

CHALMERS



GÖTEBORG UNIVERSITY

PREPRINT 2007:29

Weighted Analysis of General Microarray Experiments

ANDERS SJÖGREN
ERIK KRISTIANSSON
MATS RUDEMO
OLLE NERMAN

Department of Mathematical Sciences

Division of Mathematical Statistics

CHALMERS UNIVERSITY OF TECHNOLOGY

GÖTEBORG UNIVERSITY

Göteborg Sweden 2007

Preprint 2007:29

Weighted Analysis of General Microarray Experiments

Anders Sjögren, Erik Kristiansson, Mats Rudemo,
Olle Nerman

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden
Göteborg, September 2007

Preprint 2007:29
ISSN 1652-9715

Matematiska vetenskaper
Göteborg 2007

Weighted analysis of general microarray experiments

Anders Sjögren^{*1,2}, Erik Kristiansson^{1,2}, Mats Rudemo^{1,2}, Olle Nerman^{1,2}

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

²Mathematical 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,

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

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],

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

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],

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

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,

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

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 rhythm 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),

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

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,

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

where F_i denotes the empirical CDF from the i th 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 estimates 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 $g = 1, \dots, m$, 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

$$\mathbf{X}_g \mid c_g \sim N(\boldsymbol{\mu}_g, c_g \Sigma) ,$$

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

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 $(c_1, \mathbf{X}_1), \dots, (c_m, \mathbf{X}_m)$ 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 ,

$$\mathbf{Y}_g = \mathbf{X}_g - \tilde{\boldsymbol{\mu}}_g^0. \quad (10)$$

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

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_A = \mathbf{x}_1^T A^- \mathbf{x}_2, \quad (11)$$

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

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

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 $\mathbf{x} \in \mathbb{R}^n$, $P\mathbf{x} \in \mathcal{V}_0$ 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 $\boldsymbol{\mu}' = P'\mathbf{X}$ and $\boldsymbol{\mu}'' = P''\mathbf{X}$ 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

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

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

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

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

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

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

Since, δ is estimable based on \mathbf{Z} , the likelihood of μ_z can be examined instead of the likelihood of δ .

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

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

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

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

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 μ_z is $\mathcal{V} \cap \mathcal{V}_0^\perp$ and that $\mu_z = 0$ 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

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

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

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

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 $A^T A = \Sigma^{-1}$ and $A \mathcal{P}_{\mathcal{W}} = I_{(l)} A$ and therefore,

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

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 δ when \mathbf{X} is observed is identical to the MLE of δ when \mathbf{Z} is observed. The former is defined by

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

Define $\mathcal{G}_0 = \{\gamma : D \gamma \in \mathcal{V}_0\}$ and $\mathcal{G}_1 = \{\gamma : D \gamma \in \mathcal{V}_0^{\perp}\}$ and note that for any γ there exist $\gamma_0 \in \mathcal{G}_0$ and $\gamma_1 \in \mathcal{G}_1$ such that $\gamma = \gamma_0 + \gamma_1$. Thus,

$$\hat{\delta} = C \underset{\gamma_0 + \gamma_1 : \gamma_0 \in \mathcal{G}_0, \gamma_1 \in \mathcal{G}_1}{\operatorname{argmin}} \| \mathbf{X} - D(\gamma_0 + \gamma_1) \|_{\Sigma}^2 .$$

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

$$\begin{aligned}
\hat{\delta} &= C \underset{\gamma_0 + \gamma_1 : \gamma_0 \in \mathcal{G}_0, \gamma_1 \in \mathcal{G}_1}{\operatorname{argmin}} \left(\| \mathcal{P}_{\mathcal{V}_0}(\mathbf{X} - D(\gamma_0 + \gamma_1)) + \mathcal{P}_{\mathcal{V}_0^{\perp}}(\mathbf{X} - D(\gamma_0 + \gamma_1)) \|_{\Sigma}^2 \right) \\
&= C \underset{\gamma_0 + \gamma_1 : \gamma_0 \in \mathcal{G}_0, \gamma_1 \in \mathcal{G}_1}{\operatorname{argmin}} \left(\| \mathcal{P}_{\mathcal{V}_0}(\mathbf{X} - D \gamma_0) \|_{\Sigma}^2 + \| \mathcal{P}_{\mathcal{V}_0^{\perp}}(\mathbf{X} - D \gamma_1) \|_{\Sigma}^2 \right) ,
\end{aligned}$$

