised analysis of general paired microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, **5**(1), Article 10.

3. Anders Sjögren, Erik Kristiansson, Mats Rudemo, and Olle Nerman (2007). Weighted analysis of general microarray experiments, *BMC Bioinformatics*, **8**(387).

4. Erik Kristiansson, Michael Thorsen, Markus J Tamas, Olle Nerman (2007). Evolutionary forces act on promoter length: assessment of enriched *cis*-regulatory elements, *submitted*.

5. Michael Thorsen, Gilles Lagniel, Erik Kristiansson, Christophe Junot, Olle Nerman, Jean Labarre, Markus J Tamas (2007). Quantitative transcriptome, proteome and sulfur metabolite profiling of the Saccharomyces cerevisiae response to arsenite, *Physiological Genomics*, **30**, pp 35-43.

6. Lina Gunnarsson, Erik Kristiansson, Lars Förlin, Olle Nerman, D. G. Joakim Larsson (2007). Sensitive and robust gene expression changes in fish exposed to estrogen - a microarray approach, *BMC Genomics*, **8**(132).

* equal contribution

Introduction

Many parts of modern biology study life at cellular and molecular level. As stipulated by the central dogma; genes, the units of inheritance located in the DNA, are transcribed into mRNA which, in turn, are translated into proteins, the main building blocks of the cell (Gerstein et al., 2007). Proteins are the primary actors in the cell and are crucial for most biological processes. It has however been proven difficult to measure protein abundance in large scale, and much effort has therefore been put into measuring their originators, the mRNAs.

DNA microarrays (Schulze and Downward, 2001) were introduced in the mid-90s and have for more than ten years been popular tools for large scale measurement of mRNA abundance. Indeed, roughly 10,000 datasets have been submitted to the microarray repositories, comprising more than 100,000 microarrays (Barrett et al., 2007; Parkinson et al., 2007; Demeter et al., 2007). Microarray techniques have been successfully applied to various areas in modern biology including categorization of cell cycle genes in yeast (Spellman et al., 1998), classification of various cancers such as leukemia (Golub et al., 1999) and breast cancer (Sorlie et al., 2001), and drug target validation (Marton et al., 1998).

The popularity of the microarray technique notwithstanding, the analysis of the data is far from trivial. In fact, new methods are constantly proposed, indicating that there are many important issues left to solve. This introduction serves as a short summary of the microarray field with focus on the statistical and computational issues involved.

Overview of the DNA microarray technology

Several microarray platforms are available for large scale measurements of mRNA abundance. In this section we introduce the most common types and describe their properties.

Spotted cDNA microarrays

cDNA microarrays (Schena et al., 1995; Leung and Cavalieri, 2003) are created from cDNA libraries, i.e., collections of DNA clones containing the different mRNAs expressed in a specific tissue or cell culture. These clones, typically a few hundred base pairs long, are amplified using polymerase chain reaction (PCR) and then spotted onto a glass slide in a grid-like pattern (Figure 1). The spots have a diameter of 100-300 μm and a single slide can thus contain over 100,000 spots. Each clone, and hence each spot, corresponds to a transcript from a specific gene.

To measure gene expression in a tissue or cell culture, mRNA is extracted, reverse transcribed to cDNA, labeled with a fluorescent dye and then hybridized to the microarray. Spotted cDNA microarrays are typically of two-channel type, where two sources of mRNA, labeled with different dyes, are hybridized together in a competitive setting. Thus, for each spot there are two types of cDNA bound, one from each source.

Oligonucleotide microarrays

Oligonucleotide arrays (Wodicka et al., 1997) are based on oligonucleotides, short sequences of single stranded DNA. These sequences are either synthesized from the ground up while attached to a surface (*in situ* synthesized oligos) or pre-synthesized and then spotted onto a glass slide. The length of the oligos ranges from 25 to 70 bases. *In situ* synthesized arrays are typically of one-channel type, while spotted oligonucleotide arrays use two channels.

For *in situ* synthesized microarrays, a set of identical oligos located within a specific region is called a probe (Figure 1). If short oligos are used , many probes are usually needed for a single genes, measuring different parts of the transcripts and sometimes even different splice-variants (Cuperlovic-Culf et al., 2006). If longer oligos are used, one probe per gene is generally enough. For spotted oligonucleotide microarrays nevertheless, each gene typically corresponds to a single spot.

Oligo design is done *in silico* and several methods are available (Rouillard et al., 2003; Li and Ying, 2006; Wernersson and Nielsen, 2005). The oligos used on a oligonucleotide microarray should ideally be gene-specific in the sense that they should only be able to form complementary bindings with a single transcript from a specific gene. Simple principles are used to achieve this: for a transcript of a specific gene, test all possible oligos of a fixed length, and score each oligo based on a number of properties such as melting temperature, tendency to self-fold and position on the transcript. The oligos with highest scores are then picked to be used on the array. Little is however known about what influences the hybridization efficiency (Pozhitkov et al., 2006; Peytavi et al., 2005) which makes it hard to create a good scoring scheme. Furthermore, issues such as different splice variants (Modrek and Lee, 2002; Brett et al., 2002), paralogous genes (Spring, 2002) and sparse genome annotation (Baumgartner et al., 2007) make oligo design intricate. Further research is therefore needed to improve these methods and thus increase the general accuracy of oligonucleotide microarrays.

To hybridize mRNA to a oligonucleotide microarray, mRNA is converted to cRNA (RNA derived from cDNA) and labeled with a florescent dye, typically streptavidin, through a series of steps. For two-channel oligonucleotide microarrays, two sources of mRNA are hybridized together, while only one source is used for the one-channel arrays.

Pre-processing microarray data

Pre-processing of microarrays is the step of converting the each array in the experiment into applicable data that can be used to identify differentially expressed genes.

Quantification and image analysis

The amount of mRNA attached to the microarray is quantified using a laser scanner and a sensor. The laser emits photons at a specific wavelength which the fluorescence dyes absorb, which in turn, emit photons at a wavelength detectable by the sensor. The process results in

an image where regions with more labeled mRNA have higher intensity compared to regions sparse of mRNA (Yang et al., 2001). Scanners for two-channel arrays use two different lasers, one for each dye, and hence two images are produced. Examples of images from a two-channel spotted cDNA microarray and from a one-channel oligonucleotide microarray can be seen in Figure 1.

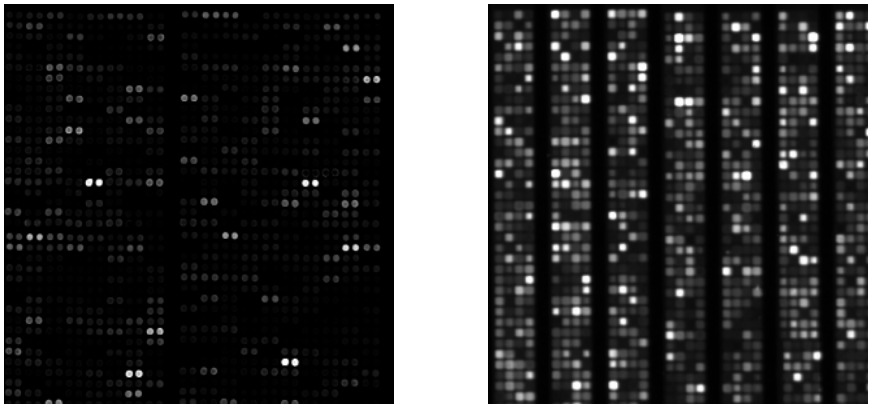


Figure 1: On the left is an image from one of the channels of a two-channel spotted cDNA microarray. This particular array is a yeast (*Saccharomyces cerevisiae*) chip containing all the 6,000 genes in the genome. To increase the power in the subsequent statistical analysis, the genes are spotted twice, resulting in roughly 12,000 spots in total. On the right is an image from the Geniom platform, showing an *in situ* hybridized one-channel microarray. Each square in this picture corresponds to a probe, each containing over 10,000 oligos. This particular array was designed to investigate genes affected by estrogen in zebrafish. Note that the scales of the two images are different. The spots on the cDNA microarray are roughly 200 μm in diameter, while the probes on the Geniom microarray measure 34×34 μm.

Image analysis tools are used to extract gene-specific values from the images. For spotted two-channel microarrays, the two images are first superpositioned and then the spots and their boundaries are located. For each spot, a foreground and a background intensity are calculated. Other data, such as the spot area and shape and various quality flags, are also extracted. Several applications have these methods implemented, for example, Imagene (BioDiscovery, 2007), GenePix (Molecular Devices, 2007), Spot (CSIRO, 2007) and ScanAlyze (Eisen Lab, 2007). Observe that all these softwares use different algorithms and will thus produce slightly different results.

For *in situ* synthesized microarrays, the position of each probe is known by design, and hence no probe localization is needed. The foreground intensity for each probe is calculated by averaging intensity over the corresponding region. Software for image analysis

of *in situ* synthesized microarrays is typically provided by the manufacturer, for example, GCOS (Affymetrix microarrays, Affymetrix (2007)) and Geniom Software (Geniom One microarrays, Febit Biomed (2007)).

Background correction

Non-specific binding, residues from washing and noise from the scanning process give rise to unwanted background signals. To remove these discrepancies from the spot/probe intensities, a background correction step is performed. For spotted microarrays, the background is estimated for each spot by taking the average intensity in an area close to, but disjoint from, the spot area (Yang et al., 2001). For *in situ* hybridized microarrays, the intensity of mismatches, i.e. oligos with one or more of the base pairs mismatched to the transcript it was designed to measure, are used to estimate the background (Urakawa et al., 2003).

Several approaches to background correction has been suggested: subtract, normexp (Ritchie et al., 2007; Irizarry et al., 2003b; McGee and Chen, 2006), variance stabilization (Huber et al., 2002), Edwards (Edwards, 2003) and Kooperberg (Kooperberg et al., 2002). The most commonly used method is subtract, where the background is simply subtracted from the foreground. This method is however questionable, since there is often a positive correlation between the foreground and background estimates. Indeed, in a recent paper (Ritchie et al., 2007), several background correction methods were compared and the performance of subtract were shown to be inferior to all other methods investigated, including using no background correction at all.

Normalization

Microarray experiments consist of a number of microarrays which need to be compared. Systematic errors, due to unequal quantities of starting material, different efficiencies in the labeling steps and different properties of the dyes, can reduce the reproducibility and hence the quality of an entire experiment. These artifacts can, however, be restrained by applying proper normalization algorithms.

The data from two-channel spotted microarrays is typically transformed before normalization (Quackenbush, 2002; Smyth and Speed, 2003). Let R_g and G_g be intensities for gene g (after background correction) for the two channel (as in "red" and "green" channel), and define the logarithmic fold-change M_g and the average logarithmic intensity A_g as

$$M_g = \log_2 R_g - \log_2 G_g, \quad A_g = \frac{1}{2}(\log_2 R_g + \log_2 G_g).$$

The upper part of Figure 2 shows a plot of the transformed data from a spotted two-channel cDNA microarray with roughly 6000 spots, each one corresponding to a gene in *Saccharomyces cerevisiae* (Baker's yeast). In this experiment, mRNA from arsenite exposed yeast cells (red channel) are compared to yeast cells without exposure (green channel) (see Paper 5 for more details). Several systematic trends can be seen in Figure 2a, most

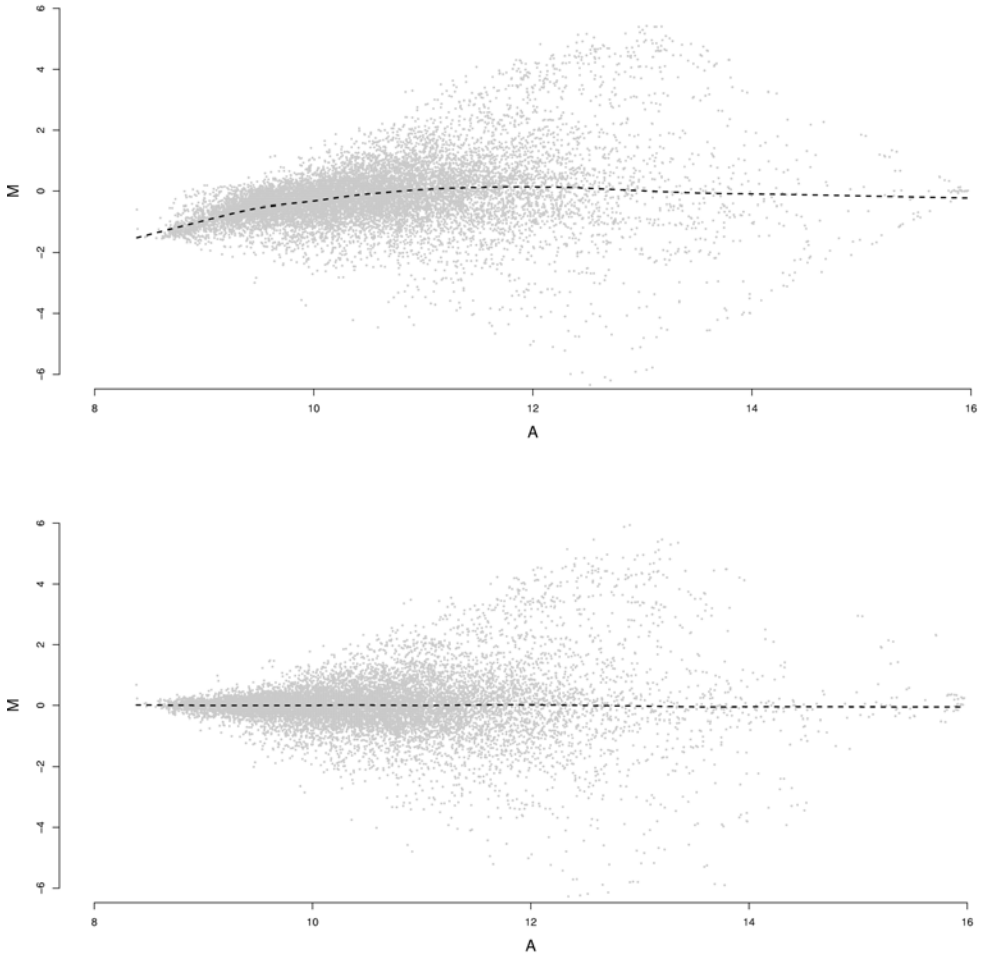


Figure 2: Plots of the M- and A-values for a yeast microarray measuring the mRNA difference between arsenic stressed and control cells for all 6000 genes in the genome. The upper plot shows the data before normalization and a clear trend can be seen (dashed line). In the lower plot, the data has been normalized with lowess normalization and the trend is removed. The cloud has also been centered around $M = 0$.

notably, that (1) the M-values are dependent on A-values and (2) there are more negative than positive M-values. These phenomena are typical for microarray data and hard to explain with biological arguments. It is therefore generally believed that they are technical artifacts from the experiments and should thus be removed.

Several normalization methods exist for normalization of two-channel spotted microarrays: median, lowess (Yang et al., 2002), print-tip lowess (Yang et al., 2002), splines (Workman et al., 2002; Baird et al., 2004), wavelets (Wang et al., 2004) and non-parametric methods based on support vector machines (Fujita et al., 2006). The commonly used lowess method fits a robust weighted regression line (Cleveland, 1979, 1981) to the MA-plot (Figure 2), which is then subtracted from the M-values. As can be seen in the lower part of Figure 2, the M-values are now centered around zero and the dependence between the M and A values is reduced.

Systematic errors can also affect groups of arrays, e.g. bad batches and time-dependent phenomena, and thus sets of arrays may require normalization. Several procedures have therefore been developed to normalize between two-channel spotted microarrays and the most commonly used are scale (Yang et al., 2002) and quantile-quantile (Bolstad et al., 2003) normalization.

Correction of one-channel *in situ* synthesized microarrays are based on a slightly different approach, where a normalization algorithm is applied to the probes (Bolstad et al., 2003; Welle et al., 2002; Åstrand, 2003), which are then combined to a gene value (Li and Wong, 2001a,b; Chu et al., 2002). Software, with both these steps implemented, include dChip (Li and Wong, 2001a), RMA (Irizarry et al., 2003a; Bolstad et al., 2003; Irizarry et al., 2003b), gcRMA (Wu et al., 2004) and MOID (Zhou and Abagyan, 2002).

Statistical analysis of microarray data

The aim of most microarray experiments is to identify genes with differences in mRNA levels between the studied conditions. This is in essence a statistical problem where many traditional methods fail due to the high dimensionality and the high levels of noise. During recent years, a vast number of methods has been suggested, and some of them are discussed here.

Identification of differentially expressed genes

Identification of differentially expressed genes from microarray data is done gene-wise, where a hypothesis test is performed for each gene. The null hypothesis usually states that there is no change in mRNA levels between the conditions of interest, but more complex hypotheses are also common. Standard methods such as linear models, including t- and F-tests, linear regression, and ANOVA, have been proposed (Kerr, 2003; Bretz et al., 2005). Even though the flexible nature of linear models fit the experimental design of many microarray experiments, these methods can perform inadequately due to the low number of replicates usually present.

One way to improve the performance is to penalize the gene-specific variance estimates. Efron et al. (2001) suggested that a small number should be added to the variance of the t-statistic, the so called *fudge factor*. Their optimal value of the *fudge factor* was derived

using cross-validation and was shown to be the 90th percentile of the sample variance distribution. Since the analytical distribution of penalized t-statistics is unknown, the null-distribution was estimated by permuting the arrays between the conditions. This approach has been implemented in the popular SAM package (Tusher et al., 2001).

Methods based on Bayesian model assumptions can also be used to control the variances. This is achieved by assuming *a priori* distributions for all the gene-specific variances with low probability close to zero. Since the observations generally are assumed to be normally distributed, the conjugate inverse gamma distribution is a popular choice. The hyperparameters can either be estimated from data by using an empirical Bayes approach (Baldi and Long, 2001; Lönnstedt and Speed, 2002; Smyth, 2004) or assumed to follow non-informative priors (Lönnstedt and Britton, 2005; Chi et al., 2007). The former approach results in moderated t- and F-statistics, which have known distributions under the null hypothesis. The empirical Bayes approach is implemented in the R-package LIMMA (R Development Core Team, 2007; Gentleman et al., 2004).

The penalized and moderated methods have been shown to outperform the standard methods (Paper 3) and have become *de facto* standard when analyzing microarray datasets with few observations. However, there are situations in which they are not suitable, e.g. when the gene-specific variance is believed to change between conditions. Furthermore, they all rely on the assumption of independent replicates, an assumption that has been shown to be too optimistic. Indeed, in Paper 1, 2 and 3 the assumption of independence is relaxed and variance heterogeneity and correlations are introduced for the different arrays. This is also further elaborated in Åstrand et al. (2007a,b).

Bioinformatical tools for gene expression data

Drawing biological conclusions from a list of hundreds of potentially differentially expressed genes is a non-trivial task and several bioinformatical tools for gene list interpretation have therefore been developed.

One example is the identification of enrichments of *cis*-regulatory motifs, which can be used to gain knowledge about the underlying regulatory mechanisms that gave rise to the observed differential expression (Bussemaker et al., 2001). The *cis*-regulatory motifs are short sequences, typically between 6 and 15 base pairs long, located upstream of the genes (Matys et al., 2003). When the cell initiates transcription, a transcription factor protein needs to interact with DNA by binding to its corresponding motif. It is therefore, given a set of differentially expressed genes, possible to search for enrichments of *cis*-regulatory motifs to deduce which transcription factors that were active in the transcription of the genes. The total number of transcription factors, and thus motifs, are however unknown and differ between species. In *Saccharomyces cerevisiae*, there are roughly 130 known transcription factors (Cherry et al., 1997) while the corresponding number is 1500 for *Arabidopsis thaliana* (Riechmann et al., 2000) and more than 2000 for *Homo sapiens* (Brivanlou and Darnell, 2002). Several methods for identification of enriched *cis*-regulatory

motifs are available (Bussemaker et al., 2001; Sharan et al., 2003a; Ettwiller et al., 2005) but there is still room for progress considering the consistently improving genome annotation (Paper 4) and availability of new types of data (ENCODE Project Consortium, 2007), e.g. chromatin prediction (Segal et al., 2006; Peckham et al., 2007).

The Gene Ontology (GO) project (Ashburner et al., 2000) is an effort to create a controlled vocabulary describing different biological processes, functions and entities. In particular, genes are annotated based on the location and function of their products. Assessing enrichments of certain GO terms is therefore a way to gain knowledge about a set of genes (Zeeberg et al., 2003). This approach has been especially successful for differentially expressed genes identified by microarrays (Stuart et al., 2003; Zimmermann et al., 2004). There are however non-trivial statistical aspects when testing overrepresentation of GO terms, due to the structure of the ontologies (Osier et al., 2004; Goeman and Buhlmann, 2007).

The field of bioinformatical tools designed for analysis of microarray results is vast. Other successful approaches include pathway analysis (Salomonis et al., 2007; Liu and Ringner, 2007), protein interaction networks (Breitkreutz et al., 2003), and *de novo cis*-regulatory motif identification (Eden et al., 2007; Gasch et al., 2004).

Summary of papers

Summary of Paper 1 - Weighted analysis of paired microarray experiments

In microarray experiments quality often varies both between samples and between arrays. In this paper, a model for analysis of paired microarray data using direct comparisions, is introduced. The model, named WAME, is an extension of a Bayesian model, originally suggested by Baldi and Long (2001) and later refined by Lönnstedt and Speed (2002) and Smyth (2004). In the WAME model, the assumption of independent samples is relaxed and variance heterogeneity and dependencies between the arrays are modelled using a covariance matrix.

More formally, assume that the experiment consists of n replicates and that a vector $\mathbf{X}_g = (X_{g1}, \ldots, X_{gn})$ is observed, consisting of the normalized $\log_2$ ratios for gene g, where $g = 1, \ldots, m$ and m is the number of genes on the arrays (typically >1000). Let Σ be an $n \times n$ covariance matrix and assume that

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

where μ_g is the expected value, c_g is a random gene specific scaling factor and α is a hyperparameter. Hence, we assume that the microarray data has a global covariance structure which, for each gene, is scaled appropriately. The aim of this model is to identify differentially expressed genes, i.e. genes with a difference in mRNA level between the tissue or cell samples. Thus, using our notation, we want to test the hypothesis

$$\begin{aligned} H_0 &: \text{ gene } g \text{ is not regulated } (\mu_g = 0) \\ H_A &: \text{ gene } g \text{ is regulated } (\mu_g \neq 0), \end{aligned} \tag{2}$$

for all genes $g = 1, \ldots, m$.

Derivation of estimators

To fit the model to a given microarray dataset, estimators of the unknown parameters need to be derived. As described in Section 4.1 in Paper 1, estimation of the covariance matrix Σ is nontrivial due to the gene specific scaling factor c_g. The approach used, is to assume that $\mu_g = 0$ for all genes and then remove the scaling by a transformation. If we let $\mathbf{U}_g = (U_{g1}, \ldots, U_{gn})$, where

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

the distribution of $(U_{g2}, \ldots, U_{gn})$ can be derived to

$$f_{U_{g2},\ldots,U_{gn} \mid \Sigma}(u_{g2}, \ldots, u_{gn}) = K \left|\Sigma\right|^{-1/2} \left[v_g^{\mathrm{T}} \Sigma^{-1} v_g\right]^{-n/2}, \tag{3}$$

where K is a constant independent of Σ and $v_g = (1, u_{g2}, \ldots, u_{gn})$. This distribution is, as expected, independent of c_g and a scale-free version of Σ, here denoted Σ^*, can be estimated from $\mathbf{U}_1, \ldots, \mathbf{U}_m$ using numerical maximum likelihood.

Estimators for the hyperparameter α and the unknown scale of Σ, denoted λ, are derived under the assumption that Σ^* is known. Define the statistic S_g as

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

where A is a an arbitrary $n-1 \times n$ matrix of full rank with each row sum equal to zero. Conditional on c_g, S_g/c_g can be shown to have a scaled χ^2-distribution with $n-1$ degrees on freedom. Thus, unconditionally, S_g has a beta prime distribution (scaled F-distribution) and hence, based on $S_1, \ldots, S_m$, both α and λ can be estimated by maximum likelihood. The parameters Σ and α are from now on assumed to be known.

Testing for differential expression

A likelihood ratio test for (2) is derived in Section 4.4 in Paper 1. The resulting test statistic can be shown to be

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

where $\bar{X}_g^w$ is the *weighted* mean value for gene g with weights defined by

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

Furthermore, under H_0, T_g is shown to follow a t-distribution with $n - 1 + 2\alpha$ degrees of freedom and T_g is therefore called the weighted moderated t-statistic.

Results on simulated data

Simulations were used to compare the proposed model to four other methods; average fold-change, ordinary t-statistic, Efron's penalized t-statistic (Efron et al., 2001) and Smyth's moderated t-statistic (Smyth, 2004). Data were generated according to the WAME model and a small proportion (10 %) of the genes were chosen to be regulated ($\mu_g \neq 0$). Under these settings, WAME performed substantially better compared to four other methods (Figure 1 and 2 in Paper 1).

Results on real data

The WAME model was also applied to three real datasets, one from an experiment using two-channel cDNA microarrays and two from experiments using one-channel Affymetrix microarrays in a paired setting. In all cases, the estimated covariance matrix contained heterogenous variances and high correlations. The resulting weights were shown to differ substantially, and some of the arrays had weights close to zero. In the polyp dataset (Section 6.2, Paper 1), one of the biopsies was considerably smaller than the others. The variance of the corresponding array was considerably larger than the variances for the other arrays indicating that the estimate of Σ contains biologically meaningful information.

Summary of Paper 2 - Quality optimised analysis of general paired microarray experiments

In Paper 1, a novel model for the analysis of microarray data from paired direct comparisions was suggested. In many situations there are, however, more suitable designs such as common references and time series. Accordingly, WAME needs to be generalized from direct comparisions to handle more general paired experimental designs.

The generalized model

To increase the flexibility and make WAME able to handle most of the existing types of paired experimental designs, the model is reparameterized as a linear model (Arnold, 1980). Assume that we perform an experiment with s different conditions ($s \geq 2$). For each gene $g = 1, \ldots, m$, 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 pair-wise differences of some of these conditions are measured, denoted by the vector

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

Let D be an $n \times s$ design matrix with rank p such that the expected value of $\mathbf{X}_g$ can be written as

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

Using this new parametrization, the generalized WAME model can be written as

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

where c_g is a gene specific scaling factor and α is a hyperparameter.

Inference in this generalized model is done by testing linear combinations of the parameter vector $\boldsymbol{\gamma}_g$. Let C be a contrast matrix of rank k and let $\boldsymbol{\delta}_g = C\boldsymbol{\gamma}_g$ be the differential

expression for the linear combinations in question. We assume that C is chosen such that $\boldsymbol{\delta}_g$ is testable. The hypotheses can then be written as

$$\begin{aligned} H_0 &: \boldsymbol{\delta}_g = \mathbf{0} \\ H_A &: \boldsymbol{\delta}_g \neq \mathbf{0}. \end{aligned} \tag{4}$$

Estimation of parameters

The estimation of the covariance matrix Σ can be performed analogously to the previous WAME model. As before, the assumption of no differential expression, i.e. $\boldsymbol{\mu}_g = \mathbf{0}$ for all g is needed. The hyperparameter α and the scale λ can also be estimated analogously. Indeed, there exists a statistic S_g (defined in Section 3.2 in Paper 2) such that

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

Hence, unconditionally, S_g has as before a beta prime distribution, which can be used to estimate α and the unknown scale λ using numerical maximum likelihood.

Inference in the generalized model

A likelihood ratio test for (4) was derived and the corresponding test statistic T can be found in Section 3.3 in Paper 2. It can be shown that T follows a F-distribution with k and $n - p + 2\alpha$ degrees of freedom under H_0.

For contrast matrices with $k = 1$, a weighted moderated t-statistic can be derived. If we define the weighted mean value $\bar{X}_g^w$ as

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

t-statistic can then be written as

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

Under H_0, T' follows a t-distribution with $n - p + 2\alpha$ degrees of freedom.

Results on real and simulated data

The generalized model was also evaluated on simulated data and was shown to perform better than average fold-change, the ordinary t-statistics and the moderated t-statistic. WAME was also applied to two real datasets and different variances and correlations were identified.

Summary of Paper 3 - Weighted analysis of microarray experiments

In Paper 1 and 2, the estimation of the covariance matrix Σ is based on the assumption that $\boldsymbol{\mu}_g = \mathbf{0}$ for all genes, an assumption only realistic for paired microarray data. In this study we relax this assumption, and make WAME applicable to experiments using one-channel microarray data.

The model assumed is the same as in the previous paper, i.e.

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

where $\mathbf{X}_g$ is the vector of observations, $\boldsymbol{\gamma}_g$ is the parameter vector, $\boldsymbol{\mu}_g$ is the expected value of $\mathbf{X}_g$, D is a design matrix, Σ is a covariance matrix and c_g a gene specific scaling factor. As before, the aim of the model is to identify differentially expressed genes, i.e., to test

$$\begin{aligned} &\mathrm{H}_0 : \boldsymbol{\delta}_g = \mathbf{0} \\ &\mathrm{H}_A : \boldsymbol{\delta}_g \neq \mathbf{0} \ . \end{aligned} \tag{6}$$

where $\boldsymbol{\delta}_g = C\,\boldsymbol{\gamma}_g$ ($\boldsymbol{\delta}_g$ is assumed to be testable).

Estimation of Σ for non-paired data

We will now describe how Σ can be estimated under the less strict assumption that $\boldsymbol{\delta}_g = 0$ for $g = 1, \ldots, m$. Let

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

where $\tilde{\boldsymbol{\mu}}_g^0$ is any linear estimator of $\boldsymbol{\mu}_g$ which is unbiased under H_0. Under the null hypothesis, the expected vector $\mathbf{Y}_g$ is zero for all genes. Thus, the covariance structure matrix of $\mathbf{Y}_g$, Σ_Y can be estimated according to the method described in Paper 1. If, for example, we have two different conditions that we want to compare,

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

we can set $\tilde{\boldsymbol{\mu}}_g^0$ to be the gene-wise mean value over all arrays. Thus,

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

Furthermore, in this paper we also show that the t- and F-statistics described the in Paper 2 are *(i)* a likelihood ratio test of (6) and *(ii)* only dependent on Σ_Y and not the full covariance matrix Σ.

Simulations based on resampling

The statistical power of WAME is evaluated by a resample based simulation study. Given a microarray dataset with sufficient biological replicates from a single condition, two subgroups, with four arrays in each, were sampled repeatedly. A small proportion of the genes were chosen to be regulated and had a signal added to one of the groups. WAME and the other methods LIMMA, weighted LIMMA and ordinary t-statistic were applied to the data and the number of correctly identified genes was counted. The power of WAME was shown to be higher, especially for genes with low expression (Figure 6 in Paper 3). Furthermore, the p-values generated by WAME were also shown to be more accurate while the p-values from the other methods typically were too optimistic (Figure 3 and 4 in Paper 3).

Summary of Paper 4 - Evolutionary forces act on promoter length: assessment of enriched *cis*-regulatory elements.

There is a plethora of bioinformatical tools and methods to infer biological knowledge from a list of genes regulated in a microarray experiment. One popular approach is to search for enrichments of *cis*-regulatory motifs within the promoters of the genes and hence deduce which transcription factors that are involved in their regulation. Although the increasing preciseness of genomic data in general improves these methods, they will on certain occasions perform inaccurately.

In this paper we show that the length of the promoter and the function of the gene are related. We also show that this can lead to false positives when assessing enrichments of *cis*-regulatory motifs using common methods, such as Fischer's exact test (Agrestri, 2002) and logistic regression (Hosmer and Lemeshow, 2000). Finally we propose a new regression-type method, which includes the promoter lengths as a critical element.

Evolutionary forces act on promoter length

The simplest definition of a gene's promoter region is the entire intergenic region located upstream of the open reading frame (ORF). Applying this definition to the genome of *Saccharomyces cerevisiae* resulted in a median promoter length of 455 base pairs (bp) (Figure 1 in Paper 4). Furthermore, we show that genes with a potentially complex regulatory pattern, such as genes involved in many different types of stress, have substantially longer promoters. For example, the common environmental response (CER) genes, defined by Causton et al. (2001), have a median promoter length of 552, more than 20% longer than for the other genes in the genome. Moreover, genes with a less complex regulatory pattern, such as essential genes, have short promoters (see Table 1 in Paper 4 for more examples).

These phenomena are also present in the fungi *Schizosaccharomyces pombe* and the plant *Arabidopsis thaliana* suggesting that evolutionary forces act on promoter length.

Finding enriched *cis*-regulatory motifs

One of the most frequently used methods to test for overrepresented *cis*-regulatory motifs is Fisher's exact test (Hughes et al., 2000; Sharan et al., 2003b; Nelander et al., 2005). For a given set of genes, this test is performed by observing the number of genes in the set that has the motif present. Under the assumption of independence, this number has a known distribution (hypergeometic) and hence significance can be derived.