where the second equality follows from the generalised Theorem of Pythagoras [19], the Lemma, and the fact that $\mathcal{P}_{\mathcal{V}_0^{\perp}} D \gamma_0 = 0$ and $\mathcal{P}_{\mathcal{V}_0} D \gamma_1 = 0$. Now since γ_0 and γ_1 minimise the expression independently of each other and since $C \gamma_0 = 0$ by construction,

$$\begin{aligned}
\hat{\delta} &= C \left(\underset{\gamma_0 \in \mathcal{G}_0}{\operatorname{argmin}} \| \mathcal{P}_{\mathcal{V}_0}(\mathbf{X} - D \gamma_0) \|_{\Sigma}^2 + \underset{\gamma_1 \in \mathcal{G}_1}{\operatorname{argmin}} \| \mathbf{Z} - D_z \gamma_1 \|_{\Sigma_z}^2 \right) \\
&= C \underset{\gamma_1 \in \mathcal{G}_1}{\operatorname{argmin}} \| \mathbf{Z} - D_z \gamma_1 \|_{\Sigma_z}^2 .
\end{aligned}$$

For all $\gamma_0 \in \mathcal{G}_0$, $C \gamma_0 = 0$ and $D_z \gamma_0 = 0$, so the area of minimisation can be extended,

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

which is the MLE of δ based on \mathbf{Z} by definition. \square

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 $\tilde{\mu}^0$ is thus truly irrelevant.

Remark 2 Sometimes, additional linear combinations of γ can be assumed to be zero for most genes, $C^* \gamma = 0$ for some matrix C^* with rowspace being a superspace of the rowspace of C . Let P^* be any projection matrix on the corresponding space $\mathcal{V}_* = \{\mu : \mu = D \gamma, C^* \gamma = \mathbf{0}, \gamma \in \mathbb{R}^q\}$ 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 $C \gamma$ 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 $1-\lambda$ 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 $\delta = \mathbf{0}$.

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, \mathbf{Y} , for the different arrays, in the two datasets.

Colour-coding according to sample variance is used for increased clarity (blue for low variance, red for high variance). Differences in variability can be noted for both datasets.

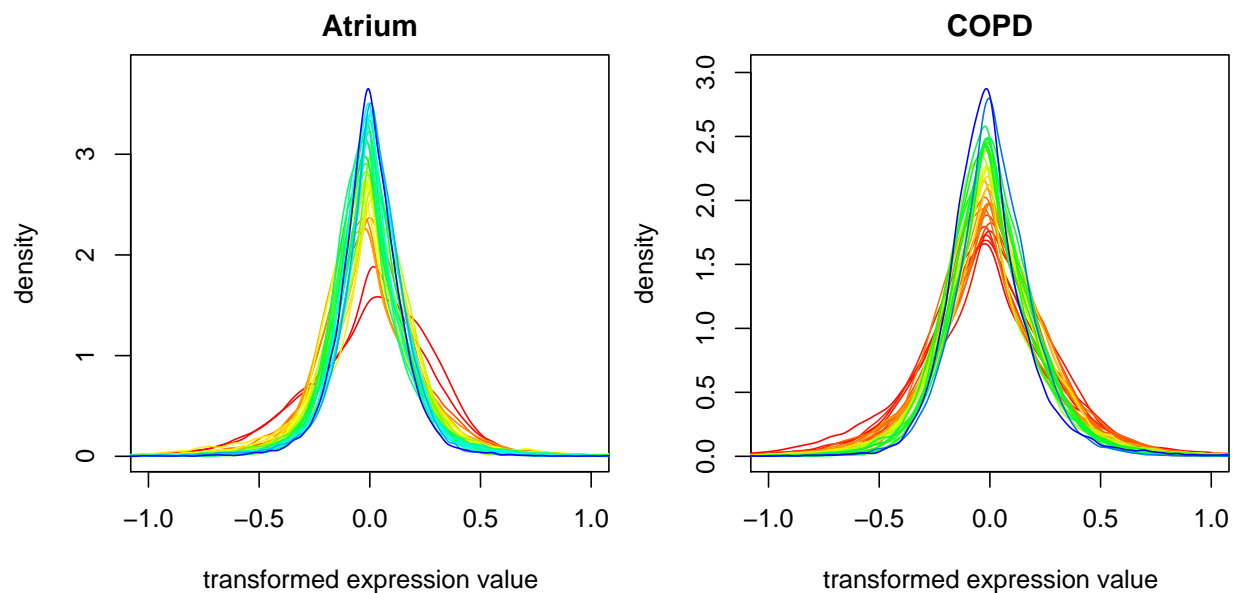


Figure 2 - Pairwise plots

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.