Various kinds of regression models have also been suggested for identification of enriched motifs. In particular, logistic regression is commonly used (Copley, 2005; Keles et al., 2004). In contrast to Fisher's exact test, significance in regression models is derived by large sample approximations.

Their popularity notwithstanding, both the hypergeometric test and the logistic regression model assume, directly or indirectly, that there is no difference in promoter length between the sets of genes. We therefore proposed a new method to find enrichments of *cis*-regulatory motifs in the promoters of a set of genes. The procedure extends logistic regression and includes the promoter lengths as a critical element. The method is formulated as a generalized additive model (GAM) with a logit link (Hastie and Tibshirani, 1990; Wood, 2006). In general, GAMs can be seen as an extension of generalized linear models (GLMs) (McCullagh and Nelder, 1983) where non-linear relations of covariates are modelled non-parametrically by smoothing functions.

Assume that there are N genes in the genome and that we want to search for enrichment of a fixed motif in a subset A, consisting of n genes. For $g = 1, \ldots, N$, let y_g be a binary 0-1 valued variable indicating whether the motif is present in the promoter of g. Let x_g be another binary variable indicating if gene g is in the set A. The regression model can then be stated as

$$\log \frac{\mathbb{P}(y_g = 1)}{\mathbb{P}(y_g = 0)} = \alpha + \beta x_g + f(l_g),$$

where α and β are coefficients and f is an unknown smooth function. Our main objective is to test if there are more motifs among the genes in A compared to the remaining genes. This will be done by testing the hypothesis

$$\begin{aligned} &\mathrm{H}_0 : \alpha = 0 \\ &\mathrm{H}_A : \alpha > 0. \end{aligned} \tag{7}$$

Observe that if f is removed, this model becomes an ordinary logistic regression model. In such a case, iterated reweighted least squares (IRLS) is typically used to find numerical maximum likelihood estimates of the coefficients α and β. For GAMs, this approach is extended and an outer iteration, which estimate the smoothness parameter λ of f, is added to the procedure. In the IRLS step λ is kept fixed and f is estimated using penalized

regression splines. For fixed values of α, β and f, λ is then estimated by minimizing the mean square error of the model. These two steps are repeated until convergence.

It is possible to show that $\hat{\beta}$, the estimate of β, is approximately normally distributed when n is large enough. The estimate is however biased, which makes confidence intervals troublesome. Nevertheless, under the null hypothesis, $\mathbb{E}[\hat{\beta}] = 0$, which makes it possible to test (7) using standard procedures. Further details regarding GAMs are available in Chapter 4 in Wood (2006).

The proposed model is evaluated in a simulation study and is shown to generate fewer false positives than Fisher's exact test and logistic regression. The model also identifies several relevant enriched *cis*-regulatory elements among genes induced under arsenic stress (see Paper 5).

Summary of Paper 5 - Quantitative transcriptome, proteome and sulfur metabolite profiling of the *Saccharomyces cerevisiae* response to arsenite

Arsenic is toxic but still used as a drug against a number of diseases such as leukemia. Since this metalloid element is ubiquitously present in the environment, all organisms have evolved a defense system protecting them against exposure. However, little is known about these defense mechanisms and their function. In this paper, we explored how *Saccharomyces cerevisiae* (Baker's yeast) responds to arsenic exposure by measuring gene expression using microarrays.

Experimental setup

Two-channel spotted cDNA microarrays, containing all of the roughly 6,000 genes in yeast, were used to identify differences in mRNA levels between control and arsenite (As^{3+}) treated yeast. Measurements were made at five different time points after exposure (0, 15 min, 30 min, 60 min and 120 min) and for two different concentrations (0.2 and 1 mM As^{3+}). Additionally, the effects of arsenite were also studied in two gene deletion mutants ($\Delta YAP1$ and $\Delta MET4$). All experiments were repeated at least three times, resulting in 30 arrays in total.

Analysis

The microarray data was analyzed using the R-package LIMMA (R Development Core Team, 2007; Gentleman et al., 2004). No background correction was performed, due to the high correlation between estimates of the background and the spot intensities. Intensity dependent trends were removed by lowess normalization (Yang et al., 2002), where a robust regression line is fitted between the M- and A-values and then subtracted from the M-values. Regulated genes were identified using the empirical Bayes model with the corresponding moderated t-statistic described by Smyth (2004).

Enrichment analysis of all transcription factors available at the Saccharomyces Genome Database (SGD) (Cherry et al., 1997, 1998) was performed among genes with high difference in mRNA levels (absolute M-value greater than 2). Since the promoter lengths of these genes were substantially longer than the other genes in the genome, an early version of the procedure described in Paper 4 was used.

Results

Many genes had their expression stimulated by arsenite. Among those were genes involved in arsenic detoxification and oxidative stress defense as well as genes encoding components of the sulfur assimilation and GSH biosynthesis pathways. Transcription factors that were enriched among the regulated genes include Yap1 and Msn2/4, which both are important in the regulation of stress responses in yeast. Furthermore, the transcription factors Cbf1, Met31 and Met32, which controls the expression of sulfur assimilation and GSH biosynthesis encoding genes, were also enriched.

Summary of Paper 6 - Sensitive and robust gene expression changes in fish exposed to estrogen - a microarray approach

An environmental biomarker is a biological response to a chemical or chemicals which gives a measure of exposure (Peakall, 1994).For estrogens, important contributors to the feminization of fish downstream from sewage treatment works, vitellogenin (VTG) has long been a well established biomarker. However, recent studies performed on fish (Örn et al., 2003; Parrott and Blunt, 2005) suggest that effects of estrogen can be seen at concentrations which are not sufficient to give rise to a measurable VTG response. Hence, more sensitive biomarkers could be useful.

A good biomarker should in general be specific, sensitive and robust. Specificity and sensitivity means that the biomarker should respond to a single substance even at low concentrations. Robustness means, for example, that it should be measurable at different temperatures, different exposure times, by different techniques, in different labs and preferably also in different species.

The main objective of this study was to use microarrays to screen for novel, sensitive and robust biomarkers for estrogenic exposure in male juvenile rainbow trout (*Oncorhynchus mykiss*). To identify sensitive gene responses, fish exposed to high and low concentrations of estrogen were compared to control fish. Additionally, a meta-analysis, using three other recently published microarray datasets, were performed to identify robust responses. The result from these two approaches were then combined to identify candidate biomarkers.

Description of the microarray

Sequence information about the transcriptome of the species of interest is vital for microarray based studies. For species without a fully sequenced genome such as rainbow

trout, spotted cDNA microarrays are therefore used, since they only need a sequenced cDNA library to be analyzed. The microarray used in this study was manufactured by the Consortium for Genomic Research on All Salmonids Program (cGRASP) (Rise et al., 2004) and contains 13,000 cDNAs from Atlantic salmon (*Salmon salar*) and 2,500 cDNAs from rainbow trout. The lack of rainbow trout cDNAs on the chip was not believed to be a problem since the Atlantic salmon and rainbow trout are very similar genome-wise and cross-species use of the array work satisfactory (Rise et al., 2004).

Experimental setup

Rainbow trouts, divided into three aquaria, were exposed to none, a low and a high concentration of a estrogen. The low concentration was chosen to correspond to the levels observed in water downstream of sewage treatment works. After two weeks, the fish were sacrificed and mRNA was extracted from the liver and hybridized to microarrays. Each fish exposed to the estrogen was paired against a control fish based on individual weights and lengths and both were hybridized to the microarray. In total, four microarrays for the high concentration and eight microarrays for the low concentration were used.

Analysis

The microarray data were analyzed using the R package LIMMA (R Development Core Team, 2007; Gentleman et al., 2004). The data was normalized using lowess (Yang et al., 2002) and no background correction was performed. Regulated genes were identified using the moderated t-statistic (Smyth, 2004).

Identification of robust biomarkers using meta-analysis

To identify robust gene responses, our results were compared to three other microarray experiments on estrogen exposed fish (Table 2 in Paper 6). Performing such a meta-analysis was however non-trivial due to the fact that two different species and three different microarray platforms were used (Jarvinen et al., 2004). Furthermore, none of the species had fully sequenced genomes, which further complicated the comparison.

The approach used in this study was to use *Danio rerio* (zebrafish) as a reference species. The zebrafish had its full genome sequenced and more than 60% of the genes were known (Hubbard et al., 2007), making it, by far, the best annotated fish genome-wise. Regulated genes from the four experiments were mapped to the zebrafish transcriptome using tBLASTx (Altschul et al., 1997). The zebrafish genes were then ranked according to the number of studies with hits. Finally, these hits were clustered together based on sequence similarity of their proteins.

Results

As expected, VTG was highly up-regulated in fish exposed to the high concentration of the estrogen while no regulation could be seen in fish exposed to the low concentration (Figure 1, Paper 6). Furthermore, other known estrogen responsive gene were also regulated (zona pellucida proteins). These genes were, in addition, verified to be robust by the meta-analysis (Figure 2, Paper 6). Candidates for novel, sensitive and robust biomarkers were found by combining genes that were regulated in both high and low concentration with genes found robust by the meta-analysis. One such gene was nucleoside disphosphate kinase (nm23), which was also verified to be up-regulated by quantitative PCR. However, nm23 needs to be further evaluated to find if it is useful as a biomarker.

Additional papers

1. Sven Nelander, Erik Larsson, Erik Kristiansson, Robert Månsson, Olle Nerman, Magnus Sigvardsson, Petter Mostad and Per Lindahl (2005). Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals, *BMC Genomics*, **6**(68).

2. R Henrik Nilsson, Erik Kristiansson, Martin Ryberg and Karl-Henrik Larsson (2005). Approaching the taxonomic affiliation of unidentified sequences in public databases - an example from the mycorrhizal fungi, *BMC Bioinformatics*, **6**(88).

3. R Henrik Nilsson, Martin Ryberg, Erik Kristiansson, Kessy Abarenkov, Karl-Henrik Larsson and Urmas Kõljalg (2006). Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective, *PLoS ONE*, **1**(1): e59.

4. Martin Ryberg, R Henrik Nilsson, Erik Kristiansson, Mats Töpel, Stig Jacobsson, Ellen Larsson (2007). Mining metadata from unidentified ITS sequences in GenBank: a case study from Inocybe (Basidiomycota), *to appear in BMC Evolutionary Biology*.

5. R Henrik Nilsson*, Erik Kristiansson*, Martin Ryberg, Nils Hallenberg, Karl-Henrik Larsson (2007). Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequences databases, *submitted*.

6. Olga Kourtchenko, Erik Kristiansson, Andreas Czihal, Helmut Bäumlein, Mats Ellerström (2007). Transcriptional control of defense responses to AvrRpm1 effector protein in *Arabidopsis*: a microarray study, *submitted*.

* equal contribution.

References

Affymetrix (2007). Genechip operating software. http://www.affymetrix.com.

Agrestri, A. (2002). *Categorical Data Analysis*. John Wiley & Sons.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402.

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

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29.

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

Åstrand, M., Mostad, P., and Rudemo, M. (2007a). Empirical bayes models for multiple probe type arrays at the probe level. Technical report, Chalmers University of Technology and Göteborg University, Department of Mathematical Statistics, http://www.math.chalmers.se/Math/Research/Preprints/2007/32.pdf.

Åstrand, M., Mostad, P., and Rudemo, M. (2007b). Improved covariance matrix estimators for weighted analysis of microarray data. *Journal of Computational Biology*. To appear.

Baird, D., Johnstone, P., and Wilson, T. (2004). Normalization of microarray data using a spatial mixed model analysis which includes splines. *Bioinformatics*, 20(17):3196–3205.

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

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., and Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*, 35(Database issue):760–765.

Baumgartner, W. A. J., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., and Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):41–48.

BioDiscovery (2007). Imagene 6. http://www.biodiscovery.com.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Breitkreutz, B.-J., Stark, C., and Tyers, M. (2003). Osprey: a network visualization system. *Genome Biol*, 4(3):R22.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. *Nat Genet*, 30(1):29–30.

Bretz, F., Landgrebe, J., and Brunner, E. (2005). Design and analysis of two-color microarray experiments using linear models. *Methods Inf Med*, 44(3):423–430.

Brivanlou, A. H. and Darnell, J. E. J. (2002). Signal transduction and the control of gene expression. *Science*, 295(5556):813–818.

Bussemaker, H. J., Li, H., and Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nat Genet*, 27(2):167–171.

Causton, H. C., Ren, B., Koh, S. S., Harbison, C. T., Kanin, E., Jennings, E. G., Lee, T. I., True, H. L., Lander, E. S., and Young, R. A. (2001). Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell*, 12(2):323–337.

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Res*, 26(1):73–79.

Cherry, J. M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., Mortimer, R. K., and Botstein, D. (1997). Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, 387(6632 Suppl):67–73.

Chi, Y.-Y., Ibrahim, J. G., Bissahoyo, A., and Threadgill, D. W. (2007). Bayesian hierarchical modeling for time course microarray experiments. *Biometrics*, 63(2):496–504.

Chu, T. M., Weir, B., and Wolfinger, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments. *Math Biosci*, 176(1):35–51.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of The American Statistical Association*, 74:829–836.

Cleveland, W. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, 35:54.

Copley, R. R. (2005). The EH1 motif in metazoan transcription factors. *BMC Genomics*, 6:169.

CSIRO (2007). Spot: software for dna microarray image analysis. http://www.csiro.au/products/ps1ry.html.

Cuperlovic-Culf, M., Belacel, N., Culf, A. S., and Ouellette, R. J. (2006). Data analysis of alternative splicing microarrays. *Drug Discov Today*, 11(21-22):983–990.

Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., Maier, D., Matese, J. C., Nitzberg, M., Wymore, F., Zachariah, Z. K., Brown, P. O., Sherlock, G., and Ball, C. A. (2007). The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res*, 35(Database issue):766–770.

Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol*, 3(3):e39.

Edwards, D. (2003). Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*, 19(7):825–833.

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

Eisen Lab (2007). Scanalyze. http://rana.lbl.gov/EisenSoftware.htm.

ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

Ettwiller, L., Paten, B., Souren, M., Loosli, F., Wittbrodt, J., and Birney, E. (2005). The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biol*, 6(12):R104.

Febit Biomed (2007). Geniom software. http://www.geniom.com.

Fujita, A., Sato, J. R., Rodrigues, L. d. O., Ferreira, C. E., and Sogayar, M. C. (2006). Evaluating different methods of microarray data normalization. *BMC Bioinformatics*, 7:469.

Gasch, A. P., Moses, A. M., Chiang, D. Y., Fraser, H. B., Berardini, M., and Eisen, M. B. (2004). Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol*, 2(12):e398.

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., Li, F. L. C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Res*, 17(6):669–681.

Goeman, J. J. and Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.

Gunnarsson, L., Kristiansson, E., Förlin, L., Nerman, O., and Larsson, D. G. J. (2007). Sensitive and robust gene expression changes in fish exposed to estrogen–a microarray approach. *BMC Genomics*, 8:149.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalizzed Additive Models*. Chapman & Hall.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley & Sons.

Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A., and Birney, E. (2007). Ensembl 2007. *Nucleic Acids Res*, 35(Database issue):610–617.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:96–104.

Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, 296(5):1205–1214.

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

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

Jarvinen, A.-K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallioniemi, O.-P., and Monni, O. (2004). Are data from different gene expression microarray platforms comparable? *Genomics*, 83(6):1164–1168.

Keles, S., van der Laan, M. J., and Vulpe, C. (2004). Regulatory motif finding by logic regression. *Bioinformatics*, 20(16):2799–2811.

Kerr, M. K. (2003). Linear models for microarray data analysis: hidden similarities and differences. *J Comput Biol*, 10(6):891–901.

Kooperberg, C., Fazzio, T. G., Delrow, J. J., and Tsukiyama, T. (2002). Improved background correction for spotted DNA microarrays. *J Comput Biol*, 9(1):55–66.

Kristiansson, E., Sjögren, A., Rudemo, M., and Nerman, O. (2006). Quality optimised analysis of general paired microarray experiments. *Stat Appl Genet Mol Biol*, 5:Article10.

Kristiansson, E., Thorsen, M., Tamas, M. J., and Nerman, O. (2007). Evolutionary forces act on promoter length: assessment or enriched cis-regulartory motifs. *submitted*.

Kristiansson, E and Sjögren, A and Rudemo, M and Nerman, O (2005). Weighted analysis of paired microarray experiments. *Stat Appl Genet Mol Biol*, 4:Article30.

Leung, Y. F. and Cavalieri, D. (2003). Fundamentals of cDNA microarray data analysis. *Trends Genet*, 19(11):649–659.

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

Li, C. and Wong, W. H. (2001b). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol*, 2(8):RESEARCH0032.

Li, W. and Ying, X. (2006). Mprobe 2.0: computer-aided probe design for oligonucleotide microarray. *Appl Bioinformatics*, 5(3):181–186.

Liu, Y. and Ringner, M. (2007). Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis. *Genome Biol*, 8(5):R77.

Lönnstedt, I. and Britton, T. (2005). Hierarchical Bayes models for cDNA microarray gene expression. *Biostatistics*, 6(2):279–291.

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

Marton, M. J., DeRisi, J. L., Bennett, H. A., Iyer, V. R., Meyer, M. R., Roberts, C. J., Stoughton, R., Burchard, J., Slade, D., Dai, H., Bassett, D. E. J., Hartwell, L. H., Brown, P. O., and Friend, S. H. (1998). Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med*, 4(11):1293–1301.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1):374–378.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Chapman & Hall.

McGee, M. and Chen, Z. (2006). Parameter estimation for the exponential-normal convolution model for background correction of affymetrix GeneChip data. *Stat Appl Genet Mol Biol*, 5:Article24.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nat Genet*, 30(1):13–19.

Molecular Devices (2007). Genepix. http://www.moleculardevices.com.

Nelander, S., Larsson, E., Kristiansson, E., Mansson, R., Nerman, O., Sigvardsson, M., Mostad, P., and Lindahl, P. (2005). Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals. *BMC Genomics*, 6(1):68.

Örn, S., Holbech, H., Madsen, T. H., Norrgren, L., and Petersen, G. I. (2003). Gonad development and vitellogenin production in zebrafish (Danio rerio) exposed to ethinylestradiol and methyltestosterone. *Aquat Toxicol*, 65(4):397–411.

Osier, M. V., Zhao, H., and Cheung, K.-H. (2004). Handling multiple testing while interpreting microarrays with the Gene Ontology Database. *BMC Bioinformatics*, 5:124.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., and Brazma, A. (2007). ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, 35(Database issue):747–750.

Parrott, J. L. and Blunt, B. R. (2005). Life-cycle exposure of fathead minnows (Pimephales promelas) to an ethinylestradiol concentration below 1 ng/L reduces egg fertilization success and demasculinizes males. *Environ Toxicol*, 20(2):131–141.

Peakall, D. B. (1994). The role of biomarkers in environmental assessment (1). Introduction. *Ecotoxicology*, 3:157–160.

Peckham, H. E., Thurman, R. E., Fu, Y., Stamatoyannopoulos, J. A., Noble, W. S., Struhl, K., and Weng, Z. (2007). Nucleosome positioning signals in genomic DNA. *Genome Res*, 17(8):1170–1177.

Peytavi, R., Liu-Ying, T., Raymond, F. R., Boissinot, K., Bissonnette, L., Boissinot, M., Picard, F. J., Huletsky, A., Ouellette, M., and Bergeron, M. G. (2005). Correlation between microarray DNA hybridization efficiency and the position of short capture probe on the target nucleic acid. *Biotechniques*, 39(1):89–96.

Pozhitkov, A., Noble, P. A., Domazet-Loso, T., Nolte, A. W., Sonnenberg, R., Staehler, P., Beier, M., and Tautz, D. (2006). Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. *Nucleic Acids Res*, 34(9):e66.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat Genet*, 32 Suppl:496–501.

R Development Core Team (2007). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, http://www.R-project.org.

Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O. J., Samaha, R. R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J. Z., Ghandehari, D., Sherman, B. K., and Yu, G. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, 290(5499):2105–2110.

Rise, M. L., von Schalburg, K. R., Brown, G. D., Mawer, M. A., Devlin, R. H., Kuipers, N., Busby, M., Beetz-Sargent, M., Alberto, R., Gibbs, A. R., Hunt, P., Shukin, R., Zeznik, J. A., Nelson, C., Jones, S. R. M., Smailus, D. E., Jones, S. J. M., Schein, J. E., Marra, M. A., Butterfield, Y. S. N., Stott, J. M., Ng, S. H. S., Davidson, W. S., and Koop, B. F. (2004). Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Res*, 14(3):478–490.

Ritchie, M., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*.

Rouillard, J.-M., Zuker, M., and Gulari, E. (2003). OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res*, 31(12):3057–3062.

Salomonis, N., Hanspers, K., Zambon, A. C., Vranizan, K., Lawlor, S. C., Dahlquist, K. D., Doniger, S. W., Stuart, J., Conklin, B. R., and Pico, A. R. (2007). GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, 8:217.

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

Schulze, A. and Downward, J. (2001). Navigating gene expression using microarrays–a technology review. *Nat Cell Biol*, 3(8):190–195.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., Wang, J.-P. Z., and Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778.

Sharan, R., Ovcharenko, I., Ben-Hur, A., and Karp, R. M. (2003a). CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, 19 Suppl 1:283–291.

Sharan, R., Ovcharenko, I., Ben-Hur, A., and Karp, R. M. (2003b). CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, 19 Suppl 1:283–291.

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

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

Smyth, G. K. and Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, 31(4):265–273.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lonning, P., and Borresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19):10869–10874.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–3297.

Spring, J. (2002). Genome duplication strikes back. *Nat Genet*, 31(2):128–129.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255.

Thorsen, M., Lagniel, G., Kristiansson, E., Junot, C., Nerman, O., Labarre, J., and Tamas, M. J. (2007). Quantitative transcriptome, proteome, and sulfur metabolite profiling of the Saccharomyces cerevisiae response to arsenite. *Physiol Genomics*, 30(1):35–43.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121.

Urakawa, H., El Fantroussi, S., Smidt, H., Smoot, J. C., Tribou, E. H., Kelly, J. J., Noble, P. A., and Stahl, D. A. (2003). Optimization of single-base-pair mismatch discrimination in oligonucleotide microarrays. *Appl Environ Microbiol*, 69(5):2848–2856.

Wang, J., Ma, J. Z., and Li, M. D. (2004). Normalization of cDNA microarray data using wavelet regressions. *Comb Chem High Throughput Screen*, 7(8):783–791.

Welle, S., Brooks, A. I., and Thornton, C. A. (2002). Computational method for reducing variance with Affymetrix microarrays. *BMC Bioinformatics*, 3:23.

Wernersson, R. and Nielsen, H. B. (2005). OligoWiz 2.0–integrating sequence feature annotation into the design of microarray probes. *Nucleic Acids Res*, 33(Web Server issue):611–615.

Wodicka, L., Dong, H., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat Biotechnol*, 15(13):1359–1367.

Wood, S. N. (2006). *Generalized Additive Models*. Chapman & Hall/CRC.

Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielser, H. B., Saxild, H.-H., Nielsen, C., Brunak, S., and Knudsen, S. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol*, 3(9):research0048.

Wu, Z., Irizarry, R., Gentleman, R., Murillo, F., and F., S. (2004). A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 99.

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

Yang, Y. H., Buckley, M. J., and Speed, T. P. (2001). Analysis of cDNA microarray images. *Brief Bioinform*, 2(4):341–349.

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., and Weinstein, J. N. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28.

Zhou, Y. and Abagyan, R. (2002). Match-only integral distribution (MOID) algorithm for high-density oligonucleotide array analysis. *BMC Bioinformatics*, 3:3.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol*, 136(1):2621–2632.

Paper 1

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-

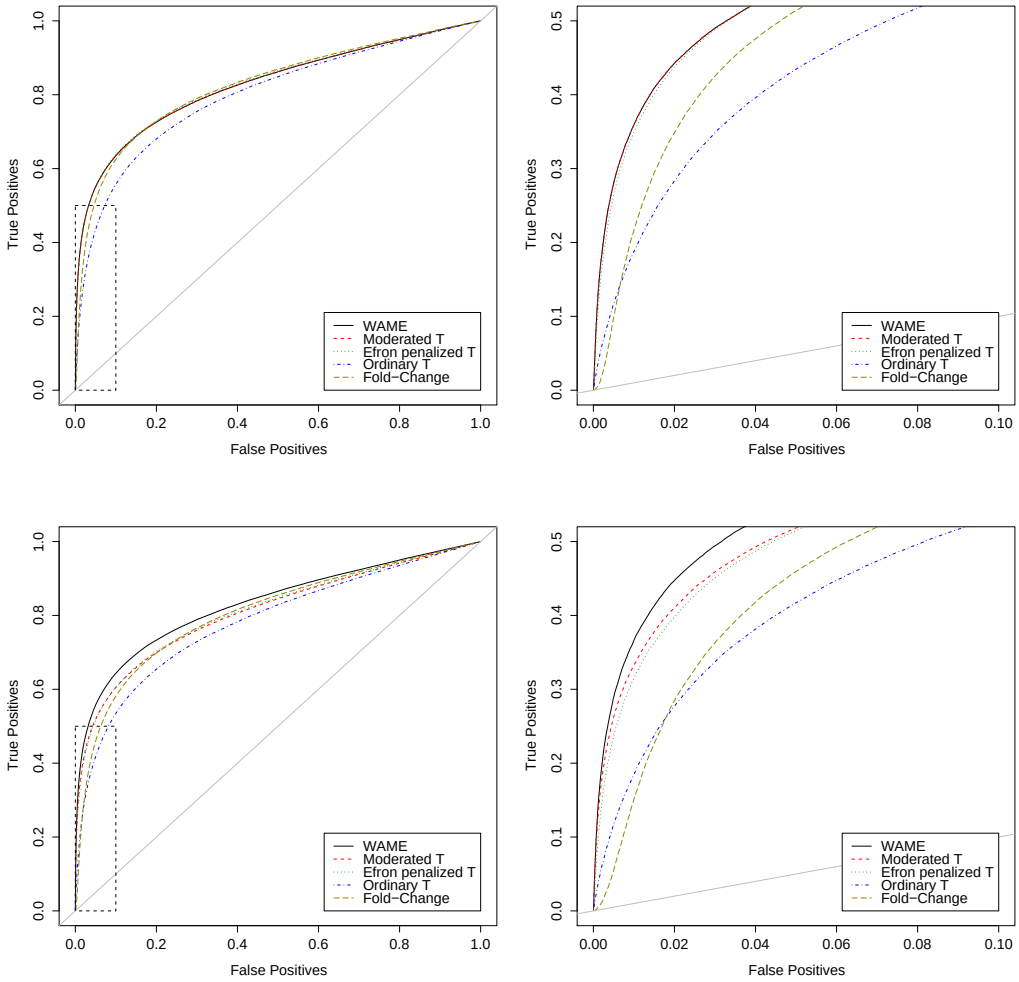


Figure 1: ROC curves from simulated data. The pair at the top, from the first case, show the performance of the evaluated methods on data with equal variances of 1 for all replicates and no correlations. The pair at the bottom, from the second case, analogously show the performance on data with different 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.

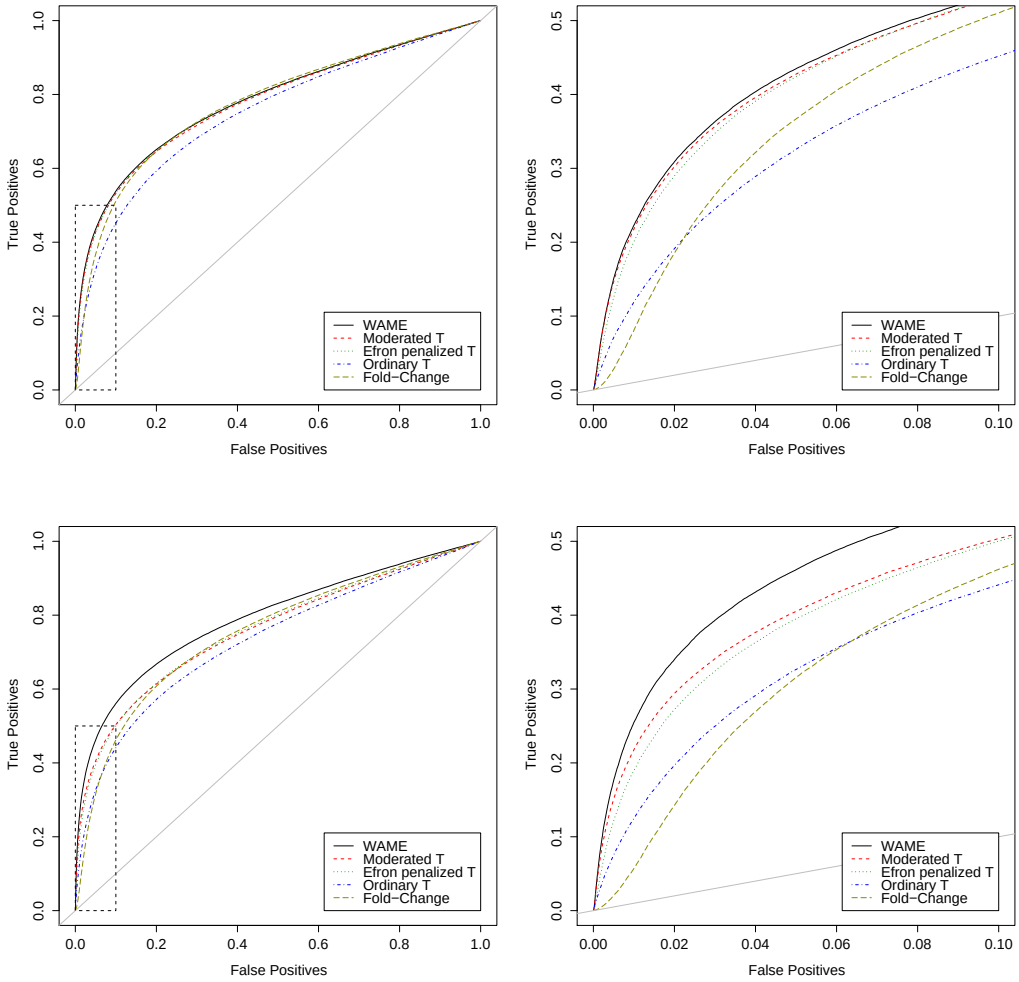


Figure 2: ROC curves from simulated data. The pair at the top, from the third case, show the performance of the evaluated methods on data with equal variances of 1 for all replicates and correlations of 0, 0.2 and 0.4, placed according to (13). The pair at the bottom, from the fourth case, analogously 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

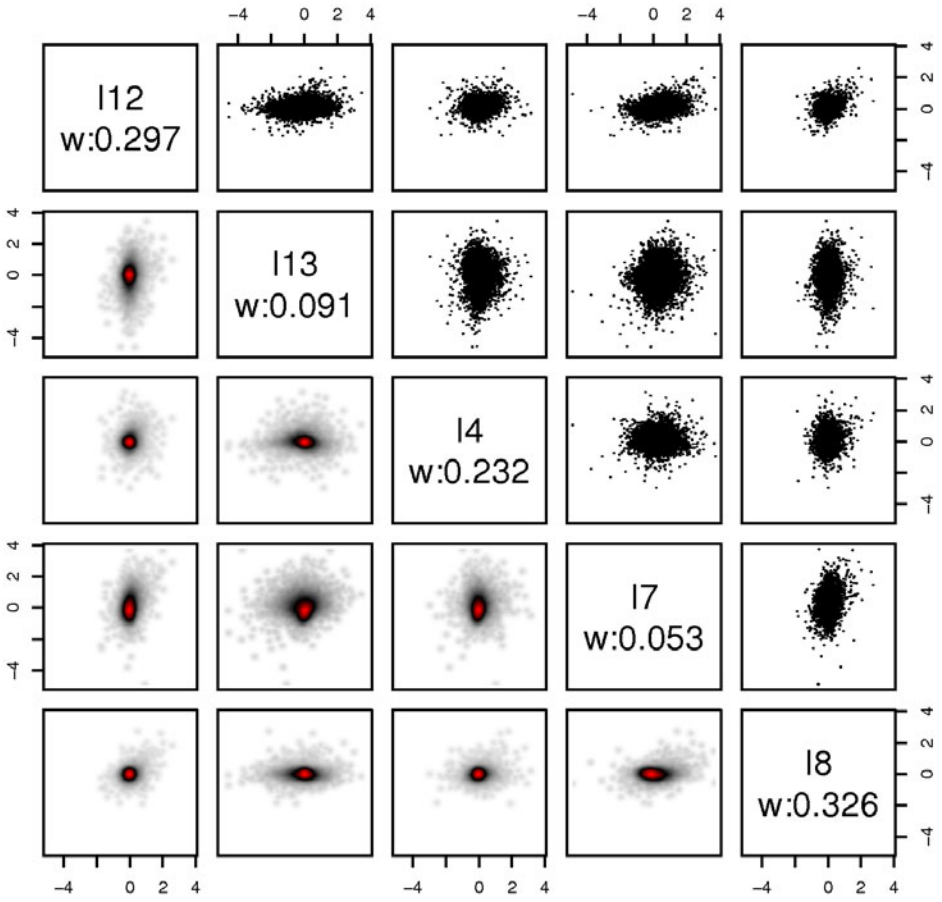


Figure 3: Pair-wise plots of the $\log_2$-ratios of the patients in the Cardiac dataset. The plots to the lower-left show two-dimensional kernel density estimates of the distribution of $\log_2$-ratios in each pair of patients. This provides information in the central areas where the corresponding scatterplots 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.

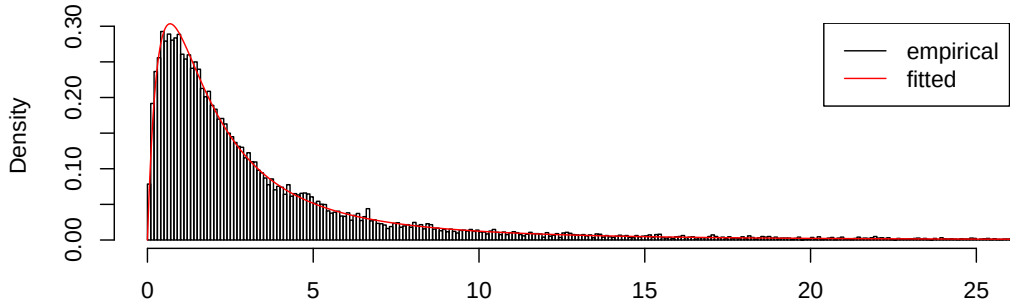


Figure 4: Empirical distribution of S_g in the Cardiac dataset, together with the density of S_g given $\alpha = 1.92$.

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

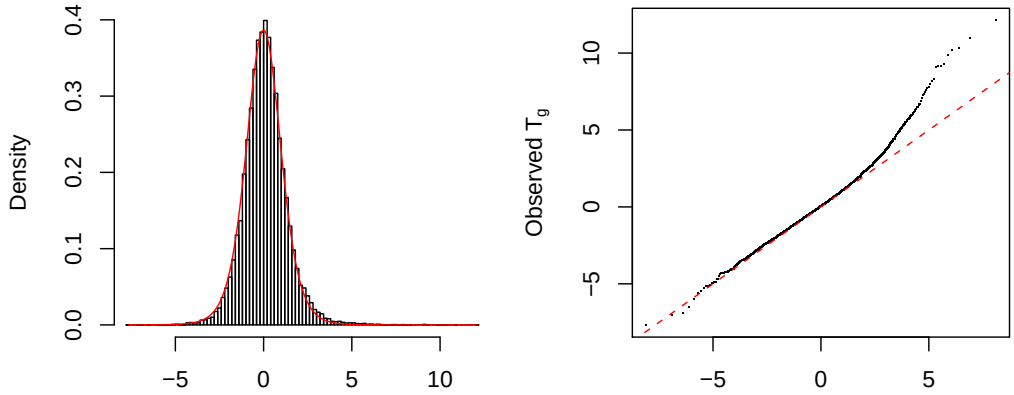


Figure 5: To the left, a histogram of the observed T_g-values together with the density of the null distribution (in red), in the Cardiac dataset. To the right, a quantile-quantile plot where the observed values of T_g are plotted against the quantiles of T_g under the null hypothesis. The central part 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

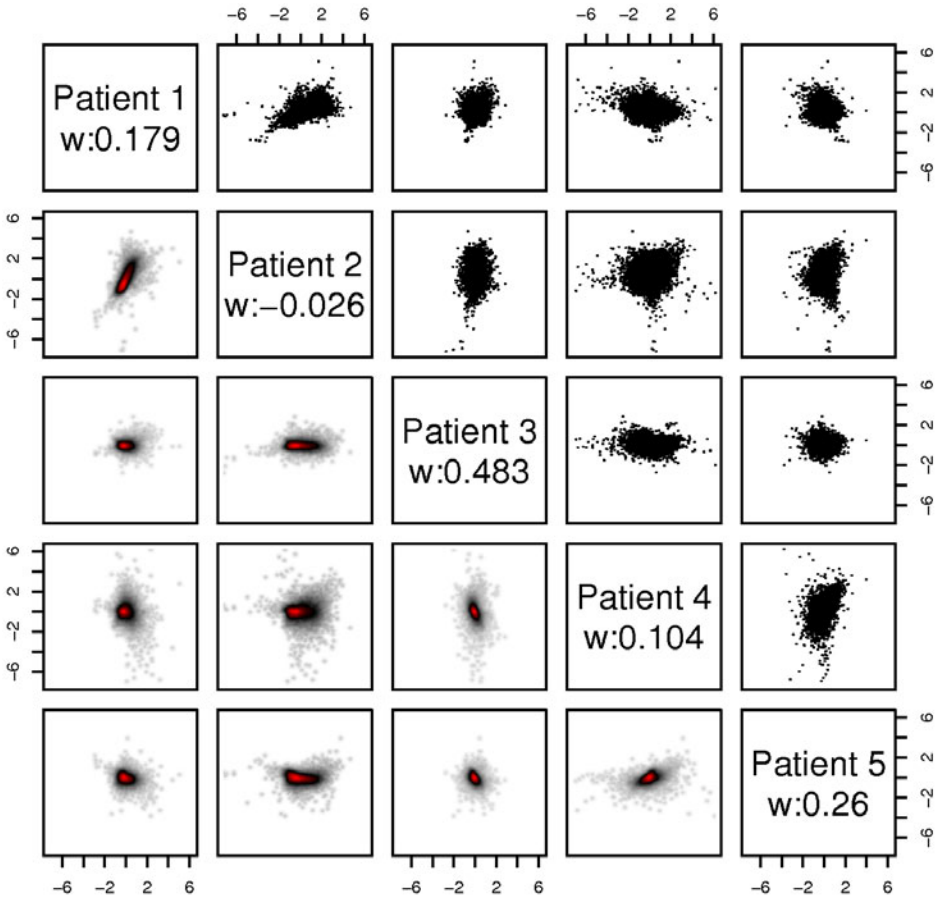


Figure 6: Pair-wise plots of the $\log_2$-ratios of the patients in the Polyp dataset. The plots to the lower-left show two-dimensional kernel density estimates of the distribution of $\log_2$-ratios in each pair of patients. This provides information in the central areas where the corresponding scatterplots 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).

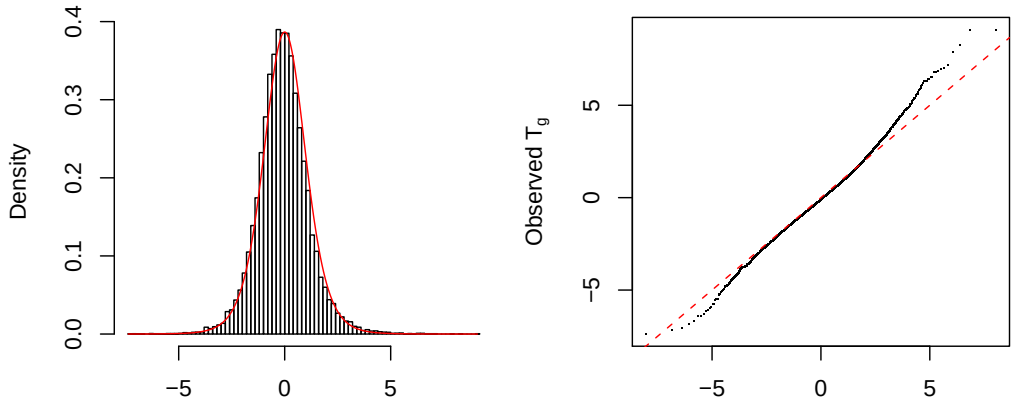


Figure 7: To the left, a histogram of the observed T_g-values together with the density of the null distribution (in red), in the Polyp dataset. To the right, a quantile-quantile plot where the observed values of T_g are plotted against the quantiles of T_g under the null hypothesis. The central part 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

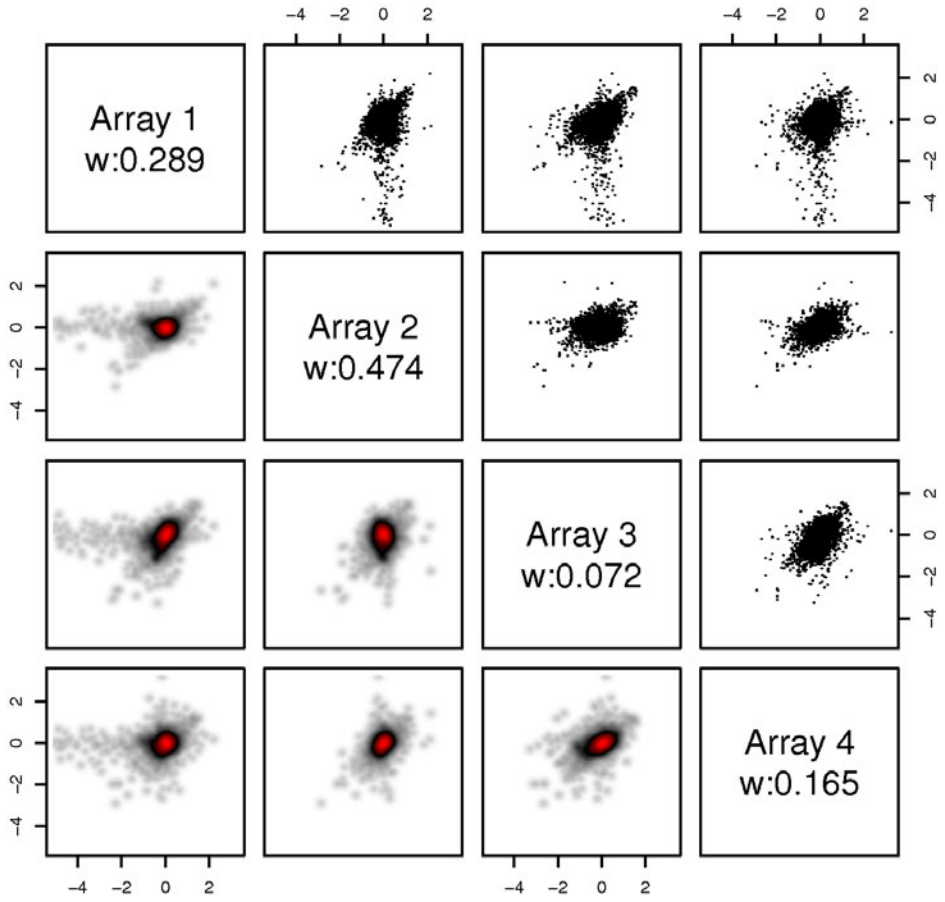


Figure 8: Pair-wise plots of the $\log_2$-ratios of the arrays in the Swirl dataset. The plots to the lower-left show two-dimensional kernel density estimates of the distribution of $\log_2$-ratios in each pair of patients. This provides information in the central areas where the corresponding scatterplots 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.

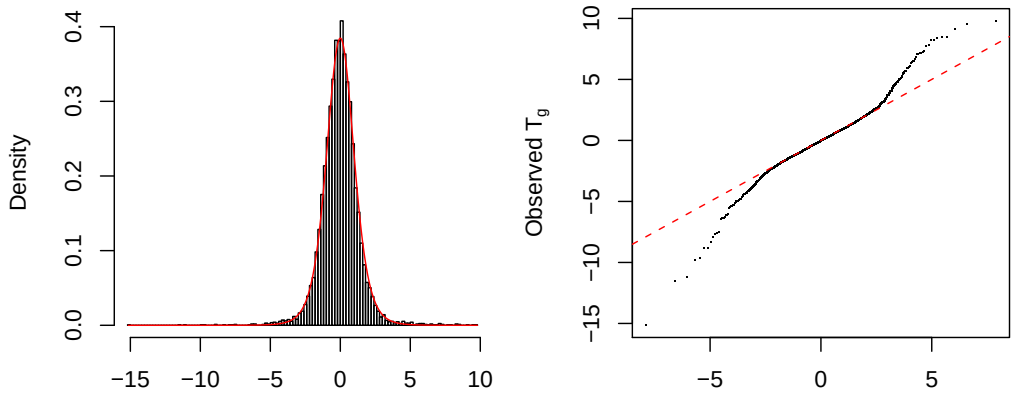


Figure 9: To the left, a histogram of the observed T_g-values together with the density of the null distribution (in red), in the Swirl dataset. To the right, a quantile-quantile plot where the observed values of T_g are plotted against the quantiles of T_g under the null hypothesis. The central part 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

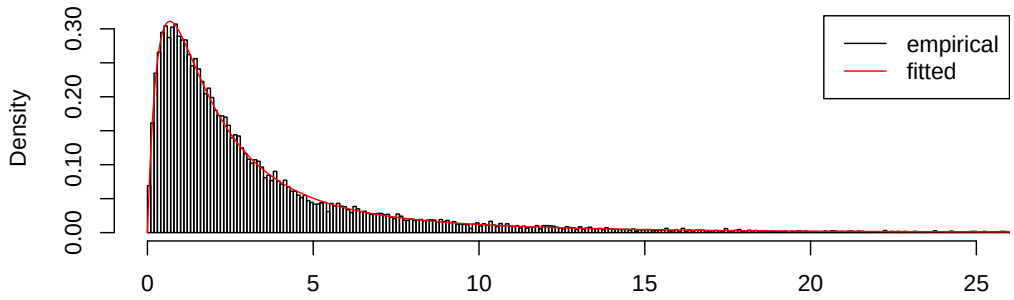


Figure 10: Empirical distribution of S_g in the Polyp dataset, together with the density of S_g given $\alpha = 1.97$.

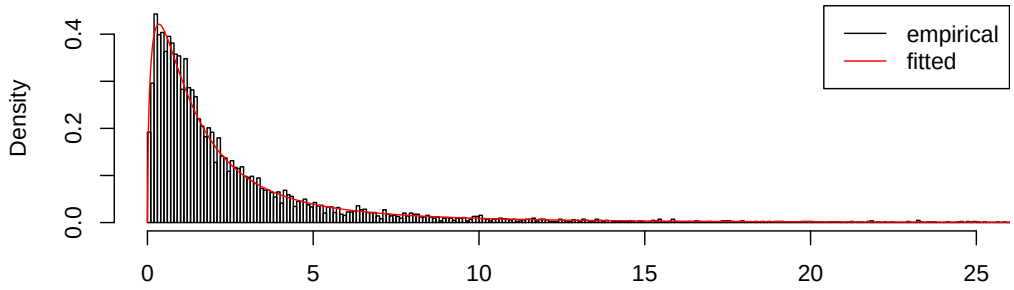


Figure 11: Empirical distribution of S_g in the Swirl dataset, together with the density of S_g given $\alpha = 1.89$.

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 2

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).

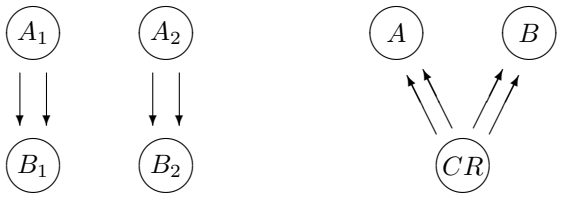
2.2 Examples of parametrisation

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.

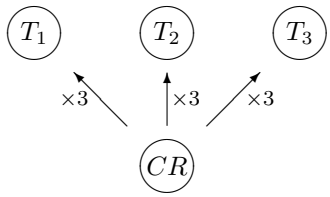


Figure 2: The experimental design used for the simulation studies.

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.

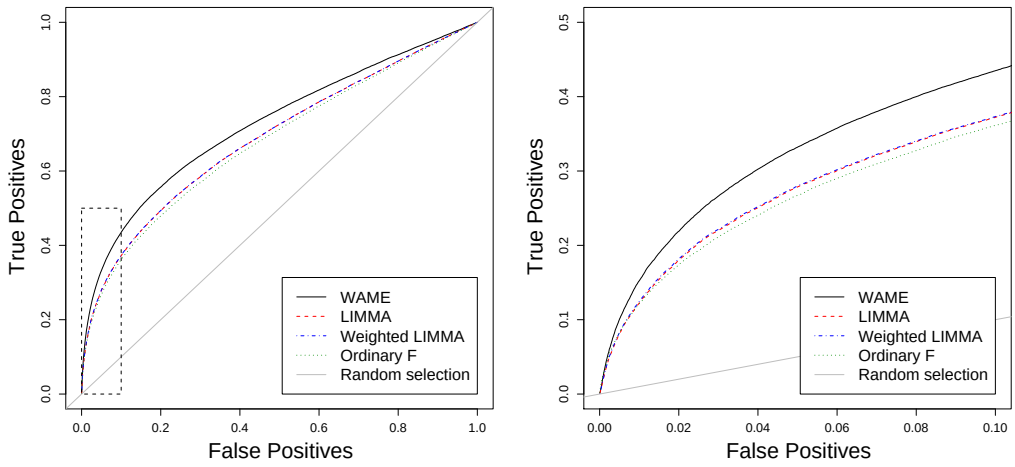


Figure 3: Receiver Operating Characteristic curves for testing regulation at time point two. Four methods are compared on simulated data; WAME, LIMMA with and without weights and the ordinary F-statistic. The figure to the right is a magnification of the dashed box to the left.

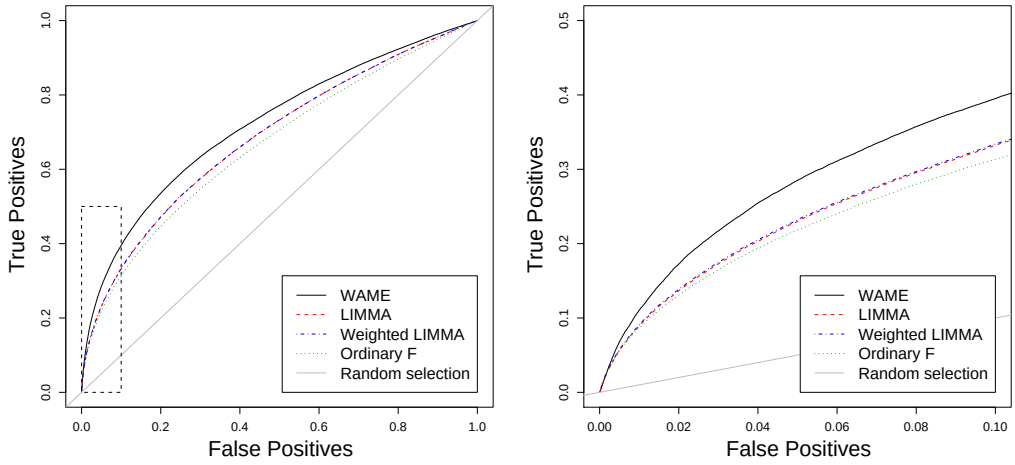


Figure 4: Receiver Operating Characteristic curves for testing regulation at all time points. Four methods are compared on simulated data; WAME, LIMMA with and without weights and the ordinary F-statistic. The figure to the right is a magnification of the dashed box to the left.

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.

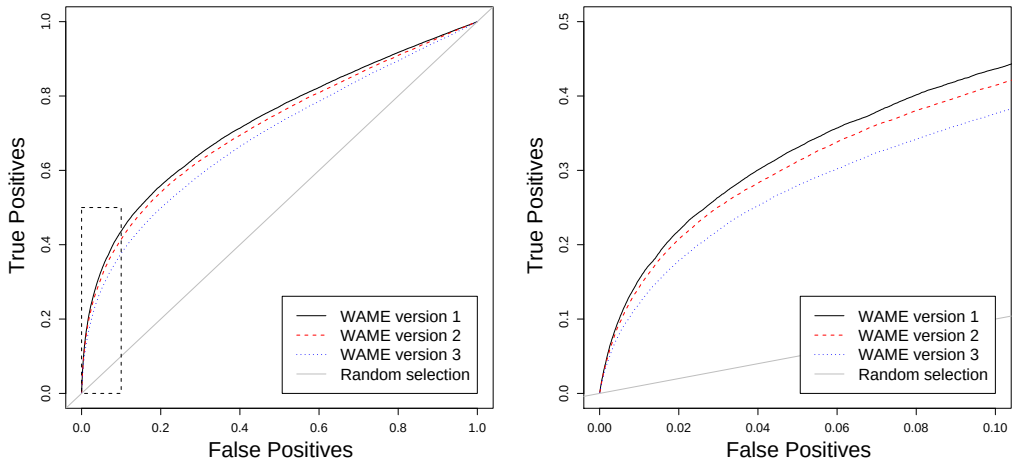


Figure 5: Simulated Receiver Operating Characteristic curves for testing three versions of WAME, all using given covariance matrices. Version 1 uses the correct covariance matrix, version 2 uses the correct covariance matrix but with all correlations between the time points set to zero and finally, version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

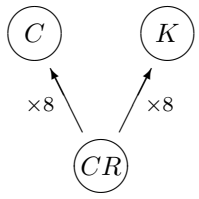


Figure 6: Experimental design for the apoAI dataset. The conditions C and K correspond to the control mice and knockout mice respectively. CR is the common reference which was created by pooling the mRNA from the controls. The mRNA from the control mice and the knockout mice were labeled with Cy5 and 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

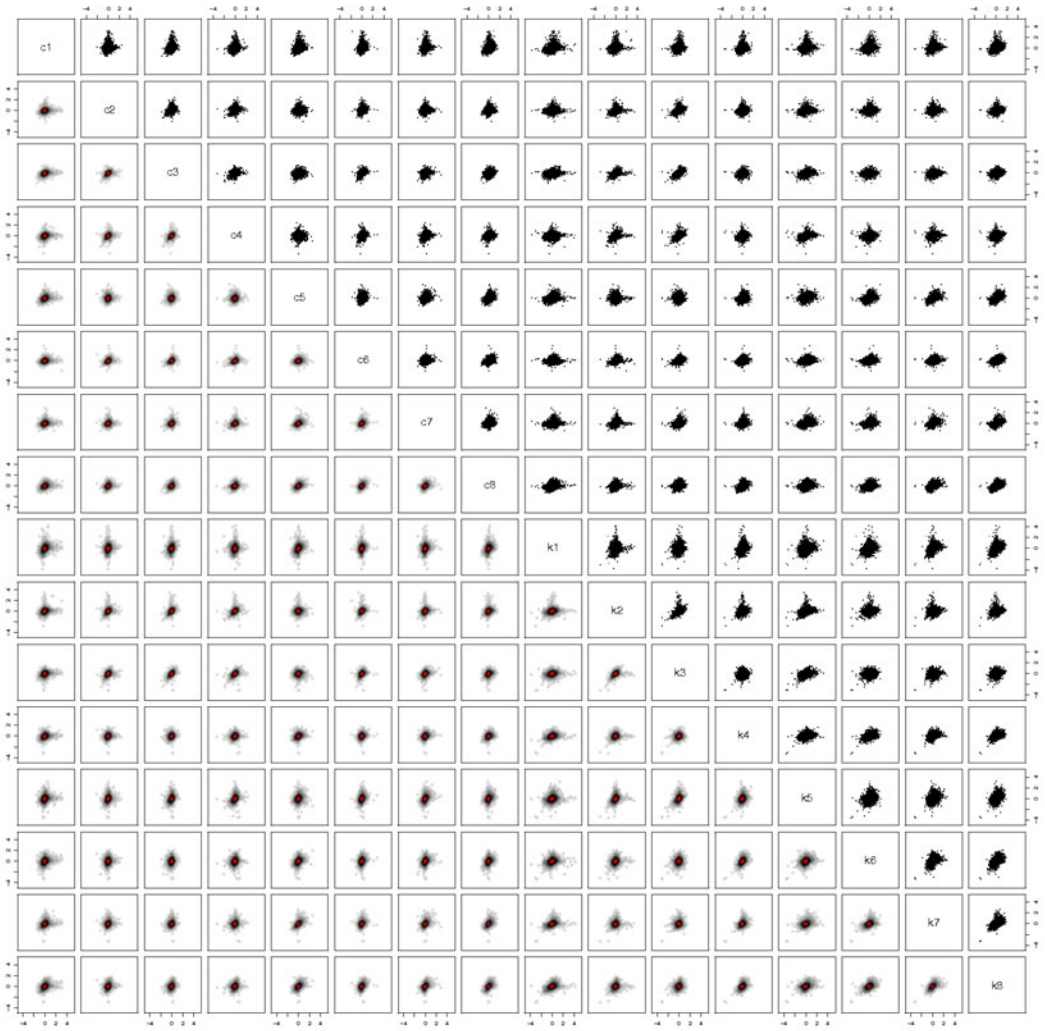


Figure 7: Pair-wise plots of the $\log_2$-ratios for all the 16 arrays in the apoAI dataset. The lower left half shows kernel density estimates of the two-dimensional distribution according to the colour-scale: white, grey, black and red (in increasing level of density). A standalone image in the lossless 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

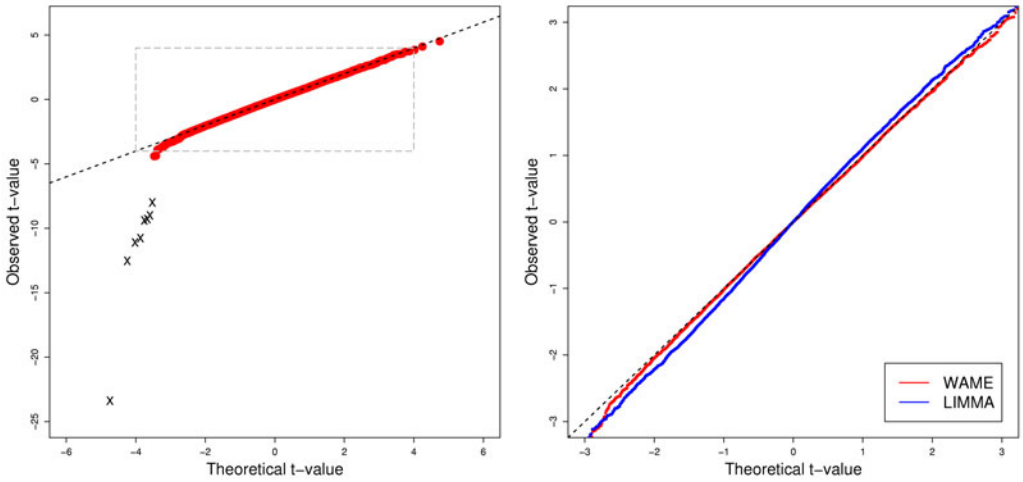


Figure 8: To the left a quantile-quantile plot comparing the estimated weighted moderated t-statistic (WAME) to its corresponding theoretical null distribution. The eight elements marked with a cross corresponds to the four genes that were verified to be differentially expressed. To the right is a blow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

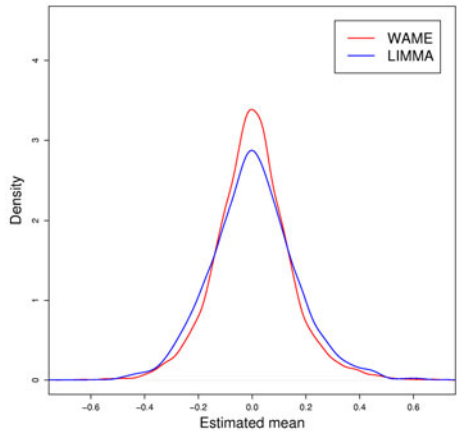


Figure 9: Kernel density estimates of the distributions of the mean values from WAME and LIMMA. The spread of the former is smaller which suggests that the estimator in WAME is more precise.

(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.

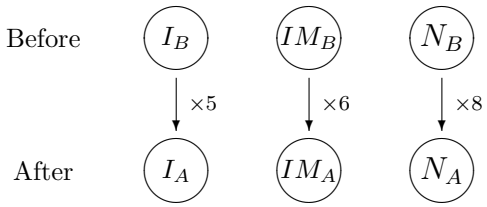


Figure 10: Experimental design for the cardiac dataset. All three groups have a condition before the treatment (I_B, IM_B and N_B) and a condition after the treatment (I_A, IM_A and N_A). The dataset consists of 19 observations in total.

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

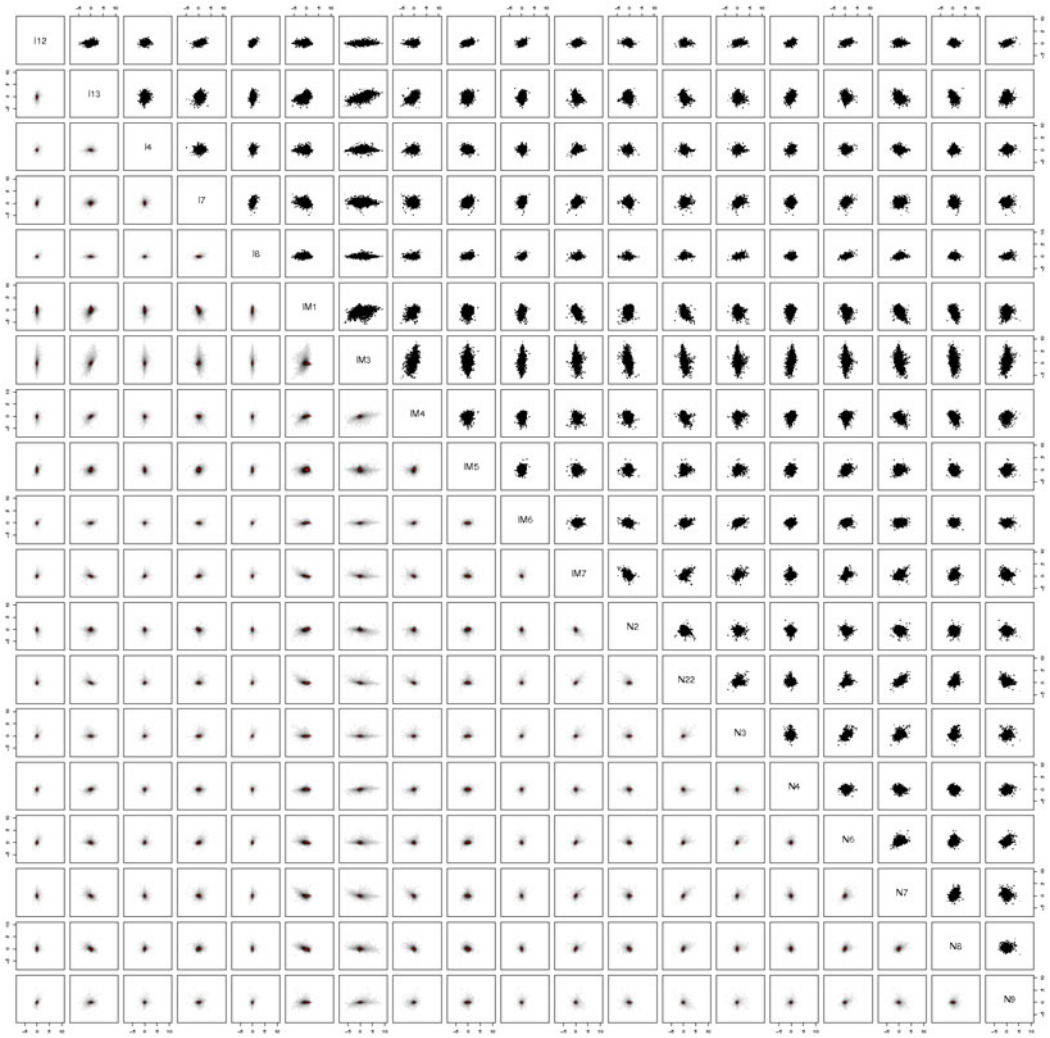


Figure 11: Pair-wise plots of the $\log_2$-ratios for all the 19 patients in the cardiac dataset. The lower left half shows kernel density estimates of the two-dimensional distribution according to the colour-scale: white, grey, black and red (in increasing level of density). A standalone image in the 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.

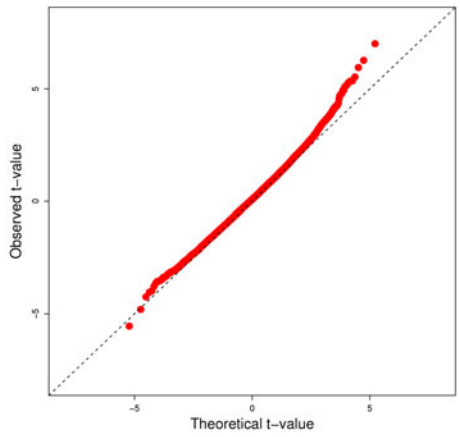


Figure 12: This figure shows the quantile-quantile plots of the weighted moderated t-statistic for the test of differential expression in group I.

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

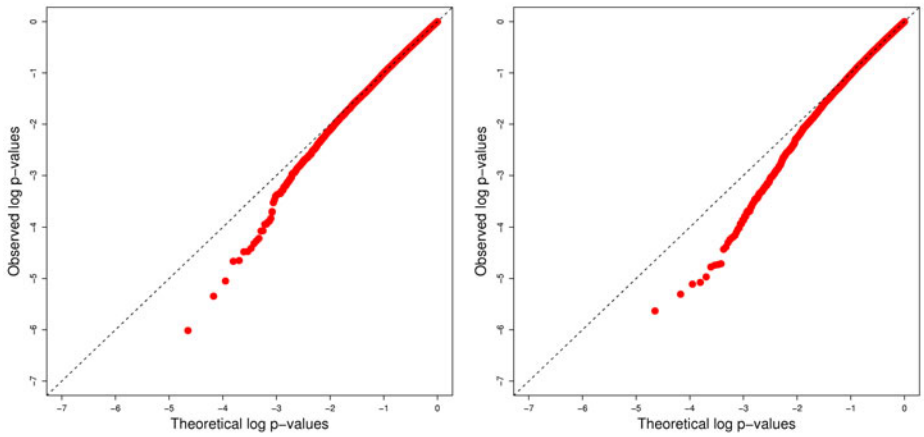


Figure 13: This figure shows two quantile-quantile plots of the $\log_{10}$ p-values for the test of differential expression in group I. The plot to the left is based on the general paired model using all 19 patients and the plot to the right is based on the restricted model from Kristiansson et al. (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 3

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, the p-values generated by WAME are found to be substantially more correct than the alternatives when a relatively small proportion of the genes is regulated. 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. The examined other methods produce invalid p-values in many cases, while WAME is shown to produce essentially valid p-values when a relatively small proportion of genes is regulated. WAME is also shown to have higher power than the examined alternative 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).

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 (data not shown) 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, both when using parametric and when using non-parametric analysis methods, 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 M, 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 S, Rhodes D, Perner S, Dhanasekaran S, Mehra R, Sun X, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie J, Shah R, Pienta K, Rubin M, Chinnaiyan A: **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.

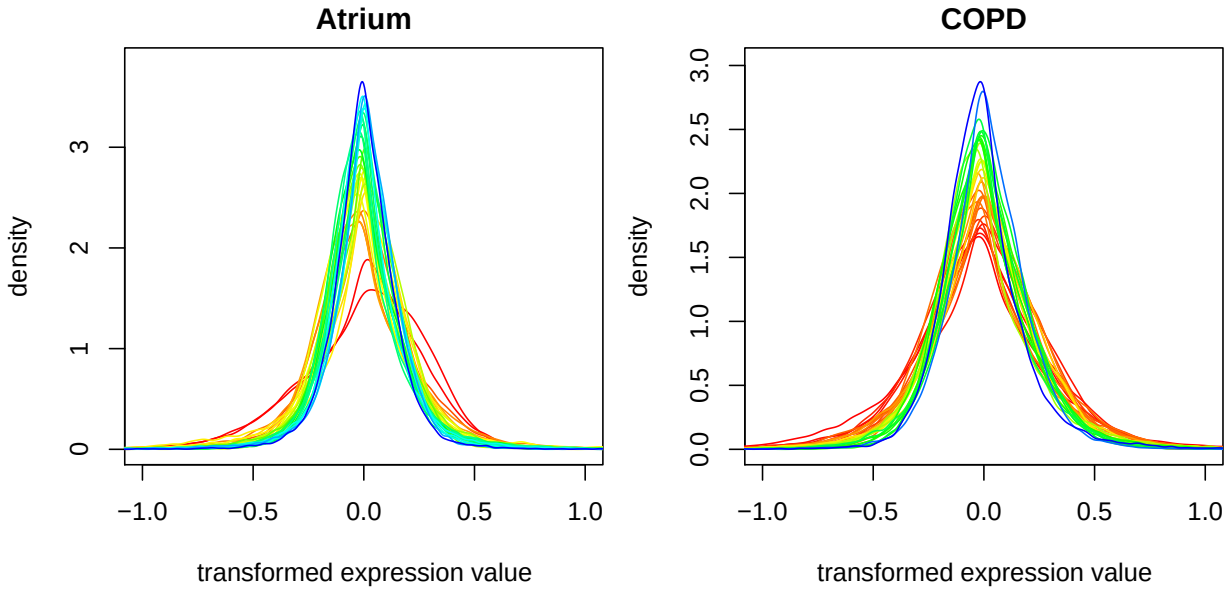


Figure 2 - Pairwise plots

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. Diagonal red clouds in the heat maps reveal correlations between arrays. Off-diagonal numbers show estimated correlations from WAME. Diagonal boxes contain sample names and weights as well as estimated variances from WAME.

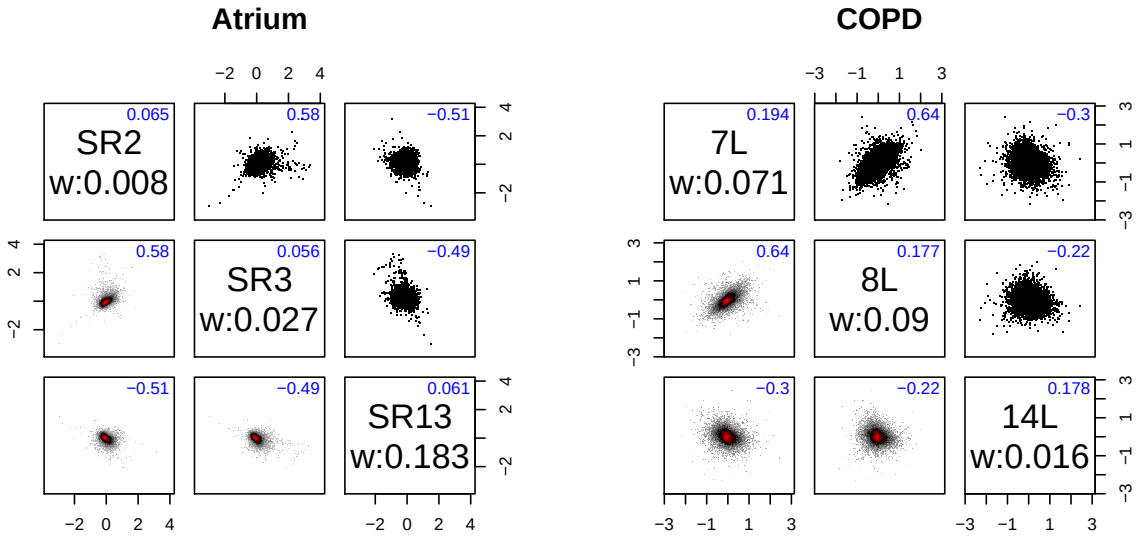


Figure 3 - Observed probability plots

Empirical distribution of p-values compared to the distribution expected for non-differentially expressed genes. The OLM and LIMMA curves largely coincide, as does the identity line and the WAME curve.

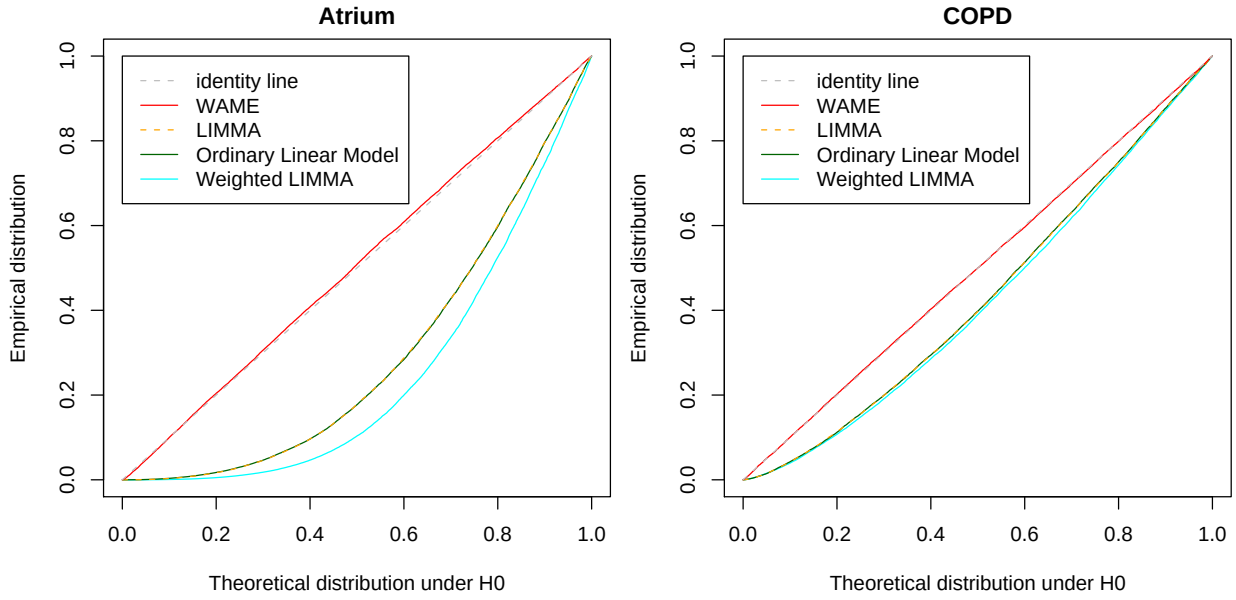


Figure 4 - Probability plots

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.

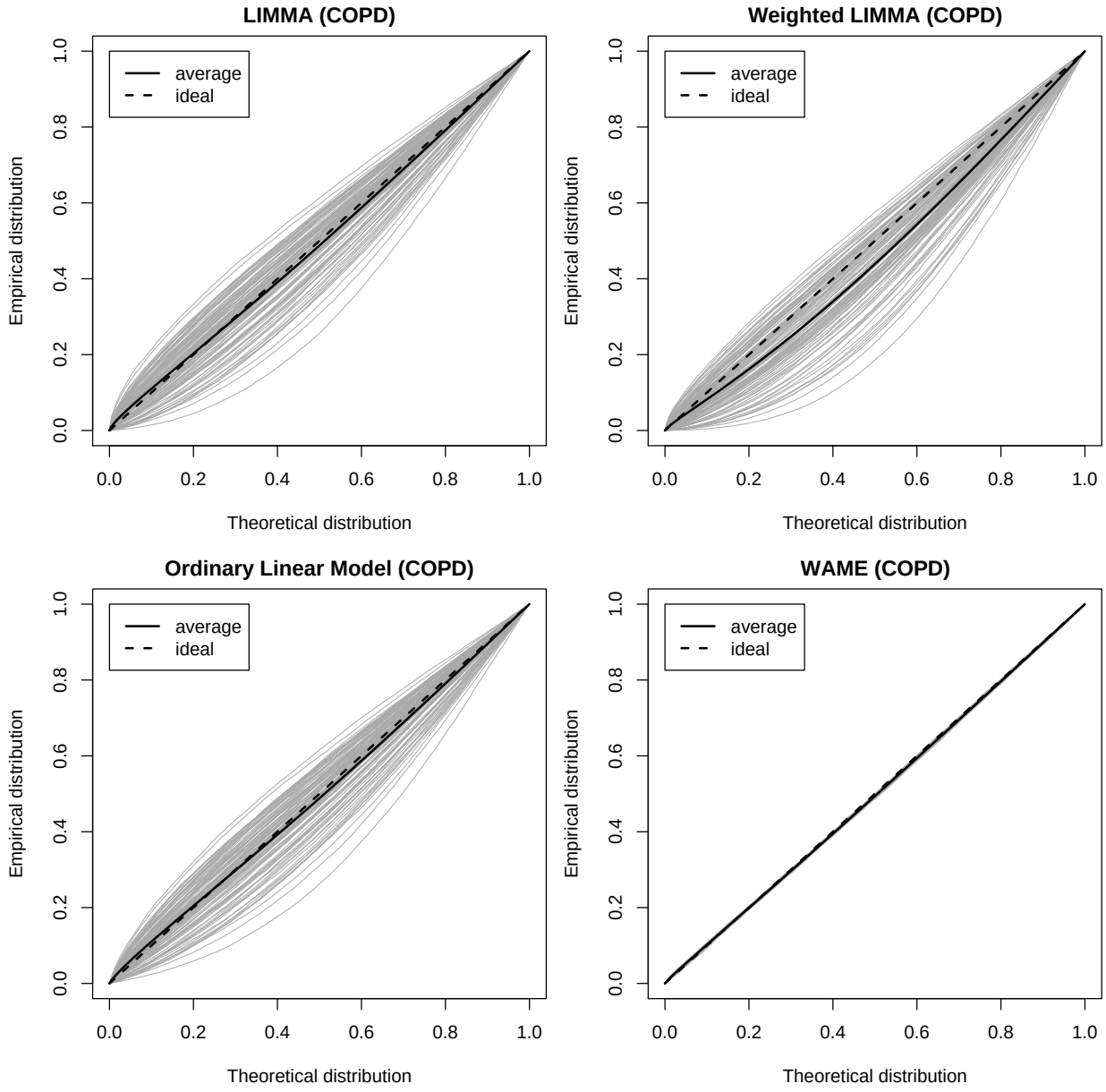


Figure 5 - Average empirical p-value distribution for WAME under regulation

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.

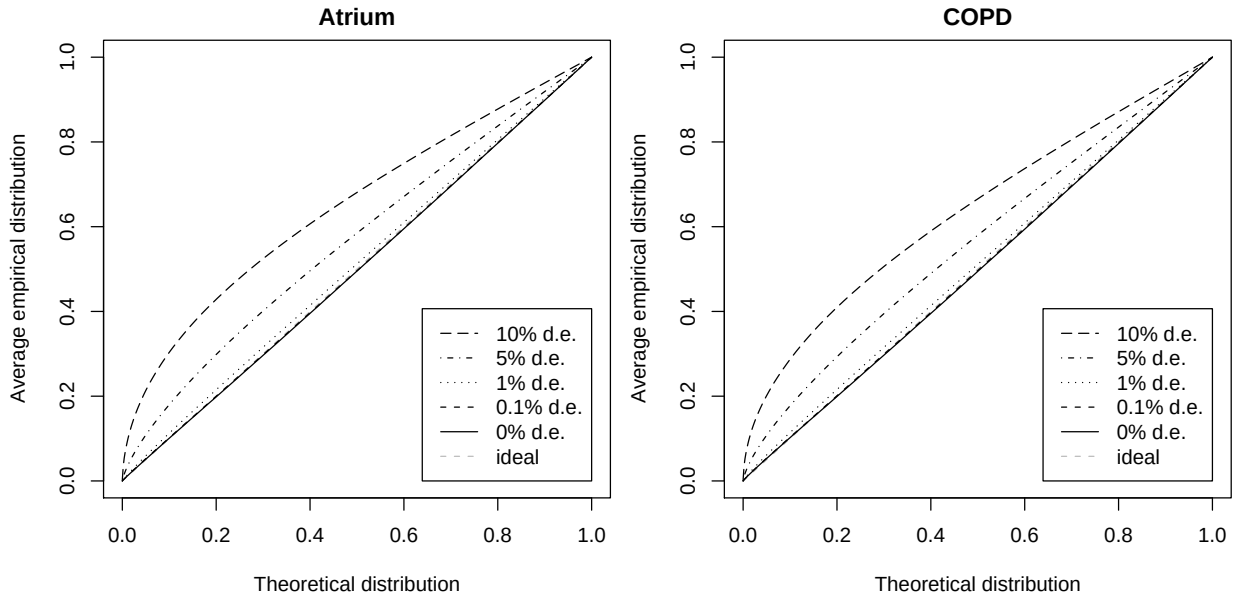
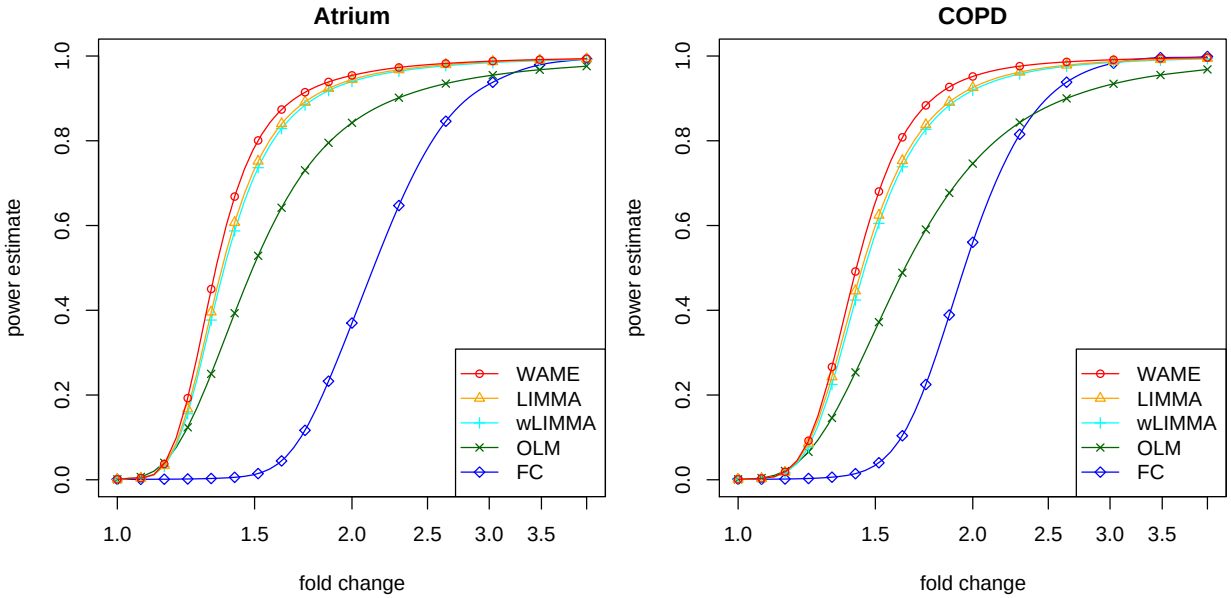


Figure 6 - Estimated power

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 4

Evolutionary forces act on promoter length: assessment of enriched *cis*-regulatory motifs

Erik Kristiansson[1*], Michael Thorsen[2]
Markus J. Tamás[2], Olle Nerman[1]

[1] Department of Mathematical Statistics, Chalmers University of Technology, S-412 96 Göteborg, Sweden.

[2] Department of Mathematical Statistics, Göteborg University, S-405 30, Göteborg, Sweden.

[3] Department of Cell and Molecular Biology/Microbiology, Göteborg University, S-405 30 Göteborg, Sweden.

[*] Corresponding author, erikkr math.chalmers.se

Abstract

Background
Transcription factors regulate transcription by binding to specific DNA sequences (cis-regulatory motifs) in the promoters of their target genes. Several methods for finding enrichments of such motifs within a set of genes, for example genes regulated in a microarray experiment, have been proposed. Although the increasing preciseness of genomic data in general improves these methods, they will on certain occasions perform inaccurately.

Results
We have found that promoter length and gene function are related. In particular, there are differences in promoter length between stress responsive and unresponsive genes. Our analysis suggests that genes with complex transcriptional regulation tend to have longer promoters than genes responding to fewer signals. Furthermore, this phenomenon is shown to be conserved in yeast, fungi and plants, thus evolutionary forces may act on promoter length. These new findings are utilized in a novel method for assessing enrichments of *cis*-regulatory motifs in promoter regions. The procedure extends logistic regression and includes the promoter lengths as a critical element. Evaluations on several datasets show that the proposed method generates more accurate p-values and less false positives compared to other common procedures.

Conclusions
Promoter length is associated with gene function and genes responsive to several stimuli generally have longer promoters. A novel method for assessing enrichment of *cis*-regulatory motifs is introduced and is shown to generate fewer false positives for sets of genes with biases in promoter length. For analysis of *S. cerevisiae* datasets, the method is available as a web service located at http://enricher.math.chalmers.se.

Background

Interactions between proteins and DNA are central in most aspects of genetic activity including gene transcription and DNA packaging, replication and repair. Consequently, it is of great importance to further develop the technologies needed to identify and/or predict these interactions. In transcriptional regulation, the protein-DNA interaction consists of the binding of a transcription factor to a specific *cis*-regulatory motif in the promoter

of its target genes, thereby regulating the recruitment of the transcriptional machinery. The transcriptional patterns of most genes are thus dependent on the presence of *cis*-regulatory motifs in their promoters. Hence, analysis of promoter sequences can aid in predicting the expression pattern of transcriptionally regulated genes. In a reverse approach, it is possible to examine expression patterns and deduce transcription factors involved in the transcriptional regulation of genes. Typically, this is done by identification of genes with similar transcriptional profiles and consecutive analysis of enrichments of the motifs located in the promoters of these genes [1, 2]. To this end, a number of procedures have been suggested, such as the hypergeometric test [3-5] and several tests based on regression models [6-8].

A multitude of sensing mechanisms has evolved to monitor the intra and extra cellular environment. Changes in the physical/chemical milieu will trigger specific sensing and signaling mechanisms allowing the cell to respond appropriately. In many cases, the signaling will initiate a change in transcription of relevant genes resulting in an optimised composition of the cell's proteome. The physiological adaptation to a new environment typically requires turning on or off the expression of several genes. Additionally, modulations in the expression levels of many other genes may be required. For the lower eukaryotic model organism *Saccharomyces cerevisiae*, many investigations have shown that environmental perturbations change the expression levels of up to 15-20% of the entire genome [1, 2, 9-11]. When we investigated the regulatory responses of the genes with altered transcription levels after arsenite exposure [2], we found randomly generated motifs to be enriched compared to the other genes in the genome. Here we argue that this peculiar bias stems from differences in the length of the promoters.

Even in a compact genome such as that of *S. cerevisiae*, it is evident that the length of the intergenic regions varies substantially. This is a potential pitfall that the enrichment studies described hitherto have failed to address. Still, most promoter analyses circumvent the problem by defining the promoter as a fixed number of base pairs (bp) upstream of the transcriptional start site. Typically 1000 bp is used [10, 12, 13] but both 800 bp [14, 15] and 600 bp [16] are also common. To our knowledge, no regulatory motifs situated in the coding region of a gene have been described in *S. cerevisiae* and it is therefore likely that *cis*-regulatory elements predominately occur in the intergenic sequence [17]. The median intergenic distance in *S. cerevisiae* is 455 bp. However, since the variation in promoter length is large [18], a fixed promoter length (e.g. 1000 bp) is a non-optimal solution.

In this paper we show that functional information is encoded in the promoter length. In particular, we identify several groups of genes with longer

or shorter promoters than the rest of the genes in the genome. Furthermore, we show that these differences are conserved in other yeast species and may therefore be a result of evolutionary forces. Consequently, we propose a regression model to assess enrichments of *cis*-regulatory motifs in the promoters in a group of genes. The method takes the promoter length into account and can thus perform satisfactory when there is a difference in promoter length between the group the remaining genes in the genome. *Cis*-regulatory motif enrichment studies using the proposed method can be performed with the web service Enricher located at http://enricher.math.chalmers.se.

Results

Promoter length in *Saccharomyces cerevisiae*

The simplest definition of a gene's promoter region is the DNA sequence located 5' of the corresponding ORF and stretching to the upstream ORF, i.e. the entire upstream intergenic region. Extracting all promoters by this definition revealed that the median length of all promoter regions in the *S. cerevisiae* genome is 455 bp and that the length of promoters varies greatly (Figure 1).

Promoter length differs between subsets of functionally related genes

By analyzing biologically meaningful gene sets, we wanted to explore if functional information can be extracted from promoter lengths. In two recent genomic studies in *S. cerevisiae*, a large set of genes has been identified as differentially expressed in response to a variety of environmental perturbations. This phenomenon was called the environmental stress response (ESR) comprising 800 genes [10] or the common environmental response (CER) comprising 600 genes [9]. The ESR and the CER gene sets are largely overlapping and the genes respond similarly in response to environmental stimuli, including temperature shock, osmotic stress, oxidative stress, low pH, high pH, and nutrient starvation. Our analyses show that the ESR and CER genes have relatively long promoters. The median promoter length of ESR genes is 493 bp which should be compared to the median length of the unresponsive genes of 450 bp ($p = 2 \times 10^{-3}$). Similarly, the median promoter length of CER genes is 552 bp compared to 445 bp for the promoters of the remaining genes ($p = 5 \times 10^{-11}$) (Table 1 and Figure 2). The low p-values imply a statistically significant bias towards longer promoters for genes of

the environmental stress response compared to other genes (see Materials and Methods for a description of the testing procedure).

We went on to look for other categories of genes that differ in promoter length as compared to the rest of the genes in the genome. Interestingly, the promoters of several classes of genes were found to have significantly longer or shorter promoters (Table 1). According to the Saccharomyces Genome Database (SGD) [19], 1033 genes are essential for growth in rich medium. Our analysis revealed that these genes have very short promoters. The median promoter length for the essential genes is 385 bp whereas the median promoter length for non-essential genes is 468 bp ($p = 8 \times 10^{-9}$). Investigating whether functional classes of genes differ in their promoter lengths, e.g. genes annotated similarly in the Yeast GO Slim terms [20], we found that most functional classes do not exhibit biases. However, certain classes such as cell wall (median length 899 bp, $p = 2 \times 10^{-11}$), transporter activity (median length 583 bp, $p = 4 \times 10^{-9}$) and RNA metabolic process (median length 340 bp, $p = 6 \times 10^{-9}$) do differ significantly in their promoter lengths. Lists of all GO Slim terms and their corresponding promoter lengths are available as supplementary information (Supplementary Table 1).

Relatively long promoters of stress regulated genes may be a conserved phenomenon

To investigate if the relationship between promoter length and gene function is an evolutionary conserved feature, we extended our analysis to include other species. The filamentous fungus *Ashbya gossypii* has the smallest genome of any free-living eukaryote reported so far. The genome is extremely compact with a median distance between open reading frames of only 342 base pairs. *A. gossypii* and *S. cerevisiae* diverged more than 100 million years ago, and their genomes differ substantially in GC content. Still, 95% of the protein-coding sequences of *A. gossypii* have homologues in the *S. cerevisiae* genome with the vast majority at syntenic locations [21]. Therefore we could easily find 405 of the homologous genes from the common environmental stress response in *S. cerevisiae* in the *A. gossypii* genome. We find that these putative CER genes in *A. gossypii* have a median promoter length of 407 bp, which is significantly longer than the median promoter length of remaining genes of 337 bp ($p = 4 \times 10^{-5}$) (Table 1). When we look at *A. gossypii* homologues of the genes essential in *S. cerevisiae* we find the promoters to be much shorter than the promoters of non-essential genes ($p = 4 \times 10^{-9}$). These differences may be a consequence of evolutionary constraints but can also be explained as an indirect effect of the high degree of synteny between

the species.

The evolutionary divergence between the fission yeast *Schizosaccharomyces pombe* and the budding yeast *S. cerevisiae* is estimated to 1,14 billion years [22], approximately equal to the distance between *S. cerevisiae* and man. Therefore, an analysis of the promoters in the *S. pombe* genome could be a better indication for any potential evolutionary constraints on promoter length. We extracted the promoters of the *S. pombe* genome similarly to our *S. cerevisiae* approach. The median length of the promoter regions in the *S. pombe* genome is 829 bp and the length of promoters varies greatly (Supplementary Figure 1). Similar to *S. cerevisiae*, the environmental stress response of *S. pombe* has been reported. Analyses of the *S. pombe* transcriptome revealed that many genes display analogous transcriptional profiles in response to oxidative stress, cadmium stress, temperature shock, osmotic stress, and to a DNA damaging agent. Here, the authors named it the core environmental stress response (CESR) [23]. A conservative definition of the CESR comprises 237 genes whereas a more loose definition resulted in 704 genes. The conservative definition of the CESR is genes that are induced more than two fold in at least four of the five conditions tested. The loose definition is the genes that were consistently induced in response to all stresses, but failed to make the twofold cutoff in some of the responses. With the loose definition of CESR, the genes have a median promoter length of 863 bp, compared to the median of the remaining promoters of 826 bp ($p = 2 \times 10^{-2}$). However, analyzing the genes of the conservatively defined CESR gives a more striking result. Here, the median length of the regulated genes is 1243 bp while the median length of the unregulated genes is 816 bp ($p = 7 \times 10^{-15}$) (Supplementary Figure 2).

Next, we analyzed promoter lengths in the genome of the plant *Arabidopsis thaliana*. Identification of genes transcriptionally up-regulated in response to environmental stress treatments (heat, cold, drought, salt, high osmolarity, UV-B light and wounding) has been reported [24]. When performing the same promoter analysis as above, we found that the median promoter length for *A. thaliana* environmental stress responsive genes were 1672 bp compared to the 1113 bp of the promoters of the remaining genes (p¡10-16) (Table 1). Taken together our findings suggest that evolutionary forces act on promoter length.

Correlation between promoter length and *cis*-regulatory motifs

The examples presented above warrant for caution when trying to assess enrichment of *cis*-regulatory motifs in promoter sequences. Naturally, the

likelihood of a random occurrence of any motif increases with the promoter length as, illustrated in Figure 3 (dotted line). In this figure, the number of different *cis*-regulatory motifs and the promoter length are shown for all genes in the genome of *S. cerevisiae*. A higher number of different motifs in longer promoters is also observed when the noise level is reduced by phylogenetic filtering (Figure 3, solid line) and may thus be a biologically relevant phenomenon.

In a reversed approach, we analysed the transcriptional profiles of the genes in the CER and divided all genes into classes depending on the number of stress conditions to which they respond. Plotting the median promoter length of each group against the number of stress conditions, a highly positive trend (p-value ¡ 10-5) and a good correlation (r2=0.96, $p = 1 \times 10^{-3}$) is observed (Figure 4). The same was done for the Arabidopsis promoters and again we found a positive trend between the promoter length and the number of different conditions which induces transcription of the gene (Supplementary Figure 3). This indicates that the length of promoters correlates with the number of signals that is integrated onto the promoters.

Assessing enrichment of *cis*-regulatory motifs under promoter length bias

Procedures for finding enrichments of *cis*-regulatory motifs are today well established tools in many aspects of modern biology. These methods are applied to a set of genes, typically from a microarray or chromatin immunoprecipitation assay, but other assortments are also possible.

One of the most frequently used methods to find overrepresented motifs is the hypergeometric test (also known as Fishers exact test) [3, 4, 25]. For a given set of genes, the hypergeometric test is performed by observing the number of genes in the set that has the motif present. Under the assumption of independence, this number has a known distribution and hence significance can be derived. Various kinds of regression models have also been suggested for identification of enriched motifs [6, 26]. In particular, logistic regression, where the presence of the motif and the selection of a gene is considered to be binary variables (i.e. present or not, selected or not), is commonly used [7, 8]. In contrast to the hypergeometric test, significance in the regression models is derived by large sample approximations and they are thus not suitable for sets with very few genes.

Their popularity notwithstanding, both the hypergeometric test and the logistic regression model assume, directly or indirectly, that there is no difference in promoter length between the sets of genes. We therefore propose a

novel procedure to assess enrichments of *cis*-regulatory motifs in the promoters of a set of genes. The method includes the promoter lengths as covariates and can thus handle length biases in a proper manner. The model is based on the generalized additive model (GAM) framework [27, 28] and can be seen as an extension of logistic regression. Further details for all three methods discussed here can be found in Materials and Methods.

Promoter length bias can result in false positives

A simulation study was performed to compare the proposed model to the hypergeometric test and logistic regression. Datasets, divided into two groups of artificial promoters, each consisting of 500 and 5000 promoters respectively, were generated as follows. The length of the promoters were sampled from uniform distributions such that the promoters in the first group were, on average, 25% longer than the promoters in the second group (median promoter length were 500 bp and 400 bp respectively). Given the length, each promoter was generated by sampling the four nucleotides with equal probabilities. In total, 1000 datasets were created. Further details are available in Materials and Methods. The simulated promoters were searched for a six nucleotide long motif and tests for enrichments within the smaller group were performed for all three methods. Histograms of the resulting p-values are shown in Figure 5. Since the promoter sequences are completely random, no enrichments should exist and the p-values should therefore be uniformly distributed between 0 and 1. The results from both the hypergeometric test and logistic regression are clearly skewed towards lower p-values. However, this is not the case for the proposed model, which suggests that it results in less false positives when there is a bias in promoter length.

Transcription factors in arsenic stress

To evaluate our model on real biological data, we assessed enrichments of *cis*-regulatory elements in a microarray dataset on transcriptional changes in response to arsenite exposure [2]. Based on this experiment, putatively up- and down-regulated genes with an absolute log2 fold-change greater than 2 where chosen for enrichment analysis. The promoter lengths of these genes were significantly longer than the promoters in the rest of the genome (Table 1). Accordingly, the model proposed in this paper is suitable for a proper analysis of this data.

Enrichment analysis of all experimentally verified *cis*-regulatory motifs described in SGD [19] were performed. Phylogenetic filtering was used to re-

move spurious motifs (see Materials and Methods). Examples of significant motifs according to the proposed model can be seen in Table 2 together with p-values from the hypergeometric test and logistic regression. In general, the p-values for highly significant motifs seem to be lower for the regression based methods for than the hypergeometric test. Furthermore, the proposed method is, as expected, more conservative than logistic regression. The last two rows in Table 2 show examples of motifs that, according to the proposed models, are candidate false positives. These motifs have low p-values for both the hypergeometric test and the logistic regression model but get substantially higher p-values when the promoter length is included. Complete lists with p-values for all *cis*-regulatory motifs in yeast can be found in Supplementary Table 2 and 3.

Discussion

Using the *S. cerevisiae* genome as a model, we analysed the length of promoter regions of genes in different datasets and we found a striking relationship between long promoters and responsiveness to a variety of stresses. Complementary to this, we analyzed analogous datasets of promoters in *S. pombe* and *A. thaliana* found the same relationship. Thus, longer promoters of stress inducible genes seem to be a conserved phenomenon. The simplest explanations may be that relatively long promoters allow for integration of a large number of regulatory inputs onto the promoter region of these genes and, conversely, that other promoters may have been shortened to reduce genome size.

The integration of many signals into a promoter requires a number of motifs. If several signalling pathways target the same promoter, binding sites must exist for all relevant transcription factors. Additionally, the regulation can be combinatorial and thus a number of auxiliary factors may also bind. In these cases, the promoters must be long enough or steric hindrances may prevent the binding of all the necessary factors. We speculate that the level of complexity of the regulation of a gene is related to the function of the gene product. Hence, promoters of certain classes of genes may have evolved to be longer due to complex regulation, whereas other classes of genes may have evolved shorter promoters to reduce genome size. A smaller genome is believed to be an advantage as replication is faster, allowing for faster proliferation.

Evolution may not have acted to reduce the intergenic stretches of all genomes. Therefore, a relationship between promoter length and regula-

tory complexity might not always exist. We have argued here that utilising the information about the length of the individual promoters improves the validity of the p-values generated when estimating enrichments of *cis*-regulatory motifs in *S. cerevisiae* and *S. pombe* datasets. Similar analyses of mammalian promoters may be difficult, as they in general are much longer and poorly defined. However, the annotation of mammalian genomes is improving and several resources for extracting promoter sequences from mammalian genomes exist. One interesting study on the organisation of the human genome clearly demonstrates that both the introns and the intergenic space around housekeeping genes are shorter than corresponding sequences for tissue specific genes. Similar tendencies are also reported for the fly Drosophila melanogaster and the worm Caenorhabditis elegans. The author speculated that the tissue specific genes have a more composite transcriptional regulation, and thus require longer sequences to facilitate a more dynamic chromatin structure [29]. These findings corroborate our conclusions and indicate that promoter length is an important factor when analysing genomes of higher eukaryotes.

Assessing enrichments of de novo motifs is a process where the rate of false positives has been particularly bothersome. A recent report [30] identified a number of problems which needs to be addressed to improve these methods. Among them were: (I) the lack of rigorous models and of an exact p-value measuring motif enrichment; (II) the tendency, in many of the existing methods, to report presumably significant motifs even when applied to randomly generated data. We believe that our model is a novel approach to (I) and a clear improvement of (II).

Our model may also be of use in other types of enrichment studies. In post-transcriptional regulation, RNA binding proteins (RBPs) interact with sites in the untranslated regions (UTRs) of the mRNAs and regulate localization, degradation, and translational control. In *S. cerevisiae*, the length of the UTRs has been shown to depend on gene function [31] and the number of identified RBP binding motifs is increasing [32]. Hence, biases in UTR lengths between groups of genes cannot be ruled out, and the proposed model should be straight-forward to apply in these situations.

Eukaryotic genomes are packaged into nucleosome particles that prevent the DNA from interacting with other DNA binding proteins. Experimental evidence shows that DNA packaging can control accessibility of specific sequences by blocking access to irrelevant non-functional sites [33, 34]. A future direction would therefore be to incorporate chromatin structure prediction into the enrichment analysis. A computational method modelling nucleosome positioning has recently been developed, that can predict roughly 50% of the

genome-wide nucleosome organisation in yeast [35]. Additional information can be extracted by predict acetylation and methylation of nucleosomes to identify sequences with regulatory activity [36]. These are important areas of research and further development of these tools is needed to improve their accuracy. Combining robust prediction algorithms for nucleosome organisation with phylogenetic based comparative genomics holds great promises for vastly enhanced prediction of true *cis*-regulatory elements.

Conclusion

We have presented evidence that promoter lengths include functional information and several categories of genes have been shown to have promoters longer or shorter than promoters of the other genes in the genome. This was seen for stress induced genes in *S. cerevisiae*, *A. gossypii*, *S. pombe* and *A. thaliana* thus evolutionary forces may act on promoter length. Consequently, we have developed a method for assessing enrichments of *cis*-regulatory motifs in sets of promoters with varying lengths. To our knowledge, this is the first model which utilizes information encoded in the promoter lengths to find overrepresentations of transcription factor binding sites. Simulations show that the proposed model performs adequately, in contrast to the hypergeometric test and logistic regression, which both can lead to false positives.

Materials and Methods

Sequence data and analysis

The intergenic regions for both *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* were downloaded from the GeneDB database [37] (March, 2007). For each gene, the 5' upstream sequence from the start of the open reading frame (ORF) until the end of the previous ORF (on any strand) was extracted. In this study, these sequences are referred to as promoters. Genes with ORFs overlapping other genes such that the intergenic region is nonexistent were removed from the analysis. This resulted in 5735 promoters in *S. cerevisiae* and 5484 promoters in *S. pombe*. Sequence data for *Saccharomyces mikatae*, *Saccharomyces kudriavzevii*, *Saccharomyces bayanus*, *Saccharomyces castellii*, and *Saccharomyces kluyveri*, all yeast species closely related to *S. cerevisiae* [38], were downloaded from SGD [19] (April, 2007). Intergenic regions for every gene with a homologue in *S. cerevisiae* were extracted and cut at the same position as the corresponding promoter in

S. cerevisiae, resulting in 2983, 3606, 4733, 4090 and 2798 promoters for *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii*, and *S. kluyveri*, respectively. Intergenic regions for *A. gossypii* were extracted from the Ashyba Genome Database (AGD) [39] and *S. cerevisiae* homologues were mapped using the table published in Dietrich et al. 2004 [21]. Intergenic regions for *A. thaliana* were retrieved from the TAIR database [40]. The total number of promoters for *A. gossypii* and *A. thaliana* were 4683 and 27736 respectively.

All tests of differences in promoter lengths between groups of genes performed in this study were done using the Wilcoxon rank sum test. This is a non-parametric procedure designed to test differences between medians and is therefore robust against exaggerated promoter lengths.

Transcription factor binding site data and phylogenetic filtering

129 known transcription factor binding sites (motifs), corresponding to 92 distinct transcription factors in *S. cerevisiae*, were downloaded from SGD (March, 2007). The motifs were described by IUPAC-codes. The length of the motifs varied between 5 bp and 20 bp (9 bp in average). A motif was defined as present in a promoter in *S. cerevisiae* given at least one exact match. Under phylogenetic filtering, a motif was defined as present if it was (1) present in the promoter of *S. cerevisiae* and (2) present in at least half of the available corresponding promoters in the closely related species. Note that the position of a motif can vary in the promoters between the different species.

Mathematical details and model descriptions

Assume that there are N genes in the genome and that we are interested in testing overrepresentation of a motif within a subset A consisting of n genes. For $g = 1, \ldots, N$, let x_g be a binary 0,1 valued variable indicating whether the motif is present in the promoter of gene g. Furthermore let y_g be another binary variable indicating if gene g belongs to the subset A. Let $z_g = x_g y_g$, i.e. z_g is one if gene g is in the subset A and has the motif present in the promoter. Finally, let l_g be the length of the promoter for gene g.

The hypergeometric test can be seen as drawing a fixed number of balls (genes in A) from an urn consisting of red (genes with a motif) and blue (genes without a motif) balls. The observed number of red balls among the drawn balls (genes in A with the motif) can then be compared to all other possible draws and the significance calculated. If we let $x_{tot} = \sum x_g$ and

$z_{tot} = \sum z_g$, then

$$p_{hyper} = \sum_{i=z_{tot}}^{\min(x_g,n)} \frac{\left(\begin{array}{c} x_{tot} \\ i \end{array} \right) \left(\begin{array}{c} N - x_{tot} \\ n - i \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)}.$$

The hypergeometric test relies on the assumption that all balls are drawn with equal probability, i.e., that motifs occur with the same probability in all promoters, which is questionable if there is a bias in promoter length.

The standard logistic regression model can be formulated as a linear relationship between the log-odds of having the motif present (y_g) and the selection of the gene (x_g), i.e.

$$\log \frac{\mathbb{P}(y_g = 1)}{\mathbb{P}(y_g = 0)} = \alpha + \beta x_g.$$

The coefficients α and β are estimated from data and an enrichment can be inferred by testing whether β is positive. Note that under the null hypothesis, i.e., when β is zero, the probability of finding the motif is equal for all genes regardless of the length of the promoter.

The method proposed in this paper is an extension of the logistic regression model to a generalized additive model (GAM) [27, 28, 41]. In GAMs, non-linear relations of covariates are modelled non-parametric by smoothing functions. In our case, an unknown function of the promoter length is added to compensate for any bias in the set of genes of interest. In other words, we assume that the log-odds of the presence of the motif depends both on the selection of the gene and the length of the promoter. Using the same notation as above, the extended model can be formulated as

$$\log \frac{\mathbb{P}(y_g = 1)}{\mathbb{P}(y_g = 0)} = \alpha + \beta x_g + f(l_g),$$

where f is an unknown smooth function. The coefficients α and β and the function f are estimated from data. Extremely long promoters, typically coming for regions scarce of ORFs such as the telomeres, were truncated at the 95th percentile resulting in a maximal length of 1902 bp. A p-value for overrepresentation can then, as for the logistic regression model, be calculated by testing whether β is positive.

All calculations were performed using the statistical language R [42]. The mgcv package [28] was used to fit the GAM. For the logistic regression model,

the maximum likelihood estimates of the coefficients α and β were found using iterated re-weighted least squared (IRLS). A penalized version of IRLS (P-IRLS) was used, together with the UBRE smoothing condition, to estimate the function f and the coefficients α and β for the generalized additive model. For both models, overrepresentation was inferred by testing $\mathrm{H}_0 : \beta = 0$ versus $\mathrm{H}_A : \beta > 0$. Large sample normal approximations of the estimated values of β were used to calculate p-values. Enrichments of motifs with less than 5 hits within the gene set A were not performed by the regression models. Further details regarding the GAMs are available in [28].

Details regarding the simulation study

The data for the simulation study was generated as follows. Two sets of artificial promoters were created, one with 500 and the other with 5000 sequences. The length of the promoters in the first group was chosen uniformly between 1 and 1000 and the corresponding distribution for the second group was uniform between 1 and 900. The average length difference thus becomes 100 bp, which is typical for the different microarray datasets examined in this study. Given the length, the nucleotide sequence of the promoters was generated by repeatedly sampling from A, C, G, and T with equal probability. In total, 1000 such sets of promoters were generated (i.e. 1000 sets were each set has 5500 promoters in total) and for each set enrichment of a six-nucleotide motif was inferred.

Authors' contributions

The analyses were designed by EK and MT and performed by EK in collaboration with ON. The proposed model was formulated by EK and ON. The manuscript was drafted by EK and MT and read, edited and approved by all authors.

Acknowledgement

This research was financed by the Swedish National Research School in Genomics and Bioinformatics. We also would like to thank Alexandra Jauhiainen for extensive comments and suggestions on the manuscript and Jonas Waringer and Olga Kourtchenko for discussions.

References

[1] Boer VM, de Winde JH, Pronk JT, Piper MD: The genome-wide transcriptional responses of Saccharomyces cerevisiae grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus, or sulfur. *J Biol Chem* 2003, 278(5):3265-3274.

[2] Thorsen M, Lagniel G, Kristiansson E, Junot C, Nerman O, Labarre J, Tamas MJ: Quantitative transcriptome, proteome and sulfur metabolite profiling of the Saccharomyces cerevisiae response to arsenite. *Physiol Genomics* 2007.

[3] Hughes JD, Estep PW, Tavazoie S, Church GM: Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* 2000, 296(5):1205-1214.

[4] Sharan R, Ovcharenko I, Ben-Hur A, Karp RM: CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics* 2003, 19 Suppl 1:i283-291.

[5] Ettwiller L, Paten B, Souren M, Loosli F, Wittbrodt J, Birney E: The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biol* 2005, 6(12):R104.

[6] Bussemaker HJ, Li H, Siggia ED: Regulatory element detection using correlation with expression. *Nat Genet* 2001, 27(2):167-171.

[7] Copley RR: The EH1 motif in metazoan transcription factors. *BMC Genomics* 2005, 6:169.

[8] Keles S, van der Laan MJ, Vulpe C: Regulatory motif finding by logic regression. *Bioinformatics* 2004, 20(16):2799-2811.

[9] Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA: Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell* 2001, 12(2):323-337.

[10] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 2000, 11(12):4241-4257.

[11] Lai LC, Kosorukoff AL, Burke PV, Kwast KE: Dynamical remodeling of the transcriptome during short-term anaerobiosis in Saccharomyces cerevisiae: differential response and role of Msn2 and/or Msn4 and other factors in galactose and glucose media. *Mol Cell Biol* 2005, 25(10):4075-4091.

[12] Haverty PM, Hansen U, Weng Z: Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Res* 2004, 32(1):179-188.

[13] Liu Y, Ringner M: Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis. *Genome Biol* 2007, 8(5):R77.

[14] Daran-Lapujade P, Jansen ML, Daran JM, van Gulik W, de Winde JH, Pronk JT: Role of transcriptional regulation in controlling fluxes in central carbon metabolism of Saccharomyces cerevisiae. A chemostat culture study. *J Biol Chem* 2004, 279(10):9125-9138.

[15] Garcia R, Bermejo C, Grau C, Perez R, Rodriguez-Pena JM, Francois J, Nombela C, Arroyo J: The global transcriptional response to transient cell wall damage in Saccharomyces cerevisiae and its regulation by the cell integrity signaling pathway. *J Biol Chem* 2004, 279(15):15183-15195.

[16] Patil CK, Li H, Walter P: Gcn4p and novel upstream activating sequences regulate targets of the unfolded protein response. *PLoS Biol* 2004, 2(8):E246.

[17] Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB: Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol* 2004, 2(12):e398.

[18] Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S et al: The genome sequence of Schizosaccharomyces pombe. *Nature* 2002, 415(6874):871-880.

[19] Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK et al: Genetic and physical maps of Saccharomyces cerevisiae. *Nature* 1997, 387(6632 Suppl):67-73.

[20] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al: Gene ontology: tool

for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, 25(1):25-29.

[21] Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S et al: The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science* 2004, 304(5668):304-307.

[22] Hedges SB: The origin and evolution of model organisms. *Nat Rev Genet* 2002, 3(11):838-849.

[23] Chen D, Toone WM, Mata J, Lyne R, Burns G, Kivinen K, Brazma A, Jones N, Bahler J: Global transcriptional responses of fission yeast to environmental stress. *Mol Biol Cell* 2003, 14(1):214-229.

[24] Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K: The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *Plant J* 2007, 50(2):347-363.

[25] Nelander S, Larsson E, Kristiansson E, Mansson R, Nerman O, Sigvardsson M, Mostad P, Lindahl P: Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals. *BMC Genomics* 2005, 6(1):68.

[26] Boulesteix AL, Strimmer K: Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theor Biol Med Model* 2005, 2:23.

[27] Hastie TJ, Tibshirani RJ: Generalized additive models. London: Chapman and Hall; 1990.

[28] Wood SN: Generalized Additive Models: Chapman & Hall; 2006.

[29] Vinogradov AE: Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet* 2004, 20(5):248-253.

[30] Eden E, Lipson D, Yogev S, Yakhini Z: Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol* 2007, 3(3):e39.

[31] David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM: A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* 2006, 103(14):5320-5325.

[32] Gerber AP, Herschlag D, Brown PO: Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol* 2004, 2(3):E79.

[33] Sekinger EA, Moqtaderi Z, Struhl K: Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol Cell* 2005, 18(6):735-748.

[34] Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ: Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* 2005, 309(5734):626-630.

[35] Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J: A genomic code for nucleosome positioning. *Nature* 2006, 442(7104):772-778.

[36] Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA et al: Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 2007, 39(3):311-318.

[37] Hertz-Fowler C, Peacock CS, Wood V, Aslett M, Kerhornou A, Mooney P, Tivey A, Berriman M, Hall N, Rutherford K et al: GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res* 2004, 32(Database issue):D339-343.

[38] Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* 2003, 301(5629):71-76.

[39] Gattiker A, Rischatsch R, Demougin P, Voegeli S, Dietrich FS, Philippsen P, Primig M: Ashbya Genome Database 3.0: a cross-species genome and transcriptome browser for yeast biologists. *BMC Genomics* 2007, 8:9.

[40] Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M et al: The Arabidopsis

Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res* 2003, 31(1):224-228.

[41] Hastie T.J T, R.J: Generalized Additive Models: Some Applications. *J Am Stat Assoc* 1987, 82(398):371-386.

[42] R: A Language and Environment for Statistical Computing [http://www.R-project.org]

Gene subset	Nr. of genes	Median length	Nr. of remaining genes	Median length	P-value
		Saccharomyces cerevisiae			
Gasch ESR genes	820	493	4915	450	2×10^{-3}
Causton CER genes	588	552	5147	445	5×10^{-11}
Thorsen up-regulated	317	560	5418	450	1×10^{-7}
Thorsen down-regulated	398	561	5337	448	3×10^{-7}
Essential genes[a]	1033	385	4702	468	8×10^{-9}
Ribosomal genes	185	551	5550	453	9×10^{-3}
Cell wall	69	899	5666	453	2×10^{-11}
Transporter activity	361	583	5374	449	4×10^{-9}
RNA metabolic process	262	340	5473	461	6×10^{-9}
		Ashbya gossypii			
Causton CER genes[b]	405	407	4278	337	4×10^{-5}
Essential genes[b]	1018	298	3665	356	4×10^{-9}
		Schizosaccharomyces pombe			
Chen CESR genes	704	863	4780	826	2×10^{-2}
Chen CERS conservative genes	237	1243	5247	816	7×10^{-15}
		Arabidopsis thaliana			
Kilian stress genes[c]	1152	1672	26584	1113	$< 10^{-16}$

[a]For growth in rich medium.

[b]Homologues based on the map in Dietrich et al. 2004 [21].

[c]Genes from the environmental stress microarray data in Kilian et al. 2004 [24] with an log fold-change more than 5 in at least one condition.

Table 1: Examples of gene sets from four different species with a significant difference in promoter length. Stress associated genes tend to have longer promoters than other genes, while promoters of essential genes are shorter. GO categories such as cell wall, transporter activity, and RNA metabolic process also have a different promoter length compared to the rest of the genes in the genome.

Transcription factor	Binding site	Hypergeometric test	Logistic regression	Proposed model
RPN4	GGTGGCAAA	6×10^{-19}	8×10^{-24}	6×10^{-22}
YAP2	MTTASTMAKC	5×10^{-12}	8×10^{-15}	1×10^{-12}
YAP7	MTKASTMA	6×10^{-12}	2×10^{-13}	4×10^{-9}
YAP1	TTAGTMAGC	3×10^{-8}	1×10^{-9}	3×10^{-8}
MSN2/4	AAGGGG	3×10^{-10}	3×10^{-11}	1×10^{-7}
YAP2/3/4/5	TTACTAA	3×10^{-9}	1×10^{-10}	1×10^{-7}
YAP1	TTASTMA	6×10^{-10}	8×10^{-10}	5×10^{-7}
MET31/32	AAACTGTGG	6×10^{-6}	1×10^{-6}	1×10^{-5}
SWI5	KGCTGR	6×10^{-6}	6×10^{-6}	2×10^{-2}
YOX1	YAATTA	9×10^{-5}	1×10^{-4}	1×10^{-1}

Table 2: The upper part of the table shows the top enriched transcription factors for up-regulated genes in the arsenite microarray dataset according to the proposed model. The median promoter length of the regulated genes was more than 100 bp longer than the promoter length of the other genes. As predicted by the simulation study, the p-values for the hypergeometric test and the logistic regression are non-conservative relative to the proposed method. The bottom part shows two examples of transcription factors having substantially low- to non-significant p-values when the proposed model is used.

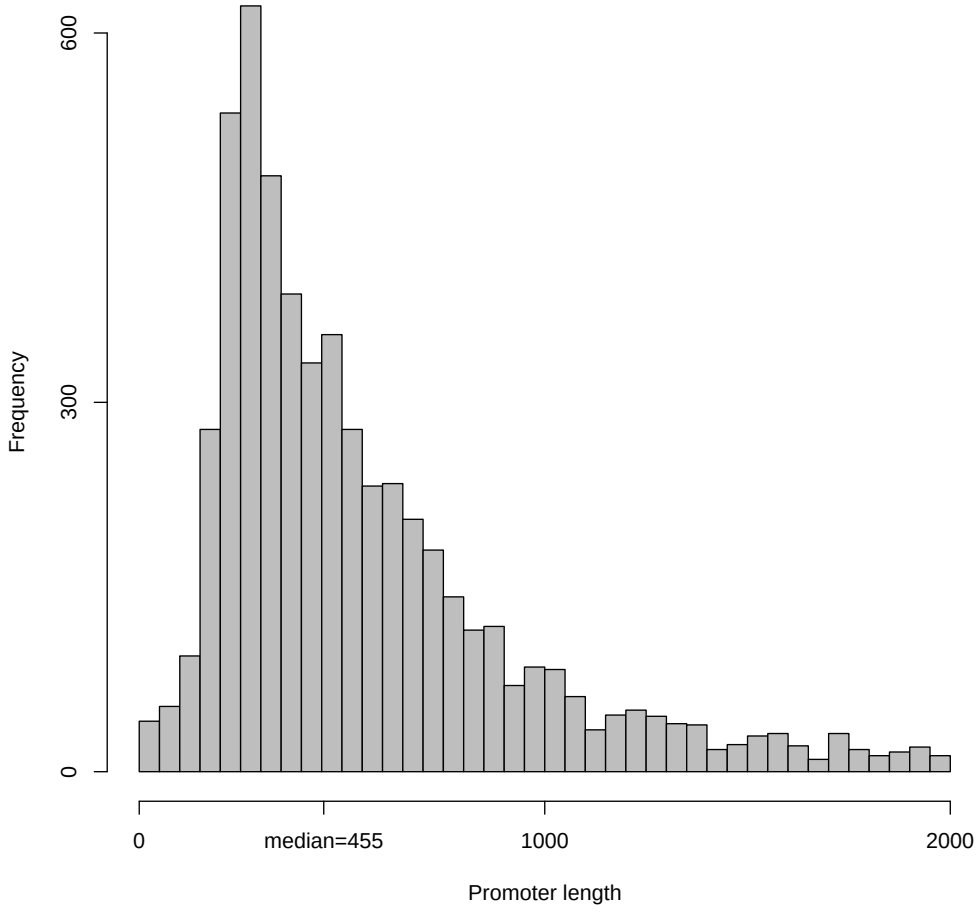


Figure 1: Histogram of the 5735 Saccharomyces cerevisiae promoters used in this study. The median promoter length is 455 bp and the distribution is asymmetric with a right tail. Roughly 5% of the promoters are longer than 2000 bp and thus not shown in this figure.

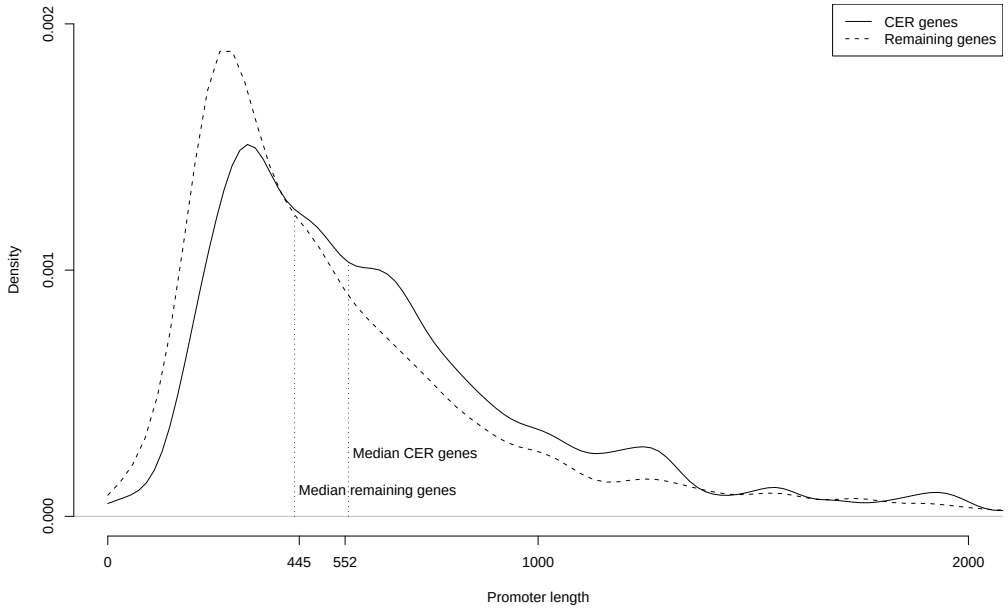


Figure 2: Estimated densities of the promoter length for the common environmental response (CER) genes (solid line) and the remaining genes (dashed line). The median length of the promoters for the CER genes was found to be significantly longer than for the remaining genes ($p = 5 \times 10^{-11}$).

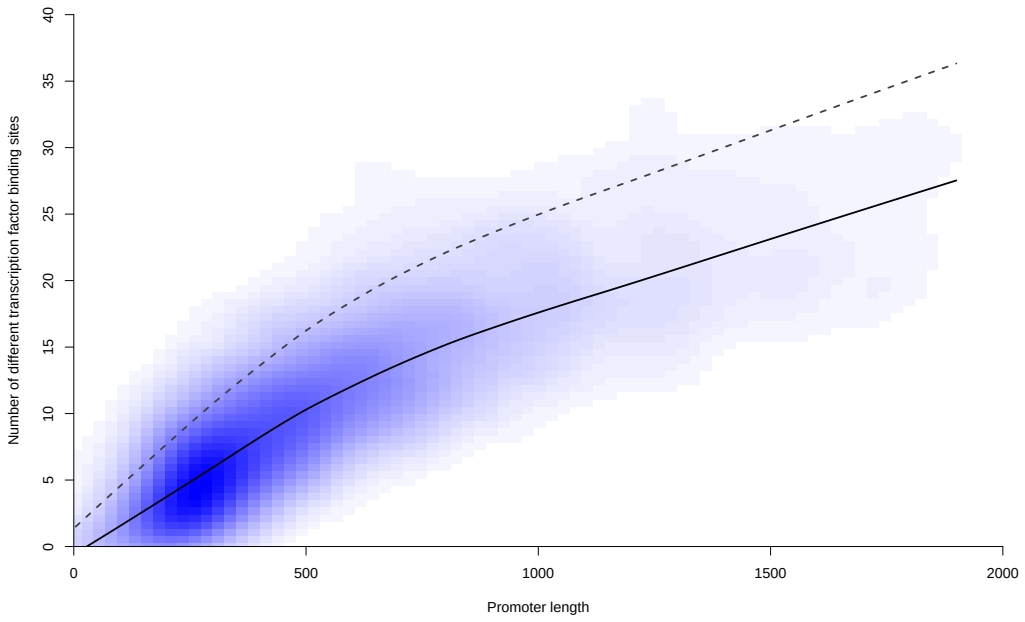


Figure 3: This figure shows the smoothed two-dimensional distribution of promoter length (x-axis) and number of different transcription factor binding sitesmotifs (y-axis). Higher intensity means more genes. Phylogenetic filtering, based on five closely relates yeast species, was used to reduce the number of spurious motifs. The general trend is shown both with and without phylogenetic filtering (solid and dashed curve respectively).

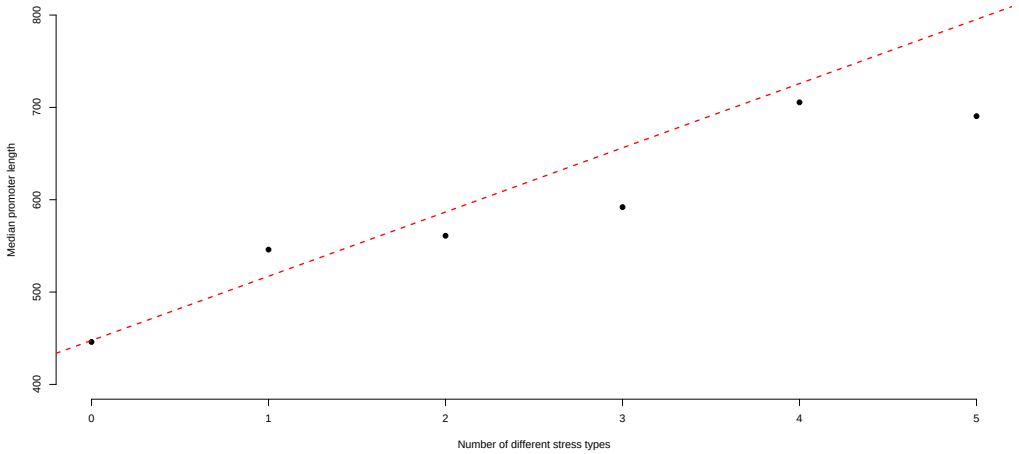


Figure 4: The number of stress conditions is correlated with the length of the promoter. Expression data from Causton et al. [9] were used to categorize genes into groups depending on the number of stress conditions they were regulated in (x-axis). For each group, the median promoter length was calculated (y-axis). The dashed line is a result of a linear regression with weights proportional to the number of genes in each group (5064, 449, 93, 32, 12, and 2 genes for the groups corresponding to 0 to 5 stress conditions, respectively). The trend is clearly positive ($p < 10^{-5}$) and genes regulated in several stress conditions have in general longer promoters. The correlation coefficient were estimated to 0.96 ($p = 10^{-3}$). Both p-values were calculated by permuting the promoters one million times while keeping the size of the groups fixed.

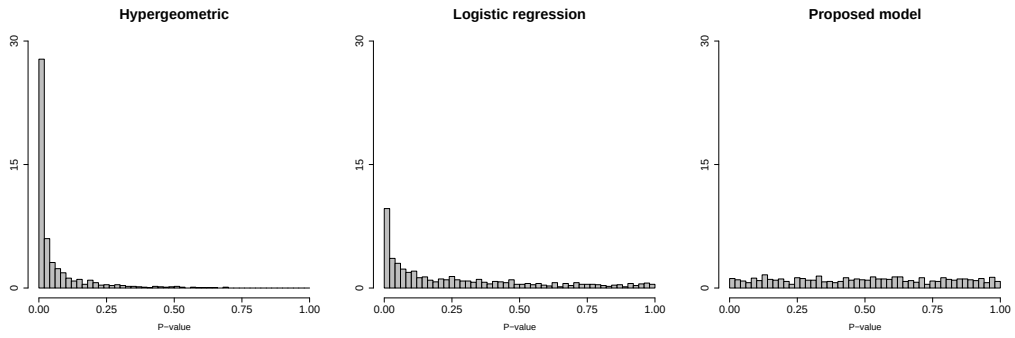


Figure 5: Results of the simulation study. Two sets of promoters with random sequences were created. The promoter length bias between the sets were set to 100 bp. Enrichment of a six-nucleotide motif was assessed in the set with long promoters using three different methods; hypergeometric test, logistic regression, and the proposed method. The hypergeometric test and the logistic regression result in several significant p-values, even though the promoters were randomly generated and no enrichment should exist. The proposed model, however, handles the bias well and reports no false positives.

Paper 5

Physiol Genomics 30: 35–43, 2007.
First published February 27, 2007; doi:10.1152/physiolgenomics.00236.2006.

Quantitative transcriptome, proteome, and sulfur metabolite profiling of the *Saccharomyces cerevisiae* response to arsenite

Michael Thorsen,[1] Gilles Lagniel,[2] Erik Kristiansson,[3] Christophe Junot,[4] Olle Nerman,[3] Jean Labarre,[2] and Markus J. Tamás[1]

[1]*Department of Cell and Molecular Biology/Microbiology, Gothenburg University, Gothenburg, Sweden;* [2]*Service de Biochimie et de Génétique Moléculaire, Département de Biologie Joliot-Curie, Commissariat à l'Énergie Atomique (CEA)/Saclay, Gif-sur-Yvette, France;* [3]*Department of Mathematical Statistics, Chalmers University of Technology, Gothenburg, Sweden; and* [4]*Service de Pharmacologie et Immunologie, Département de Recherche, Médicale, CEA/Saclay, Gif-sur-Yvette, France*

Submitted 27 October 2006; accepted in final form 22 February 2007

Thorsen M, Lagniel G, Kristiansson E, Junot C, Nerman O, Labarre J, Tamás MJ. Quantitative transcriptome, proteome, and sulfur metabolite profiling of the *Saccharomyces cerevisiae* response to arsenite. *Physiol Genomics* 30: 35–43, 2007. First published February 27, 2007; doi:10.1152/physiolgenomics.00236.2006.—Arsenic is ubiquitously present in nature, and various mechanisms have evolved enabling cells to evade toxicity and acquire tolerance. Herein, we explored how *Saccharomyces cerevisiae* (budding yeast) respond to trivalent arsenic (arsenite) by quantitative transcriptome, proteome, and sulfur metabolite profiling. Arsenite exposure affected transcription of genes encoding functions related to protein biosynthesis, arsenic detoxification, oxidative stress defense, redox maintenance, and proteolytic activity. Importantly, we observed that nearly all components of the sulfate assimilation and glutathione biosynthesis pathways were induced at both gene and protein levels. Kinetic metabolic profiling evidenced a significant increase in the pools of sulfur metabolites as well as elevated cellular glutathione levels. Moreover, the flux in the sulfur assimilation pathway as well as the glutathione synthesis rate strongly increased with a concomitant reduction of sulfur incorporation into proteins. By combining comparative genomics and molecular analyses, we pinpointed transcription factors that mediate the core of the transcriptional response to arsenite. Taken together, our data reveal that arsenite-exposed cells channel a large part of assimilated sulfur into glutathione biosynthesis, and we provide evidence that the transcriptional regulators Yap1p and Met4p control this response in concert.

DNA microarray; proteomics; glutathione

ARSENIC IS A HIGHLY TOXIC metalloid that considerably threatens the environment and human health. The most striking example is the epidemic of arsenic poisoning observed in Bangladesh and West Bengal, where arsenic contaminates the drinking water through geological sources and thereby affects millions of people (10, 26). Chronic arsenic exposure causes cardiovascular diseases, neurological disorders, and liver injury and is associated with cancers of the skin, bladder, liver, and lung. Despite its toxicity, arsenic trioxide is currently used as a treatment for acute promyelocytic leukemia, and it might also be employed against other hematological and solid cancers (7, 28).

All organisms have been exposed to toxic agents since the origin of life, and tolerance mechanisms arose early during evolution. The unicellular model eukaryote *Saccharomyces cerevisiae* (budding yeast) evades arsenic toxicity by increasing efflux of trivalent arsenite [As(III)] through the plasma membrane protein Acr3p (12, 38), by sequestering glutathione-conjugated As(III) in the vacuole through the ATP binding cassette (ABC) transporter Ycf1p (12) and by reducing As(III) influx through the aquaglyceroporin Fps1p (34, 39). In addition, cells may acquire tolerance by adjusting cytosolic redox and glutathione levels in response to As(III) (31). Two AP-1-like transcription factors have been shown to be important for yeast As(III) tolerance; Yap8p controls expression of the arsenic-specific detoxification genes *ACR2* and *ACR3*, whereas Yap1p controls transcription of genes encoding proteins with antioxidant properties. In addition, Yap1p contributes to *YCF1* and *ACR3* control (17, 24, 40).

Sulfur assimilation is essential for all organisms. In yeast, extracellular sulfate is taken up and metabolized through the sulfate assimilation pathway where sulfide is the reduced end product (33). Sulfide can then either go through the methyl cycle or into the cysteine/glutathione biosynthesis pathway (Fig. 1). Hence, the fate of assimilated sulfur is principally biosynthesis of the sulfur-containing amino acids methionine and cysteine and the low-molecular-weight thiol molecules *S*-adenosylmethionine and glutathione (GSH) (33). GSH is a key factor in the cell's defense against oxidative stress and metal toxicity. GSH may detoxify metals by *1*) chelation followed by vacuolar sequestration; *2*) protecting against oxidation caused by metals, since GSH serves as the main redox buffer of the cell; and *3*) binding to reactive sulfhydryl groups on proteins (protein glutathionylation), thereby protecting them from irreversible metal binding and/or oxidative damage (13, 27). Whether protein glutathionylation occurs in response to metals has not been investigated. Transcription of the genes encoding enzymes in the sulfate assimilation/GSH biosynthesis pathways is principally controlled by the transcriptional activator Met4p. Met4p is recruited to target promoters by the DNA binding proteins Met28p, Met31p, Met32p, and Cbf1p, forming the complexes Met4p-Met31p/Met32p and Met4p-Cbf1p-Met28p (33). Interestingly, Met4p has been shown to play a central role in cadmium tolerance by controlling expression of sulfur assimilation and GSH biosynthesis genes (8).

The aim of this study was to gain detailed insight into the yeast response to arsenite. We demonstrate that As(III)-exposed cells channel a large part of assimilated sulfur into GSH

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: M. J. Tamás, Dept. of Cell and Molecular Biology/Microbiology, Göteborg Univ., S-405 30 Göteborg, Sweden (e-mail: markus.tamas@gmm.gu.se).

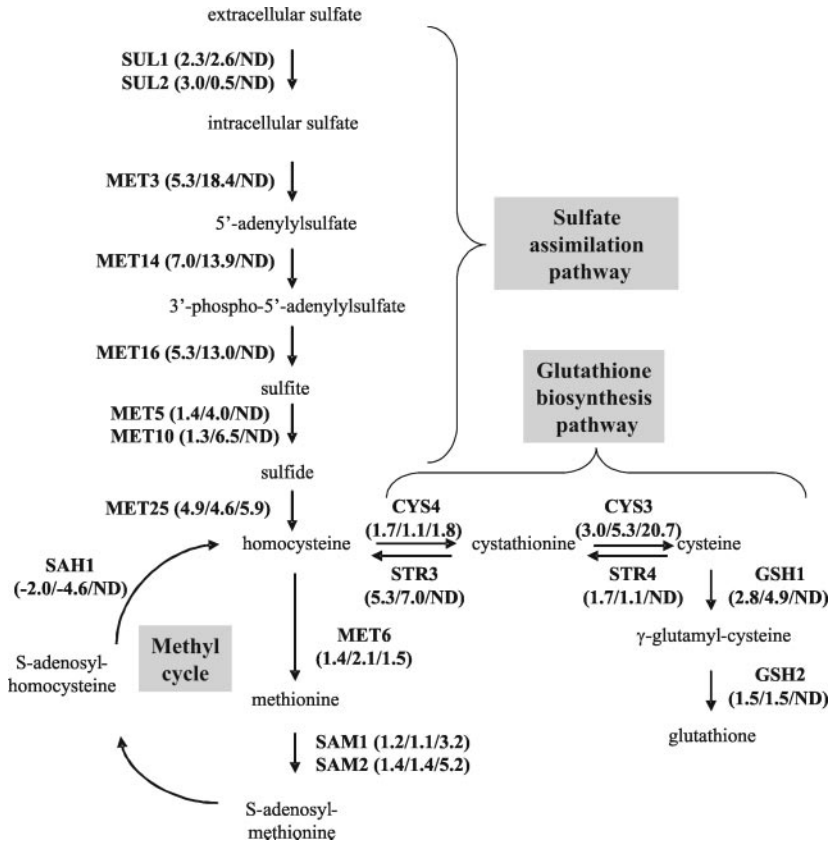


Fig. 1. Outline of the sulfur assimilation and glutathione (GSH) biosynthesis pathways in *Saccharomyces cerevisiae*. Induction levels (fold induction) of genes/proteins in the pathways in response to arsenite are indicated within brackets as follows: gene expression at 0.2 mM trivalent arsenite [As(III)] for 1 h, gene expression at 1.0 mM As(III) for 1 h, protein level at 0.2 mM As(III) for 4 h. ND, not done.

biosynthesis and provide evidence that the transcriptional regulators Yap1p and Met4p control this response in concert.

MATERIALS AND METHODS

Yeast strains, plasmids, and growth conditions. S. cerevisiae strains used in this study are summarized in Table 1. All deletion mutants were constructed according to Ref. 14, and metal sensitivity assays were carried out as previously described (39). The metals used were sodium arsenite (Sigma), cadmium chloride (Sigma), and potassium antimonyl tartrate (Acros). Yeast strains were grown at 30°C on minimal YNB medium (0.67% yeast nitrogen base) supplemented with auxotrophic requirements and 2% glucose as a carbon source or on SC medium (YNB containing 2% glucose).

RNA isolation, cDNA synthesis, microarray hybridization, and analysis. Total RNA was isolated as described previously (4) from exponentially growing yeast cells that were either untreated or exposed to sodium arsenite; 20 μg of total RNA were primed with 3 μg of random hexamer (Invitrogen) and 3 μg of anchored oligo(dT)$_{20}$ primer (ABgene) and labeled in a reverse transcription reaction with Cy3-dUTP or Cy5-dUTP (Amersham Pharmacia Biotech) in a volume of 30 μl, according to standard protocols (**http://cmgm.stanford.edu/pbrown**). Labeled cDNA was cleaned (microcon YM-30 columns, Milipore), combined, vacuum-dried, and resuspended in 80 μl of DIGeasy hybridization buffer (Roche Diagnostics). The hybridization mix was placed at 100°C for 2 min and then at 37°C for 30 min. Before hybridization, the microarray chip (Yeast 6.4k array from University Health Network Microarray Centre, Toronto, Canada) was prehybridized with 1% BSA in DIGeasy hybridization buffer at 42°C for 1 h. Hybridization was performed at 42°C for 12–18 h. After hybridization, the slides were washed at room temperature in 2× SSC plus 0.1% sodium dodecyl sulfate for 5 min, in 1× SSC for 5 min, and in 0.1× SSC for 5 min and then blow dried with N_2. The slides were scanned (VersArray ChipReader, Bio-Rad) at laser intensity and photomultiplier tube voltage settings giving the best dynamic range for each chip in the respective channel. Image segmentation and spot quantification were performed with ImaGene v.6 software (BioDiscovery, CA). The microarray data were analyzed using the linear models for microarray data (LIMMA) package (**http://www.bioconductor.org**) in the statistical language R (**http://www.R-project.org**). The data were normalized by subtracting a loess line from the M-value to remove intensity-dependent trends (41). The genes were ranked by the moderated *t*-statistic to avoid

Table 1. *Saccharomyces cerevisiae strains used in this study*

Strain	Genotype	Source
W303-1A	*MAT***a** *ura3-1 leu2-3/112 trp1-1 his3-11/15 ade2-1 can1-100 GAL SUC2 mal0*	Ref. 32
RW124	W303-1A *yap1*Δ::*loxP*	Ref. 40
CC849-1B	*MAT***a** *his3 leu2 trp1 ura3 met4*Δ::*TRP1*	Ref. 29
RW104	W303-1A *acr3*Δ::*loxP-kanMX-loxP*	Ref. 29
YPDahl166	W303-1A *acr3*Δ::*KanMX met4*Δ::*TRP1*	Present study

false positives (30). Each comparison consists of at least three independent experiments. Minimum information about a microarray experiment (MIAME)-compliant microarray data have been deposited in the microarray database Gene Expression Omnibus (GEO; **http://www.ncbi.nlm.nih.gov/geo/**) (6) with the accession number GSE6129.

Northern blot analysis. Northern analysis was performed as previously described (40). Exponentially growing cells were exposed to 0.2 mM sodium arsenite, and total RNA was extracted at the indicated time points. Blots were hybridized with ^{32}P-labeled PCR fragments of *MET3*, *MET25*, and *MET14*. 18S rRNA was used as a loading control. Primer sequences are available on request.

Proteome analysis. Exponentially growing yeast cells in YNB medium were exposed to 0.2 mM sodium arsenite for 1 h and then labeled for 30 min with [^{35}S]methionine. Protein extraction, two-dimensional gel electrophoresis, and gel analysis were performed as previously described (36).

Metabolite measurements and metabolic flux analysis. Exponentially growing yeast cells (in YNB medium) were exposed to 0.2 mM sodium arsenite. Cells were collected before and at the indicated time points after treatment, and metabolites were extracted as previously described (20). The intracellular concentrations of sulfur metabolites were determined by liquid chromatography/mass spectrometry/mass spectrometry (LC/MS/MS) using a pool of ^{15}N metabolites as internal standards (20). GSH and protein synthesis rates were determined as previously described (19).

Testing for overrepresented transcription factor binding sites. In the data set by Cliften et al. (3), promoters from >5,200 genes in *S. cerevisiae* together with corresponding promoters from orthologous genes in *S. mikate*, *S. kudriavzevii*, *S. bayanus*, *S. kluyveri*, and *S. castelli* are available. For each transcription factor binding site described in the *Saccharomyces* Genome Database (SGD) (**http://www.yeastgenome.org**), we searched for genes with the consensus motif present in the promoter of *S. cerevisiae* and in at least one-half of the available promoters of the other species. To test whether these hits were spread equally among all genes or whether they were overrepresented among the regulated genes from the microarray study [0.2 mM As(III), 1 h], a generalized additive model (GAM) with a logic link was used (16, 37)

$$\log\frac{\text{Prob}(X_g = 1)}{1 - \text{Prob}(X_g = 1)} = \beta Y_g + f(L_g)$$

Here, Prob is probability, X_g is a dichotomous random variable indicating whether gene g has the motif present (according to the search above), and Y_g is a dichotomous variable indicating whether gene g is regulated (fold change >1). L_g is the length of the promoter in *S. cerevisiae*, and f is an unknown smooth function. The coefficient β and the function f are estimated from data using the mgcv package (37) in the statistical language R (**http://www.R-project.org**), and P value is calculated based on the test of $\beta = 0$. The reason for using a GAM instead of a less complex hypergeometric test is to avoid problems with unequal promoter length between the regulated genes and the rest of the genome. Indeed, the average promoter lengths in *S. cerevisieae* of the up- and downregulated genes were 576 nucleotides and 579 nucleotides, respectively, which should be compared with an average length of 493 nucleotides for the rest of the genes in the genome. Thus the use of a hypergeometric test in this situation would lead to biased P values.

RESULTS

Transcriptional profiling of As(III)-exposed cells. The transcriptional response of yeast cells exposed to sodium arsenite was analyzed using two different concentrations: 0.2 mM, which has a moderate effect on growth, and 1 mM, which severely affects growth of wildtype cells [the minimal inhibitory concentration of As(III) on the W303-1A strain used here is 1.2 mM (39)]. Total RNA was isolated at various time points, and gene expression profiles were analyzed using cDNA microarrays. Exposing cells to 0.2 mM As(III) for 1 h altered the expression of 761 genes (differentially expressed >2-fold); mRNA levels for 428 genes were less abundant in exposed cells, whereas mRNA levels for 333 genes were more abundant (Supplemental Table S1; supplemental data are available at the online version of this article). The bulk of downregulated genes encode functions related to protein biosynthesis, i.e., genes encoding rRNA, tRNA, ribosomal proteins, and elongation factors. Among those whose expression is stimulated by As(III), we found genes related to arsenic detoxification, oxidative stress defense, redox maintenance, and proteolytic activity as well as genes encoding structural components of the sulfur assimilation and GSH biosynthesis pathways (Supplemental Table S1). The response to As(III) was largely transient: mRNA levels of most responsive genes started to change within the first 15 min of exposure, peaked at 60 min, and then stabilized at a new steady-state expression level once cells had adapted (Supplemental Table S2, *A–D*; see also Fig. 4*A*).

When comparing the transcript profiles of cells exposed to 0.2 or 1 mM As(III), we found similar responses in terms of the identity of the genes whose expression was either up- or downregulated (Supplemental Tables S1 and S2*C*). However, the lower concentration triggered a faster transcriptional response than the higher concentration. For example, the arsenic detoxification genes *ACR2* and *ACR3* responded earlier at 0.2 mM than at 1 mM As(III). In contrast, the amplitude of gene expression levels was generally larger at the higher concentration (compare Supplemental Tables S1 and S2). During the course of this study, Haugen et al. (17) reported a detailed analysis of the transcriptional response to arsenite. Since our gene expression data reported here largely confirm their results, a full analysis of gene expression changes will not be provided. Instead, we focus on characterizing the response and the control of the sulfur assimilation/GSH biosynthesis pathways in more detail.

As(III) stimulates expression of sulfur assimilation and GSH biosynthesis genes. We noted that As(III) strongly stimulated expression of genes encoding components of the sulfur assimilation and GSH biosynthesis pathways. In fact, mRNA levels of genes encoding basically all the enzymatic steps required for sulfate uptake, its reduction to sulfide, and further conversion into cysteine and GSH were elevated (Fig. 1; Supplemental Tables S1 and S2). In general, expression of these genes was induced ~2- to 5-fold at 0.2 mM As(III), whereas expression of some genes (*MET3*, *MET14*, and *MET16*) was induced up to 15- to 20-fold at 1.0 mM As(III). In addition to the sulfate permease-encoding genes *SUL1* and *SUL2*, expression of the high-affinity *S*-methylmethionine permease-encoding gene *MMP1* as well as *MUP1* and *MUP3*, encoding methionine permeases, was enhanced (Supplemental Tables S1 and S2). *CYS4* and *GSH2* were the only two genes in the pathway whose expression was not stimulated at least twofold by As(III). Similarly, expression of genes in the methyl cycle (*MET6*, *SAM1*, *SAM2*) was not significantly enhanced by As(III), whereas *SAH1* expression was reduced (Fig. 1; Supplemental Tables S1 and S2).

Proteome analysis of As(III)-exposed cells. To analyze whether the observed transcriptional changes in response to As(III) would be translated into similar changes at the proteome level, we performed two-dimensional gel analysis and quantified the abundance of selected proteins. Proteome analysis confirmed increased levels of proteins in the sulfur assimilation and GSH biosynthesis pathways (Fig. 1; Supplemental Table S3). In particular, Cys3p levels increased very strongly. On the other hand, Met6p levels were largely unaffected. Proteome analysis also confirmed enhanced levels of several proteins with antioxidant properties including Ahp1p and Sod2p, and the amount of overoxidized Tsa1p increased during As(III) exposure (Supplemental Table S3). Taken together, there appears to be a good correlation between the response of the transcriptome and the proteome in As(III)-treated cells, at least when it comes to the genes/proteins in the sulfur assimilation/GSH biosynthesis pathways. Importantly, the data furthermore suggested that As(III)-exposed cells may channel assimilated sulfur into cysteine and probably GSH biosynthesis.

Kinetics of sulfur metabolites in response to As(III). To address how As(III) affects the metabolites of the GSH biosynthesis pathway, we monitored sulfur metabolite levels in a time course experiment. The pools of homocysteine, cystathionine, cysteine, γ-glutamylcysteine, and GSH started to increase within the first 30 min of exposure and continued to rise over time (Fig. 2*A*). The strongest effect was observed for γ-glutamylcysteine, which is the precursor of GSH, increasing >10-fold after 3 h. Also, the total GSH content increased considerably (~7-fold after 4 h). In contrast, methionine levels remained largely unchanged, while *S*-adenosyl-homocysteine increased slightly (2-fold after 4 h). Collectively, these results are consistent with the notion that As(III) stimulates the sulfur pathway and that assimilated sulfur may be redirected toward GSH biosynthesis in As(III)-treated cells. We also noted that the ratio of oxidized to reduced GSH (GSSG/GSH) remained constant throughout the course of this experiment (Fig. 2*B* and below).

Flux in the sulfur pathway increases in response to As(III). The strong boost of the GSH pool, the end product of the sulfur pathway, suggested that the flux in the pathway may increase in As(III)-challenged cells. To test this, we performed a direct measurement of the GSH synthesis rate before and during As(III) exposure. The method was based on [^{35}S]sulfate label-

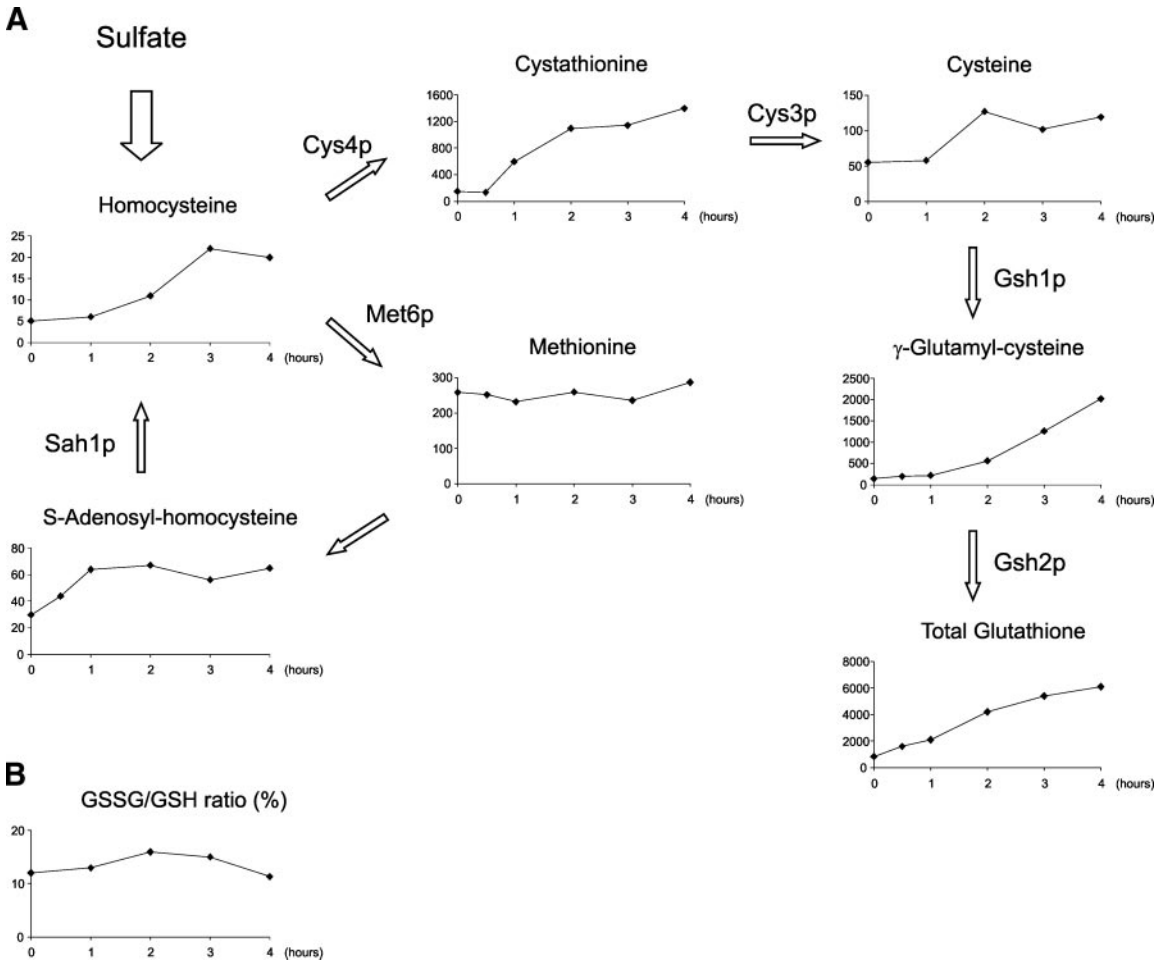


Fig. 2. Kinetic response of sulfur metabolite pools in response to arsenite. *A*: exponentially growing yeast cells (in minimal medium) were exposed to 0.2 mM As(III). Cell aliquots were collected at different time points after As(III) treatment (0, 30 min, 1 h, 2 h, 3 h, and 4 h) and processed for sulfur metabolite analysis. Intracellular sulfur metabolite concentrations are given in μM (note the differences in scales of the *y*-axes). With the exception of cysteine and homocysteine (analysis performed only one time), the values represent the mean of at least two determinations. Typical standard deviations are <20%. *B*: ratio of oxidized to reduced GSH (GSSG/GSH) was measured.

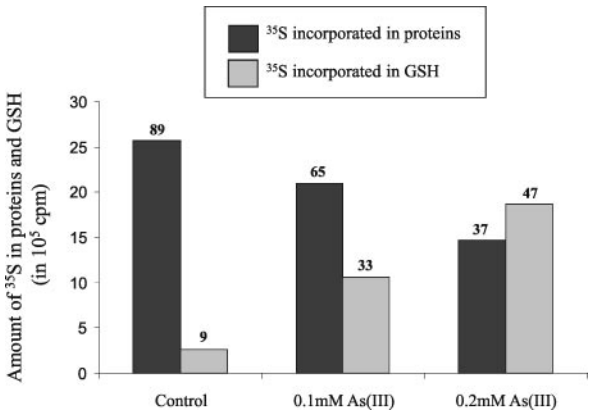


Fig. 3. Balance of sulfate utilization in As(III)-exposed cells. Exponentially growing cells in minimal medium containing 1 mM sulfate were divided into three culture aliquots: one untreated culture (control) and two cultures exposed to the indicated As(III) concentration. After 1 h of exposure, cells were labeled with [^{35}S]sulfate for 4 h. Results are reported in amount of radioactivity (cpm) incorporated into proteins or GSH. Nos. at *top* of bars are percentage of assimilated sulfate in proteins or GSH.

ing and quantification of newly synthesized GSH and proteins by counting the radioactivity in GSH and protein fractions (19). This analysis evidenced a strong increase in GSH synthesis in As(III)-exposed cells. In particular, we observed a sevenfold raise in GSH synthesis following exposure to 0.2 mM As(III) (Fig. 3). Concurrently, we found a significant reduction of sulfur incorporation into proteins, which may reflect a decrease of the protein synthesis rate and/or a decrease of the sulfur amino acid utilization in the global proteome in response to As(III). We conclude that As(III)-exposed cells channel a large part of assimilated sulfur into GSH biosynthesis.

Comparative genomics reveals transcription factor binding sites in the promoters of As(III)-regulated genes. We next sought to identify transcription factors that mediate the transcriptional response to As(III). For this, we analyzed the promoter sequences of As(III)-regulated genes with the aim of revealing motifs that are common among groups of genes that display similar expression patterns. Such analyses can reveal regulatory elements that may function as transcription factor binding sites involved in activation/repression of gene expression. We found 132 regulatory motifs reported in the literature (and curated by the SGD; **http://www.yeastgenome.org**) related to 90 different transcription factors. Using the complete current knowledge of DNA binding sites for transcriptional regulators, one can do an unbiased search for proteins that may control the transcriptional response to As(III). To this end, we searched for conserved stretches in the promoters of five related *Saccharomyces* species (*S. mikate, S. kudriavzevii, S. bayanus, S. kluyveri*, and *S. castelli*) to assess whether a motif in a given *S. cerevisiae* promoter is conserved in the corresponding promoters in these related species.

In the promoter sequence of all genes with transcripts that were found upregulated more than twofold, we discovered binding motifs of 13 transcription factors to be highly overrepresented compared with the promoters of the entire genome (Table 2, *top*). Of those, Yap1p was previously shown to be regulated by and implicated in the transcriptional response to arsenite (17, 24, 40). Since Yap2p-Yap5p and Yap7p share a DNA binding site with Yap1p (9), it is no surprise that these proteins were also identified. In addition, our analysis pinpointed Cbf1p, Met31p, and Met32p, which together with Met4p control expression of sulfur assimilation/GSH biosynthesis-encoding genes. For genes with at least twofold downregulated transcripts, we discovered binding motifs for six transcription factors to be highly overrepresented compared with the promoters of the entire genome (Table 2, *bottom*). Three of these regulate expression of genes encoding ribosomal proteins.

Expanding this analysis to include combinatorial control, we found 50 promoters in the entire genome that have both a

Table 2. *Transcription factors with overrepresented DNA binding site in the promoters of up- and downregulated genes in arsenite-challenged cells*

Transcription Factor	*P* Value	Target Genes Encode Proteins with Function in . . .
Transcription factors with overrepresented DNA binding site in the promoters of upregulated genes		
Rpn4p	1.94×10^{-23}	Proteasome function
Yap2p (Cad1p)	9.12×10^{-11}	Stress response
Yap7p	2.91×10^{-09}	Unknown
Msn4p	6.36×10^{-09}	Environmental stress response
Msn2p	6.36×10^{-09}	Environmental stress response
Yap3p	6.65×10^{-09}	Unknown
Yap5p	6.65×10^{-09}	Unknown
Yap4p (Cin5p)	6.65×10^{-09}	Unknown
Yap1p	2.14×10^{-06}	Resistance to oxidative stress
Adr1p	2.17×10^{-05}	Peroxisomal function, utilization of alternative carbon sources
Met32p	1.34×10^{-04}	Sulfur metabolism
Met31p	1.34×10^{-04}	Sulfur metabolism
Cbf1p	3.64×10^{-03}	Sulfur metabolism
Transcription factors with overrepresented DNA binding site in the promoters of downregulated genes		
Sfp1p	5.92×10^{-21}	Ribosomal function
Rap1p	4.61×10^{-17}	Ribosomal function
Fhl1p	3.33×10^{-08}	Ribosomal function
Spt23p	1.31×10^{-05}	Unknown
Aft2p	1.33×10^{-05}	Iron homeostasis, resistance to oxidative stress
Hap1p	6.27×10^{-04}	Response to cellular heme and oxygen levels

conserved Yap1p binding site and a binding site for one of the Met4p-recruiting factors. Of the 50 genes controlled by these promoters, the transcription of 11 was upregulated (at least 2-fold) in response to As(III). Interestingly, 8 of these 11 promoters regulate genes with products related to sulfur metabolism (*MET1*, *MET2*, *MET3*, *MET14*, *MET16*, *GSH1*, *SUL2*, and *GTO1*). Transcription of *GSH1* has previously been reported to be controlled by both Met4p and Yap1p in response to cadmium (5) and in response to GSH depletion (35). *CYS3* has been shown to be regulated by Yap1p in response to H_2O_2 (21) and by Met4p in response to cadmium (8). The other three genes (*HSV2*, *ICY2*, and *BNA3*) encode products with no apparent role in sulfur metabolism, although *ICY2* and *BNA3* are upregulated under sulfur starvation (2).

Yap1p and Met4p control sulfur assimilation/GSH biosynthesis pathway genes in concert. Comparative genomics strongly suggested that Yap1p and Met4p control the sulfur assimilation/GSH biosynthesis pathways in concert. To test this, and to identify gene targets of Yap1p and Met4p under As(III) exposure, we compared global gene expression profiles of *yap1*Δ and *met4*Δ mutants to that of the wildtype using microarray analysis.

Seventy-two genes displayed twofold lower expression in *yap1*Δ compared with wildtype at 0.2 mM As(III), whereas the number of Yap1p-dependent genes was larger at 1 mM As(III) (Supplemental Table S4, *A* and *B*). Many of those genes are known Yap1p targets and encode antioxidant defense functions. Importantly, lack of *YAP1* also reduced As(III)-stimulated expression of most genes related to sulfur uptake and assimilation at 0.2 mM As(III) (Fig. 4*A*; Supplemental Table S4*A*). Curiously, induced expression of these genes appeared to be largely Yap1p independent at the higher concentration.

We next analyzed the transcriptome of *met4*Δ and found 70 genes with at least twofold lower expression than in the wildtype at 0.2 mM As(III) (Supplemental Table S5); most of these genes encode functions in sulfur metabolism and functions related to protein biosynthesis (Fig. 4*A* and Supplemental Table S5). Remarkably, *MET4* deletion also resulted in enhanced expression of many genes; these genes appear to be stress responsive, and many of them are actually controlled by Yap1p. Hence, lack of *MET4* may result in hyperactivation of Yap1p. In fact, we previously observed the same phenomenon in cells with elevated cellular As(III) levels, e.g., in cells lacking *ACR3* (40). Although *MET4* deletion does not appear to affect As(III) uptake/efflux systems or to alter intracellular As(III) levels (data not shown), the amount of free As(III) is probably higher in *met4*Δ, since GSH synthesis is likely to be defective in this mutant. Finally, Northern blot analyses confirmed that *MET3*, *MET14*, and *MET25* transcripts are elevated in response to 0.2 mM arsenite and that As(III)-stimulated expression of these genes depends on both Yap1p and Met4p (Fig. 4*B*). However, while *YAP1* deletion only affected induction, *MET4* deletion affected both basal mRNA levels and As(III)-induced transcription of these genes (Fig. 4*B*). Taken together, our data demonstrate that Yap1p and Met4p activate gene expression in the sulfur/GSH pathways in concert when cells are exposed to a moderate As(III) concentration, whereas Met4p may play a more prominent role during severe As(III) stress.

Mutations in the sulfur/GSH pathways render cells As(III) sensitive. To assess the physiological importance of the sulfur/GSH pathways for As(III) detoxification/tolerance, we scored growth of mutants lacking pathway components in the presence of arsenite. We found that deletion of *MET3, MET14, MET16, MET25, CYS3*, or *CYS4* sensitized cells to As(III), whereas *gsh1*Δ cells were found hypersensitive (data not shown), confirming the importance of a functional sulfur/GSH pathway for As(III) tolerance. Also, *met4*Δ cells displayed As(III) sensitivity, although this sensitivity was not as strong as in the presence of cadmium or antimonite [Sb(III)] (Fig. 5). We hypothesized that the Met4p-mediated response of the sulfur/GSH pathways might be masked by the action of Acr3p, which efficiently mediates As(III) efflux. Corresponding cadmium- or Sb(III)-specific efflux systems have not been described. To test this, we scored growth of *acr3*Δ *met4*Δ cells in the presence of metals. Indeed, growth tests evidenced a clear additive As(III) sensitivity of the *acr3*Δ *met4*Δ double mutant compared with

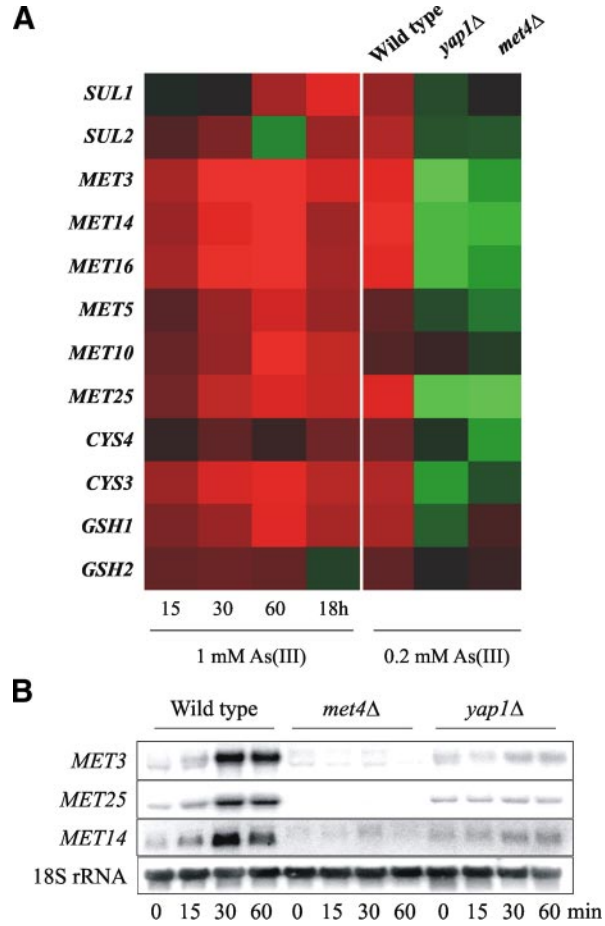


Fig. 4. Yap1p and Met4p control As(III)-induced transcription of genes in the sulfur assimilation and GSH biosynthesis pathways. *A*: heat map showing transient expression changes of genes in the sulfur assimilation and GSH biosynthesis pathways in response to 1 mM As(III) at 15, 30, and 60 min and 18 h (*left*) and the role of Met4p and Yap1p for their induced expression in response to 0.2 mM As(III) for 1 h (*right*). Red, upregulated; green, downregulated; black, unchanged. For further details, see Supplemental Tables S1, S2, S4, and S5. *B*: Northern blot analysis of total RNA extracted from wildtype, *met4*Δ, and *yap1*Δ cells before and at the indicated time points after exposure of cells to 0.2 mM As(III). Filter was hybridized to ^{32}P-labeled DNA fragments recognizing the indicated genes. 18S rRNA was used as a loading control.

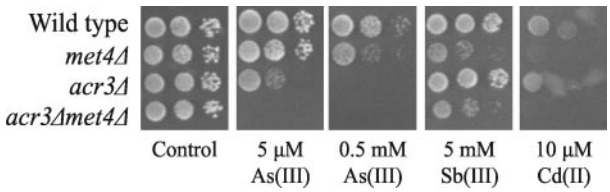


Fig. 5. Sulfur/GSH metabolism contributes to As(III) tolerance. Tenfold serial dilutions of wildtype, *met4*Δ, *acr3*Δ, and *acr3*Δ *met4*Δ cultures were spotted on SC-agar plates supplemented with the indicated metal. Growth was monitored after 2–3 days at 30°C.

each single mutant, whereas no additive growth defect on *acr3*Δ *met4*Δ was observed in the presence of Sb(III) or Cd(II) (Fig. 5). We conclude that efflux through Acr3p is likely to represent the primary As(III) defense mechanism and that cells lacking an efflux system become critically dependent on the sulfur/GSH-mediated detoxification system.

DISCUSSION

In this study, we explored the response of *S. cerevisiae* to arsenite. By combining transcriptome, proteome, and metabolite profiling with comparative genomics and physiology, we demonstrate that stimulation of the sulfur assimilation/GSH biosynthesis pathways represents an important step in cellular As(III) tolerance acquisition and that the transcription factors Yap1p and Met4p control this response in concert.

Role of the sulfur/GSH pathways under As(III) exposure. Yeast cells responded to As(III) by stimulating the sulfur assimilation/GSH biosynthesis pathways at both gene and protein levels. Furthermore, a rapid increase of the pools of all the intermediates in the GSH biosynthesis pathway was observed. The metabolites continued to accumulate over time, and the sulfur may eventually be metabolized all the way to GSH. This finding was further underscored by a sulfur flux analysis that evidenced a strong increase in GSH synthesis concomitantly with a significant reduction of sulfur incorporation into proteins. Hence, As(III)-exposed cells channel a large part of the assimilated sulfur into GSH biosynthesis. This response is likely to provide more GSH for metal conjugation, for cellular redox buffering, and possibly also for protein glutathionylation. The physiological importance of the sulfur/GSH pathway for As(III) tolerance is highlighted by the sensitivity of mutants in the pathway.

In mammals, arsenite exposure may lead to increased production of reactive oxygen species (22). Therefore, it is likely that cells launch defense mechanisms that protect the cytosol from oxidation. Indeed, we observed increased expression of several genes encoding antioxidant functions including *GSH1* and *GLR1* (GSH reductase; Ref. 40 and present study). Similarly, we observed increased levels of proteins with antioxidant properties. However, despite a strong increase in GSH levels, the ratio of oxidized to reduced GSH (GSSG/GSH) was unexpectedly found to remain constant in As(III)-exposed cells. Since GSH is produced in its reduced form, this finding might indicate that a continuously increasing number of GSH molecules are oxidized in response to As(III).

By comparing As(III) sensitivities of deletion mutants, one may get insight into the relative importance of various defense systems. The most As(III)-sensitive strains tested by us were *yap8*Δ and *acr3*Δ (39, 40), suggesting that the main line of defense is probably the Yap8p/Acr3p-mediated response. *YAP1* deletion caused a moderate sensitivity (40), implying that the antioxidative defense and/or the Ycf1p-mediated defense is less critical for tolerance. Here, we showed that *met4*Δ cells were moderately As(III) sensitive, probably because the Acr3p-mediated defense system is generally sufficient to ensure almost wildtype tolerance. Indeed, when *MET4* was deleted in an *acr3*Δ background, we observed a clear additive effect in terms of As(III) sensitivity. Hence, although the Met4p-mediated defense is important for As(III) tolerance, it is not as critically required as it is for cadmium or antimonite tolerance, possibly because of the lack of specific inducible cadmium or antimonite detoxification systems.

Regulation of the sulfur/GSH pathways under As(III) exposure. By combining transcriptome analyses with comparative genomics, we confirmed previous reports implicating Yap1p in the transcriptional response to As(III) challenge (17, 24, 40). Furthermore, we established a role for Met4p in transcriptional activation of genes in the sulfur/GSH pathways and demonstrated that this protein contributes to As(III) tolerance. Previous studies indicated that Met4p and Yap1p jointly control *GSH1* expression; induced *GSH1* expression is regulated by Yap1p in response to oxidative stress, whereas *GSH1* expression is co-regulated by Yap1p and Met4p during cadmium exposure (5) and GSH depletion (35). Here, we show that this joint control in fact extends to most genes of the sulfur/GSH pathways; deletion of either *YAP1* or *MET4* resulted in reduced gene expression, and promoter analysis confirmed the presence of Yap1p and Met4p DNA binding sites in these promoters. Hence, both transcription factors contribute to As(III)-stimulated expression of sulfur/GSH pathway genes.

During sulfur starvation, transcription of sulfur/GSH pathway genes is regulated in response to changes in the cysteine pool (15, 23). In As(III)-exposed cells, we found an increase in the cysteine pool without a downregulation of Met4p gene targets. Similarly, Met4p activation is also independent of the cysteine pool under cadmium exposure (1, 19, 42). Met4p is regulated by ubiquitination, and the SCF^{Met30} (Skp1/Cullin/F-box protein, where Met30p is the F-box protein) ubiquitin ligase complex is responsible for ubiquitination (and hence inactivation) of Met4p in response to adequate cysteine levels (18, 29). Cadmium inhibits Met4p ubiquitination by preventing proper formation of the SCF^{Met30} ubiquitin ligase complex (1, 42). Interestingly, the need for Yap1p to fully induce expression of sulfur/GSH genes at low concentrations of As(III), but not at high concentrations, may also be explained by the regulation of Met4p. Ubiquitination and degradation of Met4p are inhibited in response to As(III); however, high concentrations are needed to completely abolish Met4p ubiquitination (42). So when cells are exposed to low arsenite concentration, Met4p may still be partially ubiquitinated, and full induction of sulfur/GSH genes then requires Yap1p as an additional transactivating factor.

Other factors required for As(III) tolerance? Besides Yap1p and Met4p-recruiting factors, comparative genome analysis pinpointed a number of transcription factors that might control the transcriptional response to As(III). These factors include Msn2p, Msn4p, and Rap1p, which have been implicated in the regulation of the so-called "environmental stress response" (11), as well as Rpn4p, which controls As(III)-stimulated

expression of proteasome genes (17). However, the physiological roles of these transcription factors and their molecular mechanisms of action under As(III) challenge remain to be revealed. Our promoter analysis did not identify Yap8p, despite the fact that this transcription factor is critically involved in mediating arsenic tolerance by activating *ACR2* and *ACR3* expression (40). The exact DNA binding site of Yap8p is not known, and, hence, Yap8p was not retrieved by this analysis.

A common response to thiol-reactive metals? Yeast cells respond in a similar way to both As(III) and cadmium; expression of genes and enzymes of the sulfur/GSH pathways is strongly induced, GSH synthesis and pathway flux increase, and several pathway mutants display arsenite and cadmium sensitivity (Refs. 8, 19, 25, and 36 and present study). Since both arsenite and cadmium are thiol-reactive metals, the observed stimulation of the sulfur/GSH pathways might be of a general nature in response to this class of metals. Consistent with this notion, *met4*Δ is antimonite sensitive, and Yap1p is involved in the cellular response to antimonite (40). Although yeast cells respond to cadmium and arsenite in a similar way, metal-specific responses also exist. For instance, *ACR2* and *ACR3* respond only to As(III), while a similar cadmium-induced detoxification system has not been described. In response to cadmium, yeast launch a so-called sulfur sparing program; highly abundant proteins involved in carbohydrate metabolism (e.g., pyruvate decarboxylase and enolase) are replaced by isoenzymes with a low sulfur amino acid content, possibly to permit allocation of more sulfur to GSH production (8). Here, we did not find any clear evidence for a sulfur sparing program in response to arsenite. Whether this response is absent altogether or masked by the action of Acr3p remains to be revealed.

To conclude, this study confirms and extends previous reports on the *S. cerevisiae* response to arsenite. First, our transcriptional analysis largely corroborates previously reported gene expression data (17). Haugen et al. (17) integrated phenotypic and transcriptional profiling and mapped the data onto metabolic and regulatory networks. By using this approach, they suggested that arsenic-exposed cells channel sulfur into GSH biosynthesis (17). We now demonstrate that this is indeed the case by sulfur metabolite and metabolic flux analyses. Second, our expression data combined with the promoter analysis clearly establish that Met4p and Yap1p act together as transcriptional activators of the sulfur assimilation/GSH biosynthesis pathways in As(III)-challenged cells. Finally, we show that quantitative transcriptome, proteome, and metabolite profiling combined with comparative genomics and physiology provides a powerful means to obtain "systems level" insight into the role and regulation of entire metabolic pathways. Arsenite has a profound impact on the environment and on human health, as both a causative and a curative agent of disease, and a full understanding of global and specific responses may prove of value for use in medical therapy.

ACKNOWLEDGMENTS

We acknowledge L. Kuras (Centre National de la Recherche Scientifique, Gif-sur-Yvette, France) for providing strains, P. Dahl (Gothenburg University, Gothenburg, Sweden) for constructing the *acr3*Δ *met4*Δ mutant, and R. Genet (CEA/Saclay) for providing [^{15}N]ammonium sulfate.

GRANTS

This work was supported by grants from the Programme Toxicologie Nucléaire Environnementale, to J. Labarre; the Swedish National Research School in Genomics and Bioinformatics, to O. Nerman and M. J. Tamás; and the Swedish Research Council, to M. J. Tamás.

REFERENCES

1. **Barbey R, Baudouin-Cornu P, Lee TA, Rouillon A, Zarzov P, Tyers M, Thomas D.** Inducible dissociation of SCF^{Met30} ubiquitin ligase mediates a rapid transcriptional response to cadmium. *EMBO J* 24: 521–532, 2005.
2. **Boer VM, de Winde JH, Pronk JT, Piper MD.** The genome-wide transcriptional responses of *Saccharomyces cerevisiae* grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus, or sulfur. *J Biol Chem* 278: 3265–3274, 2003.
3. **Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M.** Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301: 71–76, 2003.
4. **de Winde JH, Crauwels M, Hohmann S, Thevelein JM, Winderickx J.** Differential requirement of the yeast sugar kinases for sugar sensing in the establishment of the catabolite repressed state. *Eur J Biochem* 241: 633–643, 1996.
5. **Dormer UH, Westwater J, McLaren NF, Kent NA, Mellor J, Jamieson DJ.** Cadmium-inducible expression of the yeast *GSH1* gene requires a functional sulfur-amino acid regulatory network. *J Biol Chem* 275: 32611–32616, 2000.
6. **Edgar R, Domrachev M, Lash AE.** Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30: 207–210, 2002.
7. **Evens AM, Tallman MS, Gartenhaus RB.** The potential of arsenic trioxide in the treatment of malignant disease: past, present, and future. *Leuk Res* 28: 891–900, 2004.
8. **Fauchon M, Lagniel G, Aude JC, Lombardia L, Soularue P, Petat C, Marguerie G, Sentenac A, Werner M, Labarre J.** Sulfur sparing in the yeast proteome in response to sulfur demand. *Mol Cell* 9: 713–723, 2002.
9. **Fernandes L, Rodrigues-Pousada C, Struhl K.** Yap, a novel family of eight bZIP proteins in *Saccharomyces cerevisiae* with distinct biological functions. *Mol Cell Biol* 17: 6982–6993, 1997.
10. **Frisbie SH, Ortega R, Maynard DM, Sarkar B.** The concentrations of arsenic and other toxic elements in Bangladesh's drinking water. *Environ Health Perspect* 110: 1147–1153, 2002.
11. **Gasch AP.** The environmental stress response: a common yeast response to diverse environmental stresses. In: *Yeast Stress Responses*, edited by Hohmann S and Mager WH. Heidelberg: Springer Verlag, 2003, p. 11–70.
12. **Ghosh M, Shen J, Rosen BP.** Pathways of As(III) detoxification in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 96: 5001–5006, 1999.
13. **Grant CM.** Role of the glutathione/glutaredoxin and thioredoxin systems in yeast growth and response to stress conditions. *Mol Microbiol* 39: 533–541, 2001.
14. **Güldener U, Heck S, Fielder T, Beinhauer J, Hegemann JH.** A new efficient gene disruption cassette for repeated use in budding yeast. *Nucleic Acids Res* 24: 2519–2524, 1996.
15. **Hansen J, Johannsen PF.** Cysteine is essential for transcriptional regulation of the sulfur assimilation genes in *Saccharomyces cerevisiae*. *Mol Gen Genet* 263: 535–542, 2000.
16. **Hastie TJ, Tibshirani RJ.** *Generalized Additive Models.* London: Chapman and Hall, 1990.
17. **Haugen AC, Kelley R, Collins JB, Tucker CJ, Deng C, Afshari CA, Brown JM, Ideker T, Van Houten B.** Integrating phenotypic and expression profiles to map arsenic-response networks. *Genome Biol* 5: R95, 2004.
18. **Kaiser P, Flick K, Wittenberg C, Reed SI.** Regulation of transcription by ubiquitination without proteolysis: Cdc34/SCF^{Met30}-mediated inactivation of the transcription factor Met4. *Cell* 102: 303–314, 2000.
19. **Lafaye A, Junot C, Pereira Y, Lagniel G, Tabet JC, Ezan E, Labarre J.** Combined proteome and metabolite-profiling analyses reveal surprising insights into yeast sulfur metabolism. *J Biol Chem* 280: 24723–24730, 2005.
20. **Lafaye A, Labarre J, Tabet JC, Ezan E, Junot C.** Liquid chromatography-mass spectrometry and ^{15}N metabolic labeling for quantitative metabolic profiling. *Anal Chem* 77: 2026–2033, 2005.

21. **Lee J, Godon C, Lagniel G, Spector D, Garin J, Labarre J, Toledano MB.** Yap1 and Skn7 control two specialized oxidative stress response regulons in yeast. *J Biol Chem* 274: 16040–16046, 1999.
22. **Liu SX, Athar M, Lippai I, Waldren C, Hei TK.** Induction of oxyradicals by arsenic: implication for mechanism of genotoxicity. *Proc Natl Acad Sci USA* 98: 1643–1648, 2001.
23. **Menant A, Baudouin-Cornu P, Peyraud C, Tyers M, Thomas D.** Determinants of the ubiquitin-mediated degradation of the Met4 transcription factor. *J Biol Chem* 281: 11744–11754, 2006.
24. **Menezes RA, Amaral C, Delaunay A, Toledano M, Rodrigues-Pousada C.** Yap8p activation in *Saccharomyces cerevisiae* under arsenic conditions. *FEBS Lett* 566: 141–146, 2004.
25. **Momose Y, Iwahashi H.** Bioassay of cadmium using a DNA microarray: genome-wide expression patterns of *Saccharomyces cerevisiae* response to cadmium. *Environ Toxicol Chem* 20: 2353–2360, 2001.
26. **Nordstrom DK.** Public health. Worldwide occurrences of arsenic in ground water. *Science* 296: 2143–2145, 2002.
27. **Pompella A, Visvikis A, Paolicchi A, De Tata V, Casini AF.** The changing faces of glutathione, a cellular protagonist. *Biochem Pharmacol* 66: 1499–1503, 2003.
28. **Ravandi F.** Arsenic trioxide: expanding roles for an ancient drug? *Leukemia* 18: 1457–1459, 2004.
29. **Rouillon A, Barbey R, Patton EE, Tyers M, Thomas D.** Feedback-regulated degradation of the transcriptional activator Met4 is triggered by the SCF^{Met30} complex. *EMBO J* 19: 282–294, 2000.
30. **Smyth GK.** Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3: Article3, 2004.
31. **Tamás MJ, Labarre J, Toledano MB, Wysocki R.** Mechanisms of toxic metal tolerance in yeast. In: *Molecular Biology of Metal Homeostasis and Detoxification: from Microbes to Man*, edited by Tamás MJ and Martinoia E. Heidelberg: Springer Verlag, 2005, p. 395–454.
32. **Thomas BJ, Rothstein R.** Elevated recombination rates in transcriptionally active DNA. *Cell* 56: 619–630, 1989.
33. **Thomas D, Surdin-Kerjan Y.** Metabolism of sulfur amino acids in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev* 61: 503–532, 1997.
34. **Thorsen M, Di Y, Tangemo C, Morillas M, Ahmadpour D, Van der Does C, Wagner A, Johansson E, Boman J, Posas F, Wysocki R, Tamás MJ.** The MAPK Hog1p modulates Fps1p-dependent arsenite uptake and tolerance in yeast. *Mol Biol Cell* 17: 4400–4410, 2006.
35. **Wheeler GL, Trotter EW, Dawes IW, Grant CM.** Coupling of the transcriptional regulation of glutathione biosynthesis to the availability of glutathione and methionine via the Met4 and Yap1 transcription factors. *J Biol Chem* 278: 49920–49928, 2003.
36. **Vido K, Spector D, Lagniel G, Lopez S, Toledano MB, Labarre J.** A proteome analysis of the cadmium response in *Saccharomyces cerevisiae*. *J Biol Chem* 276: 8469–8474, 2001.
37. **Wood SN.** *Generalized Additive Models: an Introduction with R.* Boca Raton, FL: Chapman and Hall/CRC, 2006.
38. **Wysocki R, Bobrowicz P, Ulaszewski S.** The *Saccharomyces cerevisiae ACR3* gene encodes a putative membrane protein involved in arsenite transport. *J Biol Chem* 272: 30061–30066, 1997.
39. **Wysocki R, Chéry CC, Wawrzycka D, Van Hulle M, Cornelis R, Thevelein JM, Tamás MJ.** The glycerol channel Fps1p mediates the uptake of arsenite and antimonite in *Saccharomyces cerevisiae*. *Mol Microbiol* 40: 1391–1401, 2001.
40. **Wysocki R, Fortier PK, Maciaszczyk E, Thorsen M, Leduc A, Odhagen A, Owsianik G, Ulaszewski S, Ramotar D, Tamás MJ.** Transcriptional activation of metalloid tolerance genes in *Saccharomyces cerevisiae* requires the AP-1-like proteins Yap1p and Yap8p. *Mol Biol Cell* 15: 2049–2060, 2004.
41. **Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP.** Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 30: e15, 2002.
42. **Yen JL, Su NY, Kaiser P.** The yeast ubiquitin ligase SCF^{Met30} regulates heavy metal response. *Mol Biol Cell* 16: 1872–1882, 2005.

Paper 6

BMC Genomics

BioMed Central

Research article

Sensitive and robust gene expression changes in fish exposed to estrogen – a microarray approach

Lina Gunnarsson[1], Erik Kristiansson[2], Lars Förlin[3], Olle Nerman[2] and D G Joakim Larsson*[1]

Address: [1]Department of Neuroscience and Physiology, the Sahlgrenska Academy at Göteborg University, SE-405 30 Göteborg, Sweden, [2]Department of Mathematical Statistics, Chalmers University of Technology, SE-412 96 Göteborg, Sweden and [3]Department of Zoology/Zoophysiology, Göteborg University, SE-405 30 Göteborg, Sweden

Email: Lina Gunnarsson - lina.gunnarsson@fysiologi.gu.se; Erik Kristiansson - erikkr@math.chalmers.se; Lars Förlin - lars.forlin@zool.gu.se; Olle Nerman - nerman@chalmers.se; D G Joakim Larsson* - joakim.larsson@fysiologi.gu.se

* Corresponding author

Published: 7 June 2007

BMC Genomics 2007, **8**:149 doi:10.1186/1471-2164-8-149

Received: 22 December 2006
Accepted: 7 June 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/149

Abstract

Background: Vitellogenin is a well established biomarker for estrogenic exposure in fish. However, effects on gonadal differentiation at concentrations of estrogen not sufficient to give rise to a measurable vitellogenin response suggest that more sensitive biomarkers would be useful. Induction of zona pellucida genes may be more sensitive but their specificities are not as clear. The objective of this study was to find additional sensitive and robust candidate biomarkers of estrogenic exposure.

Results: Hepatic mRNA expression profiles were characterized in juvenile rainbow trout exposed to a measured concentration of 0.87 and 10 ng ethinylestradiol/L using a salmonid cDNA microarray. The higher concentration was used to guide the subsequent identification of generally more subtle responses at the low concentration not sufficient to induce vitellogenin. A meta-analysis was performed with data from the present study and three similar microarray studies using different fish species and platforms. Within the generated list of presumably robust responses, several well-known estrogen-regulated genes were identified. Two genes, confirmed by quantitative RT-PCR (qPCR), fulfilled both the criteria of high sensitivity and robustness; the induction of the genes encoding zona pellucida protein 3 and a nucleoside diphosphate kinase (nm23).

Conclusion: The cross-species, cross-platform meta-analysis correctly identified several robust responses. This adds confidence to our approach used for identifying candidate biomarkers. Specifically, we propose that analyses of an nm23 gene together with zona pellucida genes may increase the possibilities to detect an exposure to low levels of estrogenic compounds in fish.

Background

The contraceptive estrogen, ethinylestradiol (EE_2) is an important contributor to the feminization of fish downstream from sewage treatment works [1-5]. This discovery was greatly facilitated by the use of vitellogenin (VTG) as a biomarker. VTG is produced in the liver of sexually maturing female fish under the influence of endogenous estrogen. Normally, VTG is not expressed in males or juve-

niles, unless they are exposed to estrogens via water or food. Both VTG mRNA and protein in male and juvenile fish have thus become established biomarkers for exposure to environmental estrogens [6]. However, estrogens can effect gonadal sex differentiation of fish at concentration not sufficient to give rise to a measurable VTG response [7]. It has also been shown that life cycle exposure of fathead minnow to an inordinately low concentration of EE_2 (0.32 ng/L) was sufficient to decrease the egg fertilisation and to skew the sex ratios towards female[8]. This suggests that more sensitive biomarkers would be useful. Zona pellucida (ZP) genes may be more sensitive than VTG [9] but their specificity for estrogens is not as clear [10-12]. Additional, sensitive biomarkers would thus increase our possibilities to identify exposure to low, but biologically important concentrations of estrogens.

Rapidly accumulating data on genomes and proteomes have increased the possibilities to use different types of discovery-driven methods in ecotoxicology [13,14]. The large number of potential responses that can be studied with microarrays renders the method suitable for identifying candidate biomarkers of exposure [15-20]. Such candidates may then be further evaluated to find if they are useful as biomarkers. In general, a good biomarker should be sensitive, specific and robust. A robust response implies for example that it should be measurable at complex exposure situations, at different exposure concentrations, at different temperatures, after different exposure times, by different analytical approaches, in different labs and preferably also in different species.

The main objective of the present study was to use microarrays to find novel, sensitive and robust biomarkers of estrogenic exposure in fish. We have used a salmonid cDNA microarray from cGRASP [21] to analyze hepatic expression profiles in juvenile rainbow trout (*Oncorhynchus mykiss)* exposed to EE_2 *in vivo*. The responses identified at a high concentration of EE_2 were used to guide the subsequent identification of generally more subtle responses at a low concentration of estrogen. We also identified estrogen-responses shared between fish species, experimental conditions and analytical platforms. This was achieved by a meta-analysis using our dataset together with results from three recently published articles describing hepatic gene expression profiles in fish exposed to estrogens [16,20,22].

Results

Sensitive gene-expression changes

Both male and female juvenile fish exposed to 0.87 ng EE_2/L were analyzed with microarray. The microarray analysis of female fish suggested that only three out of four females had an induced expression of the known estrogen-responsive gene ZP3. In contrast, an induction was present in all eight males. This observation suggested that some juvenile females may have sufficient endogenous estrogen to induce sensitive estrogen-responsive genes. Thus, in our search for genes responding to low concentrations of estrogens only the microarray results from male fish were used.

Thirty-six sets of cDNAs (presumably corresponding to 29 genes) were regulated in male fish both by the low and the high concentration of EE_2 (Table 1). All of the cDNAs responded in a dose-dependent manner. ZP3 was the most differentially expressed gene in fish exposed to both high and low concentrations with a fold change of 84 and 3.5 respectively. VTG was not affected by the low concentration while it was up-regulated 537 times by 10 ng EE_2/L as measured by quantitative RT-PCR (qPCR) (Figure 1).

Robust gene-expression changes

A meta-analysis was performed with the aim to identify robust estrogen-responsive genes. The microarray data from fish (both sexes) exposed to 10 ng/L from the present study and available microarray data from three other exposure studies with fish and estradiol (E_2) or EE_2 were used in the meta-analysis [16,20,22]. Information about the different studies is shown in Table 2. Transcripts (360) presumably corresponding to 55 genes or groups of paralog genes were identified as differentially expressed in at least two of the four different studies (see Additional file 1). VTG and ZP3 were differentially expressed in all four studies and nine genes had an altered expression in at least three studies (Figure 2). It should be noted that ZP1 and the estrogen receptor-α, which are well-know estrogen-responsive genes in fish, have poor sequence representation on the cGRASP microarrays and are therefore not present in Figure 2.

Confirmation of microarray data with quantitative RT-PCR

Genes that were likely to be both sensitive (Table 1) and robust (Figure 2) were chosen for subsequent qPCR analysis. Three genes fulfilled these criteria: ZP3, a nucleoside diphosphate kinase (nm23) and fatty acid binding protein 3 (fabp3 or H-FABP). In addition, VTG was subjected to the qPCR analysis as well as the reference gene ubiquitin. In accordance with the microarray results the expression of VTG, ZP3 and nm23 were significantly induced in fish exposed to the high concentration. Also, as suggested by the microarrays, ZP3 and nm23 were significantly induced by the low concentration as well, whereas VTG expression was not induced (Figure 1). In stark contrast to the microarray results fabp3 had no tendency to any regulation caused by the treatment but showed a large variation within each treatment group (data not shown). The fabp3 and nm23 qPCR products were sequenced in order to confirm the amplification of the right products and

Table 1: Estrogen-sensitive genes.

cGRASP ID	M-value 0.87 ng/L	M-value 10 ng/L	Annotation
CK991165	**1.40**	**5.60**	**[GO] [P10761] Zona pellucida sperm-binding protein 3 precursor (Zona pellucida glycoprotein ZP3) (Sperm receptor) (Zona pellucida protein C)**,
CB492227	-0.23	-0.34	[GO] [P23506] Protein-L-isoaspartate(D-aspartate) O-methyltransferase (EC 2,1,1,77) (Protein-beta-aspartate methyltransferase) (PIMT) (Protein L- isoaspartyl/D-aspartyl methyltransferase)
CA054422	0.22	0.43	UNKNOWN
CB496664	0.30	0.50	[GO] [Q9D0E1] Heterogeneous nuclear ribonucleoprotein M (hnRNP M),
CB497378	**0.27**	**0.69**	**[GO] [P15532] Nucleoside diphosphate kinase A (EC 2,7,4,6) (NDK A) (NDP kinase A) (Tumor metastatic process-associated protein) (Metastasis inhibition factor NM23) (NDPK-A) (nm23-M1)**,
CB511030	**0.27**	**0.47**	**[GO] [P15532] Nucleoside diphosphate kinase A (EC 2,7,4,6) (NDK A) (NDP kinase A) (Tumor metastatic process-associated protein) (Metastasis inhibition factor NM23) (NDPK-A) (nm23-M1)**
CK991305	**0.27**	**0.59**	**[GO] [Q01768] Nucleoside diphosphate kinase B (EC 2,7,4,6) (NDK B) (NDP kinase B) (nm23-M2) (P18)**,
CA037915	0.30	0.38	[GO] [P35505] Fumarylacetoacetase (EC 3,7,1,2) (Fumarylacetoacetate hydrolase) (Beta-diketonase) (FAA),
CA060608	0.21	0.46	[GO] [P56384] ATP synthase lipid-binding protein, mitochondrial precursor (EC 3,6,3,14) (ATP synthase proteolipid P3) (ATPase protein 9) (ATPase subunit C),
CB496562	0.16	0.32	[GO] [Q9CY58] Plasminogen activator inhibitor 1 RNA-binding protein (PAI1 RNA-binding protein 1) (PAI-RBP1),
CB516182	0.46	0.56	[GO] [O08709] Peroxiredoxin 6 (EC 1,11,1,15) (Antioxidant protein 2) (1-Cys peroxiredoxin) (1-Cys PRX) (Acidic calcium-independent phospholipase A2) (EC 3,1,1,-) (aiPLA2) (Non-selenium glutathione peroxidase) (EC 1,11,1,7) (NSGPx),
CB511422	**0.25**	**0.50**	**[GO] [P15532] Nucleoside diphosphate kinase A (EC 2,7,4,6) (NDK A) (NDP kinase A) (Tumor metastatic process-associated protein) (Metastasis inhibition factor NM23) (NDPK-A) (nm23-M1)**,
CB496931	**0.80**	**2.58**	**[GO] [P11404] Fatty acid-binding protein, heart (H-FABP) (Heart-type fatty acid- binding protein) (Mammary-derived growth inhibitor) (MDGI)**,
CB497374	**0.60**	**2.08**	**[GO] [P11404] Fatty acid-binding protein, heart (H-FABP) (Heart-type fatty acid- binding protein) (Mammary-derived growth inhibitor) (MDGI)**,
CB505692	-0.33	-0.42	UNKNOWN
CB497174	0.53	1.71	[NR] [XP_423045] PREDICTED: similar to nudix (nucleoside diphosphate linked moiety X)-type motif 7; coenzyme A diphosphatase [Gallus gallus]
CB497649	**0.21**	**0.52**	**[GO] [Q01768] Nucleoside diphosphate kinase B (EC 2,7,4,6) (NDK B) (NDP kinase B) (nm23-M2) (P18)**,
CA037988	0.23	0.41	[NT] [AJ488155] Pachymedusa dacnicolor partial mRNA for ribosomal protein S16 (rps16 gene)
CB499596	0.14	0.49	[NR] [NP_077217] hydroxysteroid dehydrogenase like 2 [Mus musculus]
CB489314	-1.01	-1.11	UNKNOWN
CA769854	**0.64**	**2.17**	**[GO] [P11404] Fatty acid-binding protein, heart (H-FABP) (Heart-type fatty acid- binding protein) (Mammary-derived growth inhibitor) (MDGI)**,
CB509453	-0.32	-0.48	[GO] [O16797] 60S ribosomal protein L3,
CB500821	-0.26	-0.36	[GO] [P62918] 60S ribosomal protein L8,
CB492885	-0.15	-0.31	UNKNOWN
CA061403	0.42	0.72	UNKNOWN
CB498219	0.24	0.51	[NR] [XP_613218] PREDICTED: similar to 24-dehydrocholesterol reductase precursor, partial [Bos taurus]
CA054168	-0.39	-0.58	UNKNOWN
CB515449	0.34	-0.60	[GO] [P50247] Adenosylhomocysteinase (EC 3,3,1,1) (S-adenosyl-L-homocysteine hydrolase) (AdoHcyase) (Liver copper binding protein) (CUBP),
CB515945	-0.36	-0.49	[NR] [NP_683732] RNA binding motif protein 5 [Mus musculus]
CA057448	-0.30	-0.33	UNKNOWN
CA036745	-0.62	-0.74	[NT] [XM_532501] PREDICTED: Canis familiaris similar to Chimerin (chimaerin) 2 (LOC475267), mRNA
CB509472	-0.29	-0.42	UNKNOWN
CB494192	0.23	0.33	[GO] [P09411] Phosphoglycerate kinase 1 (EC 2,7,2,3),
CB496589	-0.49	-0.82	UNKNOWN
CB513882	-0.48	-0.60	[NR] [XP_413822] PREDICTED: similar to normal mucosa of esophagus specific 1 [Gallus gallus]
CK990857	-0.64	-1.11	UNKNOWN

cDNAs or sets of cDNA putatively sensitive to estrogen exposure as judged by the presence on the top 250-lists ranked by moderated t-statistics on both 0.87 and 10 ng EE2/L exposure experiments in male juvenile rainbow trout. cDNAs corresponding to genes that were selected for qPCR analysis are marked with bold text. Note that several cDNAs may likely correspond to the same gene.

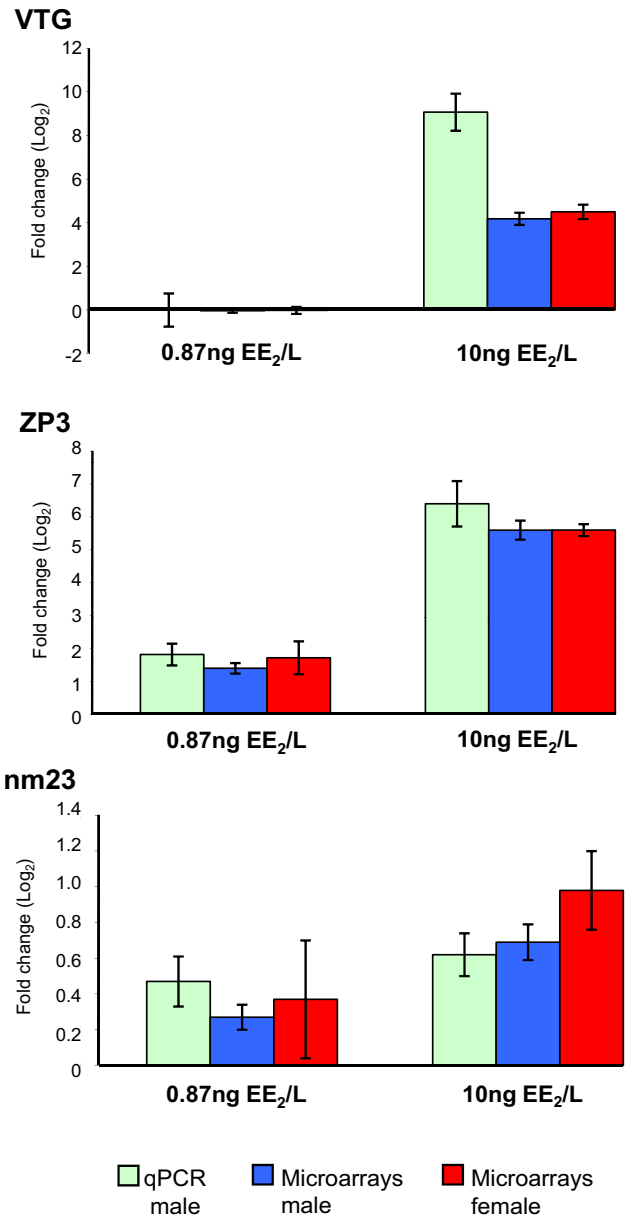


Figure 1
Gene expression changes of VTG, ZP3 and nm23 measured by qPCR and microarray. Hepatic gene expression in rainbow trout of vitellogenin (VTG), zona pellucida protein 3 (ZP3) and a nucleoside diphosphate kinase (nm23) after EE_2 exposure measured with qPCR (green bars, male fish) or microarray (blue bars, male fish: red bars, female fish). Values are expressed as fold change (log_2) compared to control fish. Paired student's t-tests (single sided) were performed on the qPCR data to confirm/test the putative regulation suggested from microarray data. VTG, ZP3 and nm23 were confirmed to be significantly up-regulated in fish exposed to 10 ng/L (p = 0.001, 0.001 and 0.007 respectively, four biological replicates in each group). ZP3 and nm23, but not VTG were up-regulated in fish exposed to 0.87 ng/L (p = 0.0004, 0.006 and 0.5 respectively, eight biological replicates in each group) in accordance with the microarray data.

they were identical to fabp3 and nm23 [EMBL:U95296, AF350241] in rainbow trout (data not shown).

Discussion

Our meta-analysis correctly identified some of the most well known estrogen-responsive genes (VTG, ZP3, ZP2). This suggests that the approach has a good potential to identify other robust, less well known estrogen-regulated genes. We also showed that ZP3 and a hepatic nucleoside diphosphate kinase nm23 are more sensitive to estrogenic exposure than the widely used biomarker VTG. As far as we know, no other microarray study has identified the effects of as low concentrations of estrogen as used here. The recognition of nm23 induction as a highly sensitive response is therefore a novel finding. Thus we propose that analyses of nm23 together with ZP genes may increase the possibilities to detect an exposure to low levels of estrogenic compounds in fish. However, more studies are required in order to fully assess the potential of nm23 as a biomarker.

Sensitive biomarkers can be used as early warning signals to indicate exposure and thus potential risk of adverse effects. It has been suggested that the induction of ZP mRNAs are more sensitive than induction of VTG [9,23]. However expression of ZP genes can, as most genes, be affected by other environmental factors, for example cortisol exposure [10-12]. The regulation of a single gene is rarely sufficient to conclusively demonstrate a specific exposure, but a combination of responses would together potentially increase the degree of evidence.

We identified nine genes (or groups of paralog genes) that were affected by estrogen in at least three out of the four studies included in the meta-analysis. The known estrogen-responsive genes VTG and ZP3 were up-regulated in all four studies. The robust gene expression changes of ZP3, nm23 and fabp3 were also tentatively identified to be sensitive. However, the induction of fabp3 was not confirmed by qPCR. The incorrect identification from microarray data might be explained by cross hybridization to related mRNAs, a known problem for cDNA microarrays. Nm23, on the other hand, was confirmed with qPCR to be significantly induced both by a low and a high water concentration of EE_2. In addition, microarray results from the other studies of rainbow trout exposed to 50 ng EE_2/L and dietary exposed to 5 μg/g of E_2 further supports an estrogen-induction of nm23 in rainbow trout during different exposure conditions [20,22]. The study of estrogen-exposed medaka did not report nm23 as an estrogen-responsive gene but it is unclear if nm23 was represented on the medaka microarray [16]. Whether nm23 is regulated by estrogen in other fish species is still an open question, although mammalian studies suggest a conserved induction mechanism [24].

Table 2: A summary of the four different studies used in the meta-analysis

	Present study	**Hook *et al*. 2006**	**Tilton *et al*. 2006**	**Kishi *et al*. 2006**
Species	*O.mykiss*	*O.mykiss*	*O.mykiss*	*O.latipes*
Sex	juvenile	male	juvenile	male
Estrogenic substance	EE_2	EE_2	E_2	E_2
Exposure	Water, 10 ng/L	Water, 50 ng/L	Dietary, 5 ppm	Water, 100 ng/L
Water temperature (C°)	10	12	12	24
Duration (days)	14	7	12	21
Platform	Two-channel spotted cDNA (GRASP 16 k v.1)	Two-channel spotted cDNA (GRASP 16 k v.1)	Two-channel spotted cDNA (GRASP 3.7 k v.1)	One-channel oligonucleotide (60 mer) (Kishi *et al*. 2006)
Experimental Setup	Direct comparison, 8 biological replicates	Direct comparison, 3 biological × 3 technical replicates	Reference design, 2 biological × 2 technical replicates	6 control and 3 exposed biological replicates
Number of cDNAs/ probes	16006	16006	3700	22587
Pre-processing	Loess, no background correction	Loess	Loess	Robust Multichip
Statistical method used for ranking	Moderated t-statistic	Student t-statistic	fold change	Student t-statistic
Selected cDNA/ probes	250 (167 induced, 83 suppressed)	189 (48 induced, 141 suppressed)	366 (127 induced, 239 suppressed)	381 (242 induced, 139 suppressed)
Source for sequences	cDNA, cGRASP	cDNA, cGRASP	cDNA, cGRASP	Transcripts, TIGR (OLGI release 4.0)
Number of matches[a] in *D.rerio*	91	89	190	184

a) A match is defined as a tblastx hit with a E-value less than 10^{-25}.

Nm23 belongs to a larger class of nucleoside disphophate kinases that exist in multiple isoforms and are highly conserved throughout evolution. The investigated nm23 have been sequenced in rainbow trout [EMBL AF350241], Atlantic salmon [EMBL AF045187] and zebrafish [EMBL AF201764]. The salmon and zebrafish nm23 shows high similarity to the human nm23-H1 and H2 genes. A phylogenetic analysis suggests that nm23-H1 and H2 have arisen by gene duplication after the speciation event that gave rise to modern teleost fish and tetrapods [25]. Therefore it is assumed that the salmonid genome would only have one gene homologue to the nm23 -H1 and -H2 genes. In mammals the nm23-H2 gene encodes the c-MYC transcription factor and the nm23-H1 gene has been shown to be metastasis associated [24]. Moreover, the nm23-H1 gene and protein is up-regulated by E_2-treatment in human breast carcinoma cell lines. This induction seems to be mediated, at least in part, at the transcriptional level via the estrogen receptor α binding to an estrogen responsive element in the promoter region of the human nm23-H1 [24]. The physiological function of nm23 in fish remains to be determined as well as the possibilities of a regulation by other factors than estrogen (specificity) and the robustness of the response during more complex exposure scenarios.

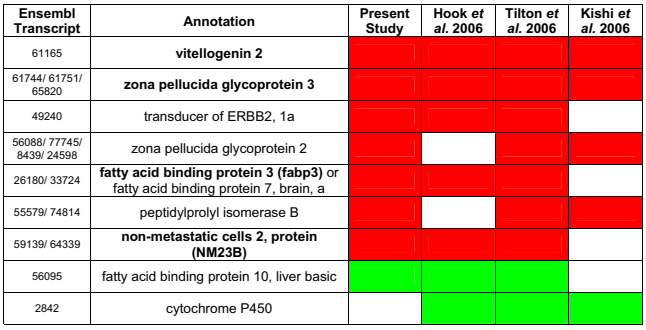


Figure 2
Robust estrogen-responsive genes. Genes affected by estrogen in at least three out of four studies used in the meta-analysis. Red refers to an up-regulation, whereas green refers to a down-regulation. Only the zebrafish transcripts with the best TBLASTX hit to each of the probes from the different studies are presented in the figure. Genes that were selected for qPCR analysis are marked with bold text.

To be useful as a biomarker, a response should ideally be as robust as possible. In the meta-analysis we tested gene responses for robustness between species, exposure conditions and analytical platforms. Combining microarray data from different species and platforms is a challenging task, particularly when sequence information and good annotations are limited. We have addressed the cross platform/cross species comparison by using the zebrafish transcriptome as a reference. In contrast to most other fish

species zebrafish has the advantage of being both well sequenced and well annotated. However, using zebrafish as a reference also has limitations, e.g. a lack of identified homologes for some genes. The results in the meta-analysis were also influenced by the shortage of available microarray raw data and therefore we had to accept the different statistical approaches used for selecting estrogen-responsive genes in the different studies.

It is certainly possible that more than nine hepatic genes are robustly regulated by estrogen in the analyzed species/conditions. We have only included comprehensive fish arrays in the meta-analysis. Nevertheless, many genes are still likely to be represented on only one or a few of the array platforms which limits the possibility of identifying robust responses. The choice of microarray platform also affects the possibilities to accurately identify differentially expressed genes. Amplicon arrays (cDNA arrays) show less concordance with other platforms, for example qPCR and commercially produced high density arrays with oligonucleotide probes or cDNA arrays with synthetic oligonucleotides [26]. Although, it has been shown that when two independent platforms give consistent results, the outcome of qPCR analysis will most often also be in agreement [27,28]. This adds confidence that several of the differentially expressed genes identified by the meta-analysis indeed are relatively robust responses.

By making the data on putative sensitive and/or robust gene responses public, it can be used as a base for further investigations on the effects of environmental estrogens in fish in order to develop biomarkers or to increase the understanding of the physiological impact of environmental estrogens.

Conclusion

We have used microarrays to identify a range of potentially sensitive and/or robust gene expression changes in fish exposed to estrogen. We have identified the induction of ZP3 and a hepatic nm23 mRNA as being both sensitive and presumably robust responses. After further evaluation, nm23 induction would therefore be a good candidate biomarker together with ZP genes to reveal exposure to low levels of estrogens not sufficient to induce VTG but still with potential to affect gonadal differentiation of fish.

Methods

Experimental animals, exposure and preparation of hepatic total RNA

Fish from a previously published study were analysed [29]. The experimental setup was in short: 15, 14 and 14 juvenile rainbow trout (weighing around 100 g) were divided into three aquaria and exposed for two weeks to measured concentrations of 0, 0.87 and 10 ng/L respectively of EE_2 in a flow-through system. Water samples were taken from the low and high EE_2 concentration aquaria, before the transfer of the fish, on day 8 and on day 13. One sample was collected from the control aquaria on day 8. Solid phase columns were used to extract and purify EE_2 from the water followed by derivatization (pentafluorobenzoylester) and further purification. EE_2-concentrations were determined using GC/MS. The limit of detection (signal-to-noise set to 5) was 0.01 ng/L. Samuelsson *et al* showed that the fish exposed to 10 ng/L EE_2 had significantly increased plasma levels of VTG, increased hepatosomatic index and the plasma metabolite profile were affected by the treatment. However, in the fish exposed to 0.87 ng/L neither induction of plasma VTG protein nor an altered metabolite pattern could be demonstrated using a specific VTG-ELISA and NMR respectively [29]. Gene responses in liver are widely used as biomarkers for environmental pollutants, i.e. estrogens, and the hepatic responses to estrogens are not restricted to a short developmental period. A prerequisite for our meta-analysis was availability of additional array-data from the same tissue in estrogen-exposed fish. Only hepatic microarray data was available in the literature, which therefore also contributed to our choice of tissue. Livers were collected and snap frozen in liquid nitrogen. Total hepatic RNA was isolated from individual trout liver using TRI reagent (Sigma chemicals Co, St Louis, MO, USA). RNA quality and quantity were assessed by agarose gel electrophoresis and spectrophotometric measurements (Nanodrop 1000, NanoDrop Technologies, USA and Spectra MAXplus, Molecular Devices, CA, USA).

Microarray chip, hybridisation and wash

Salmonid cDNA microarrays (GRASP16k v1) were purchased from cGRASP, Univerity of Victoria, BC, Canada [21]. Microarray fabrication and quality control have previously been described in von Schalburg *et al*. [30]. The array contains 13,421 Atlantic salmon (*Salmon salar*) cDNAs and 2,576 rainbow trout cDNAs that together with a few more expressed sequence tags (ESTs) from other salmonid fish results in 16 006 spotted cDNAs in total. It has previously been shown that the sequence similarity between the Atlantic salmon and rainbow trout is sufficiently high for cross species use of the array [31].

Several cDNAs on the array correspond to the same gene and to reduce redundancy, a sequence based clustering was made as follows. Each cDNA sequence was compared to all other sequences on the array using BLAST [32]. A stringent cut-off value of at least 98% sequence similarity over 250 base-pairs or more was used to define equality. Single-link clustering was then applied which resulted in 13853 sets of cDNAs.

Slide preparation have been described in detail in von Schalburg *et al*. [30]. Briefly, 8 μg of total RNA was reverse

transcribed and labelled using SuperScript Indirect cDNA labelling System kit (Invitrogen, Carlsbad, CA, USA) and fluorescent dyes Cy5 and Cy3 (GE Healthcare, Buckinghamshire, UK). cDNA from one control fish and one exposed fish were labelled and hybridised to the same array. Every other pair was dye swapped to compensate for cyanine flour effects. Eight male control fish were paired with male fish exposed to 0.87 ng EE2/L matching individual weights and lengths as closely as possible. Four of the same male control fish were also paired to fish exposed to 10 ng EE2/L. In the same way, four female control fish were paired both to females exposed to the low dose and the high dose. Hybridisation and wash were performed as described before by von Schalburg *et al*. [30] with the exception of the prehybridization that was preformed for 1.5 h in 5xSSC, 0.1% SDS, 0.2% BSA at 49°C. In total 20 microarrays were analysed.

Microarray analyses

Fluorescent images of hybridized arrays were acquired using an Agilent MicroArray Scanner (Agilent Technologies, Palo Alto, CA, USA). Intensity data were extracted from TIFF images using Imagene version 6.0 (BioDiscovery, CA, USA). The statistical analysis was performed using the R package [33] LIMMA [34], which is available at the Bioconductor repository [35,36]. For each cDNA on the chip, M-values ($\log_2$ fold change) and A-values (average $\log_2$ intensity) were calculated. Loess normalization was applied to each array to remove intensity dependent trends [37]. For each set of cDNAs (defined above), an M-value was calculated by taking the average of the M-values of all the cDNAs in the set. Next, each set was annotated based on the cDNA with highest A-value (i.e. the spot with best hybridization). Finally, the sets of cDNAs were ranked by moderated t-statistic [34] to reduce the proportion of false positives. Data from the complete microarray experiment is available according to the MIAME guidelines at Array Express [38].

Meta-analysis

Microarray data from four different studies on estrogen-exposed fish were subjected to a meta-analysis with the aim to identify robust estrogen responsive genes [16,20,22]. Another study with estrogen-exposed adult female zebrafish was excluded since the control fish presumable had high levels of endogenuous estrogen (neither VTG nor ZP3 was regulated in this study) [39]. To our knowledge, no other relevant microarray studies covering more than 3000 cDNAs/transcripts were publicly available (i.e. open access to transcript sequences) in October 2006 when we performed the meta-analyses. The studies included are summarized in Table 2.

The meta-analysis was done as follows. For each study, a list of the reported estrogen-regulated genes and the corresponding transcripts/cDNA-sequences was created. Note that the studies used different statistical methods to find the regulated genes (Table 2).

From the present study, the topmost 250 sets of cDNA from fish (both female and male) exposed to 10 ng EE2/L were chosen. To compare the platforms, zebrafish was used as a reference species. It was chosen since it is almost fully sequenced and well annotated compared to the other species involved. All transcripts/cDNAs were compared to the zebrafish transcriptome available through Ensembl release 40 (26,679 in total) using tblastx [32]. A cut-off E-value of 10^{-25} was used to define a match. This resulted in a list of 360 zebrafish transcripts that had a match to transcripts/cDNAs from at least two studies (the transcripts should be regulated in the same direction). The list of zebrafish transcripts contained both multiple transcripts from the same gene (different splice variants) and paralogs and therefore the list was clustered into groups of transcripts. A similarity indicator matrix was created by comparing all transcripts in the list to each other using tblastx. Pairs of transcripts with an E-value of 10^{-50} or less were defined to be equal. Otherwise the distance was set to zero. Single link clustering was then applied to create the groups of transcripts. Finally, all transcripts were annotated using Ensembl. The complete list of transcripts is available in Additional file 1.

Quantitative RT-PCR

To confirm regulation of four selected genes, the abundances of the mRNAs were analysed with qPCR. The qPCR was performed on isolated total RNA from the same fish used in the microarray analysis. Total RNA (0.5 μg) was reverse transcribed in duplicate with a mixture of random hexamers and oligo(dT) primers, using the iScript™ cDNA Synthesis Kit (Bio-Rad, Hercules, CA, USA). The cDNA synthesis was performed according to the manufacturer's instructions, except for a scale-down of the reaction volume to 10 μl. Pooled RNA samples were used as no reverse transcriptase controls to control for genomic contamination. It was discovered that three samples might have been contaminated with DNA. These samples were treated with DNase and new qPCR analyses were done. PCR primers for ZP3 [EMBL:AF231708], nm23 [EMBL:AF350241], fabp3 [EMBL:U95296], VTG [EMBL:X92804], β-actin [EMBL:AJ438158] and ubiquitin [EMBL:AB036060] were designed using Primer3 software [40]. Primer sequences were as follows: 5'-ccctgcgtatctttgtgga-3' and 5'-gtgggaacctgtcattttgg-3' for ZP3; 5'-ccttcttcccctggtctcgt-3' and 5'-gatgatgttcctgcccactt-3' for nm23 ; 5'-ctttccctgtttcccctcct-3' and 5'-tgctgtgtgcttcttgctactc-3' for VTG; 5'-ggggcagtatggcttgtatg-3' and 5'-ctggcaccctaatcacctct-3' for beta actin; 5'-cgatagacggtggtaagatgg-3' and 5'-aggtgtggcaaagggtagtg-3' for fabp3 ; 5'- atgtcaaggccaagatccag -3' and 5'-ataatgcctccacgaagacg -3' for ubiquitin. For

all genes except the reference gene, ubiquitin (for which the qPCR was performed according to a previously published protocol [41]), the qPCR reactions contained 1× Real Time PCR Buffer, 3 mM $MgCl_2$, 400 μM dNTP, 300 nM of each primer, 1 U TaKaRa Ex Taq™ R-PCR Version 2.1 (TaKaRa Bio Inc., Shiga, Japan), 0.25× SYBR Green I (Molecular Probes Eugene, OR, USA) and cDNA corresponding to 20 ng total RNA, in a final reaction volume of 20 μl. Real-time qPCR was performed on a Stratagene Mx3005p with 30 sec initial denaturation at 95°C, followed by 45 cycles of 95°C for 20s, 60°C for 20s and 72°C for 20s. A melting curve analysis was performed after each run to verify specific amplification. In addition the qPCR products were subjected to an agarose gel electrophoresis to confirm the expected size of the product. Both beta actin and ubiquitin were chosen as potential reference genes. Beta actin had a high variance and also a tendency to be regulated in the high dose group and therefore only ubiquitin was used. All signals were normalized against ubiquitin and ratios were calculated for exposed fish compared to control fish paired in the same manner as in the microarray analysis. Paired single-sided student's t-test were performed to test for significantly regulated genes. Since all samples could not be run at the same occasion, two standard samples were run at all occasions in order to enable a compensation for a possible run to run variation. Applying a run to run factor made little difference and the differently expressed genes VTG, ZP3 and nm23 were significant up-regulated both with and without applying the factor. The presented qPCR results are calculated without this factor.

Authors' contributions

LG participated in the planning of the study, the sampling of the fish, carried out the molecular biology, participated in the statistical analysis, in the meta-analysis and drafted the manuscript. EK performed the statistical analysis and the meta-analysis and participated in the drafting of the manuscript. LF participated in the planning of the study. ON supervised the statistical analysis. DGJL planned the study, participated in the sampling, supervised and drafted the manuscript. All authors read and approved the final manuscript.

Additional material

Additional File 1

Transcripts or groups transcripts identified as differentially expressed in at least two of the four different studies

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-8-149-S1.xls]

Acknowledgements

The authors wish to thank cGRASP (B. F. Koop and W. Davidson) for providing the cDNA microarrays and G. Cooper for technical assistance/advice. We also want to thank L. Samuelsson for performing the fish exposure experiment. In addition we would like to thank the Lundberg Laboratories for Cancer Research at the Sahlgrenska Academy at Göteborg University for letting us use their Agilent scanner and TATAA Biocenter, Göteborg, Sweden for performing the qPCR analysis. This research was supported by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS) (DGJL, LF), the Foundation for Strategic Environmental Research (MISTRA) (DGJL, LF), the Swedish Research Council (VR) (DGJL), Renova AB (LG), The Swedish Research School for Genomics and Bioinformatics (ON) and the Adlerbertska Research Foundation (LG, DGJL).

References

1. Desbrow C, Rutledge EJ, Brighty GC, Sumpter JP, Waldock M: **Identification of estrogenic chemicals in STW effluent: 1. Chemical fractionation and in vitro biological screening.** *Environ Sci Technol* 1998, **32(11):**1549-1558.
2. Larsson DGJ, Adolfsson-Erici M, Parkkonen J, Pettersson M, Berg AH, Olsson PE, Förlin L: **Ethinyloestradiol - an undesired fish contraceptive?** *Aquat Toxicol* 1999, **45:**91-97.
3. Purdom CE, Hardiman PA, Bye VJ, Eno NC, Tyler CR, Sumpter JP: **Estrogenic effects of effluents from sewage treatment works.** *Chem Ecol* 1994, **8:**275-285.
4. Jobling S, Coey S, Whitmore JG, Kime DE, Van Look KJ, McAllister BG, Beresford N, Henshaw AC, Brighty G, Tyler CR, Sumpter JP: **Wild intersex roach (Rutilus rutilus) have reduced fertility.** *Biol Reprod* 2002, **67(2):**515-524.
5. Routledge EJ, Sheahan D, Desbrow C, Brighty GC, Waldock M, Sumpter JP: **Identification of estrogenic chemicals in STW effluent. 2. In vivo responses in trout and roach.** *Environ Sci Technol* 1998, **32(11):**1559-1565.
6. Sumpter JP, Jobling S: **Vitellogenesis as a biomarker for contamination of the aquatic environment.** *Environ Health Perspect* 1995, **103(Suppl. 7):**173-178.
7. Örn S, Holbech H, Madsen HT, Norrgren L, Petersen IG: **Gonad development and vitellogenin production in zebrafish (Danio reio) exposed to ethinylestradiol and methyltestosterone.** *Aquat Toxicol* 2003, **65:**397-411.
8. Parrott JL, Blunt BR: **Life-cycle exposure of fathead minnows (Pimephales promelas) to an ethinylestradiol concentration below 1 ng/L reduces egg fertilization success and demasculinizes males.** *Environ Toxicol* 2005, **20(2):**131-141.
9. Thomas-Jones E, Thorpe K, Harrison N, Thomas G, Morris C, Hutchinson T, Woodhead S, Tyler C: **Dynamics of estrogen biomarker responses in rainbow trout exposed to 17beta-estradiol and 17alpha-ethinylestradiol.** *Environ Toxicol Chem* 2003, **22(12):**3001-3008.
10. Berg AH, Westerlund L, Olsson PE: **Regulation of Arctic char (Salvelinus alpinus) egg shell proteins and vitellogenin during reproduction and in response to 17beta-estradiol and cortisol.** *Gen Comp Endocrinol* 2004, **135(3):**276-285.
11. Larsson DGJ, Mayer I, Hyllner SJ, Forlin L: **Seasonal variations of vitelline envelope proteins, vitellogenin, and sex steroids in male and female eelpout (Zoarces viviparus).** *Gen Comp Endocrinol* 2002, **125(2):**184-196.
12. Rotchell JM, Ostrander GK: **Molecular markers of endocrine disruption in aquatic organisms.** *J Toxicol Environ Health B Crit Rev* 2003, **6(5):**453-496.
13. Snape JR, Maund SJ, Pickford DB, Hutchinson TH: **Ecotoxicogenomics: the challenge of integrating genomics into aquatic and terrestrial ecotoxicology.** *Aquat Toxicol* 2004, **67(2):**143-154.
14. Miracle AL, Ankley GT: **Ecotoxicogenomics: linkages between exposure and effects in assessing risks of aquatic contaminants to fish.** *Reprod Toxicol* 2005, **19(3):**321-326.
15. Ju Z, Wells MC, Walter RB: **DNA microarray technology in toxicogenomics of aquatic models: Methods and applications.** *Comp Biochem Physiol C Toxicol Pharmacol* 2006.

16. Kishi K, Kitagawa E, Onikura N, Nakamura A, Iwahashi H: **Expression analysis of sex-specific and 17beta-estradiol-responsive genes in the Japanese medaka, Oryzias latipes, using oligonucleotide microarrays.** *Genomics* 2006, **88(2):**241-251.
17. Lettieri T: **Recent applications of DNA microarray technology to toxicology and ecotoxicology.** *Environ Health Perspect* 2006, **114(1):**4-9.
18. Moens LN, van der Ven K, Van Remortel P, Del-Favero J, De Coen WM: **Expression profiling of endocrine disrupting compounds using a customized cyprinus carpio cDNA microarray.** *Toxicol Sci* 2006.
19. van der Ven K, Keil D, Moens LN, Hummelen PV, van Remortel P, Maras M, De Coen W: **Effects of the antidepressant mianserin in zebrafish: Molecular markers of endocrine disruption.** *Chemosphere* 2006.
20. Hook SE, Skillman AD, Small JA, Schultz IR: **Gene expression patterns in rainbow trout, Oncorhynchus mykiss, exposed to a suite of model toxicants.** *Aquat Toxicol* 2006, **77(4):**372-385.
21. **consortium Genomics Research on All Salmon** [http://web.uvic.ca/cbr/grasp]
22. Tilton SC, Givan SA, Pereira CB, Bailey GS, Williams DE: **Toxicogenomic profiling of the hepatic tumor promoters indole-3-carbinol, 17beta-estradiol and beta-naphthoflavone in rainbow trout.** *Toxicol Sci* 2006, **90(1):**61-72.
23. Celius T, Matthews JB, Giesy JP, Zacharewski TR: **Quantification of rainbow trout (Oncorhynchus mykiss) zona radiata and vitellogenin mRNA levels using real-time PCR after in vivo treatment with estradiol-17 beta or alpha-zearalenol.** *J Steroid Biochem Mol Biol* 2000, **75(2-3):**109-119.
24. Lin KH, Wang WJ, Wu YH, Cheng SY: **Activation of antimetastatic Nm23-H1 gene expression by estrogen and its alpha-receptor.** *Endocrinol* 2002, **143(2):**467-475.
25. Murphy M, Harte T, McInerney J, Smith TJ: **Molecular cloning of an Atlantic salmon nucleoside diphosphate kinase cDNA and its pattern of expression during embryogenesis.** *Gene* 2000, **257(1):**139-148.
26. Woo Y, Affourtit J, Daigle S, Viale A, Johnson K, Naggert J, Churchill G: **A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms.** *J Biomol Tech* 2004, **15(4):**276-284.
27. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J: **Independence and reproducibility across microarray platforms.** *Nat Methods* 2005, **2(5):**337-344.
28. Severgnini M, Bicciato S, Mangano E, Scarlatti F, Mezzelani A, Mattioli M, Ghidoni R, Peano C, Bonnal R, Viti F, Milanesi L, De Bellis G, Battaglia C: **Strategies for comparing gene expression profiles from different microarray platforms: application to a case-control experiment.** *Anal Biochem* 2006, **353(1):**43-56.
29. Samuelsson LM, Forlin L, Karlsson G, Adolfsson-Erici M, Larsson DGJ: **Using NMR metabolomics to identify responses of an environmental estrogen in blood plasma of fish.** *Aquat Toxicol* 2006, **78(4):**341-349.
30. von Schalburg KR, Rise ML, Cooper GA, Brown GD, Gibbs AR, Nelson CC, Davidson WS, Koop BF: **Fish and chips: various methodologies demonstrate utility of a 16,006-gene salmonid microarray.** *BMC Genomics* 2005, **6:**126.
31. Rise ML, von Schalburg KR, Brown GD, Mawer MA, Devlin RH, Kuipers N, Busby M, Beetz-Sargent M, Alberto R, Gibbs AR, Hunt P, Shukin R, Zeznik JA, Nelson C, Jones SR, Smailus DE, Jones SJ, Schein JE, Marra MA, Butterfield YS, Stott JM, Ng SH, Davidson WS, Koop BF: **Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics.** *Genome Res* 2004, **14(3):**478-490.
32. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17):**3389-3402.
33. R Development Core Team: **R: A Language and Environment for Statistical Computing.** 2006 [http://www.R-project.org]. R Foundation for Statistical Computing
34. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3(1):**Article3.
35. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5(10):**R80.
36. **Bioconductor open source software for bioinformatics** [http://www.bioconductor.org/]
37. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res* 2002, **30(4):**e15.
38. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA: **ArrayExpress--a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2003, **31(1):**68-71.
39. Hoffmann JL, Torontali SP, Thomason RG, Lee DM, Brill JL, Price BB, Carr GJ, Versteeg DJ: **Hepatic gene expression profiling using genechips in zebrafish exposed to 17alpha-ethynylestradiol.** *Aquat Toxicol* 2006.
40. **Primer3** [http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi]
41. Rise ML, Jones SR, Brown GD, von Schalburg KR, Davidson WS, Koop BF: **Microarray analyses identify molecular biomarkers of Atlantic salmon macrophage and hematopoietic kidney response to Piscirickettsia salmonis infection.** *Physiol Genomics* 2004, **20(1):**21-35.