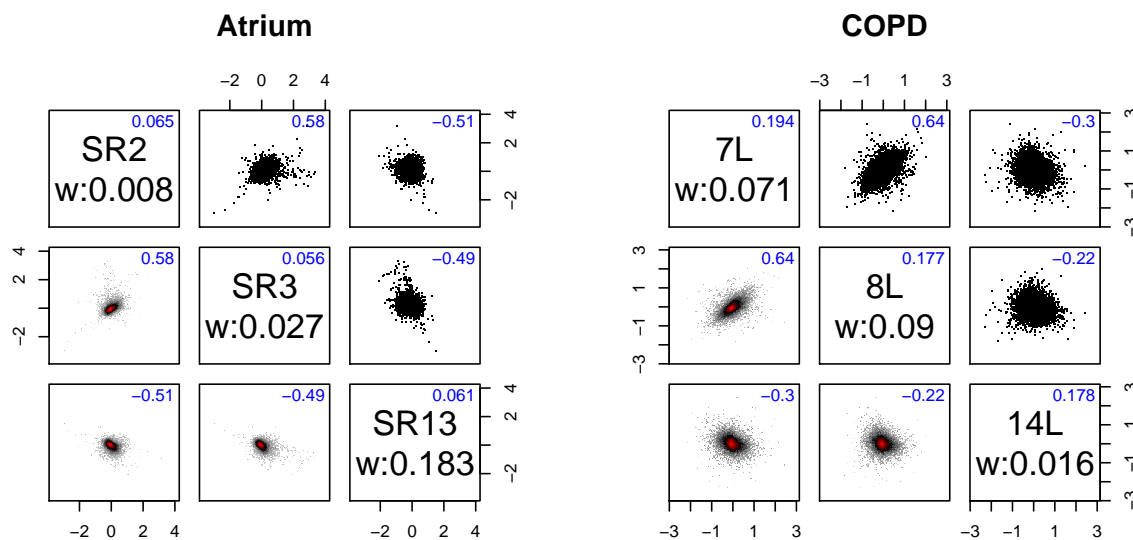


Figure 3 - Observed probability plots

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.

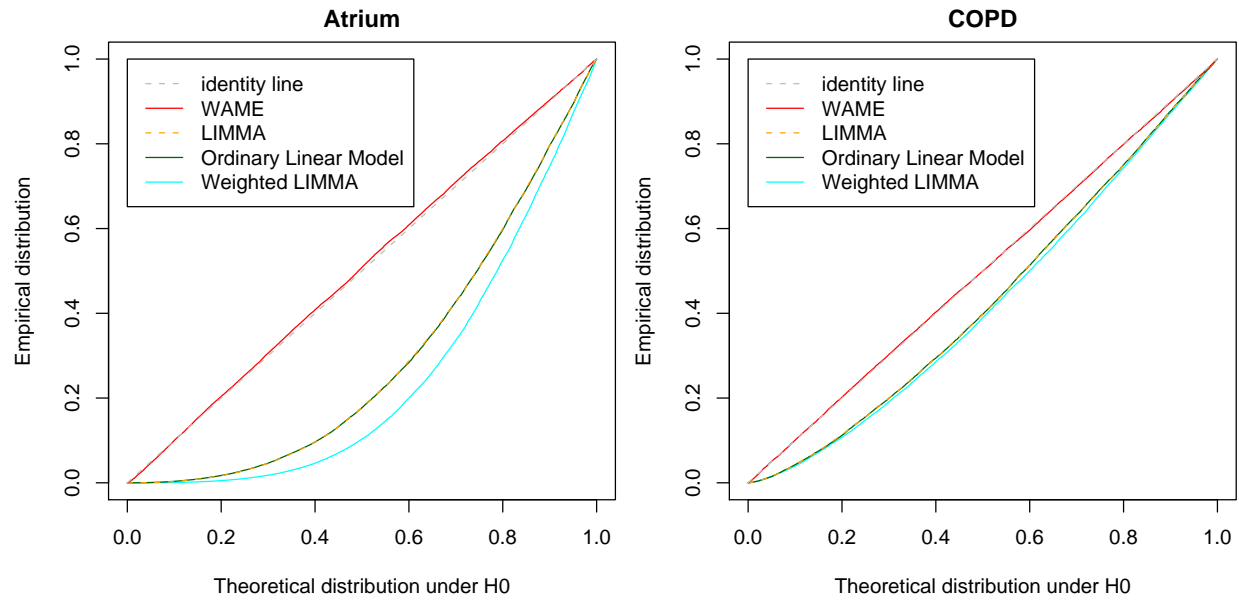


Figure 4 - Probability plots

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.

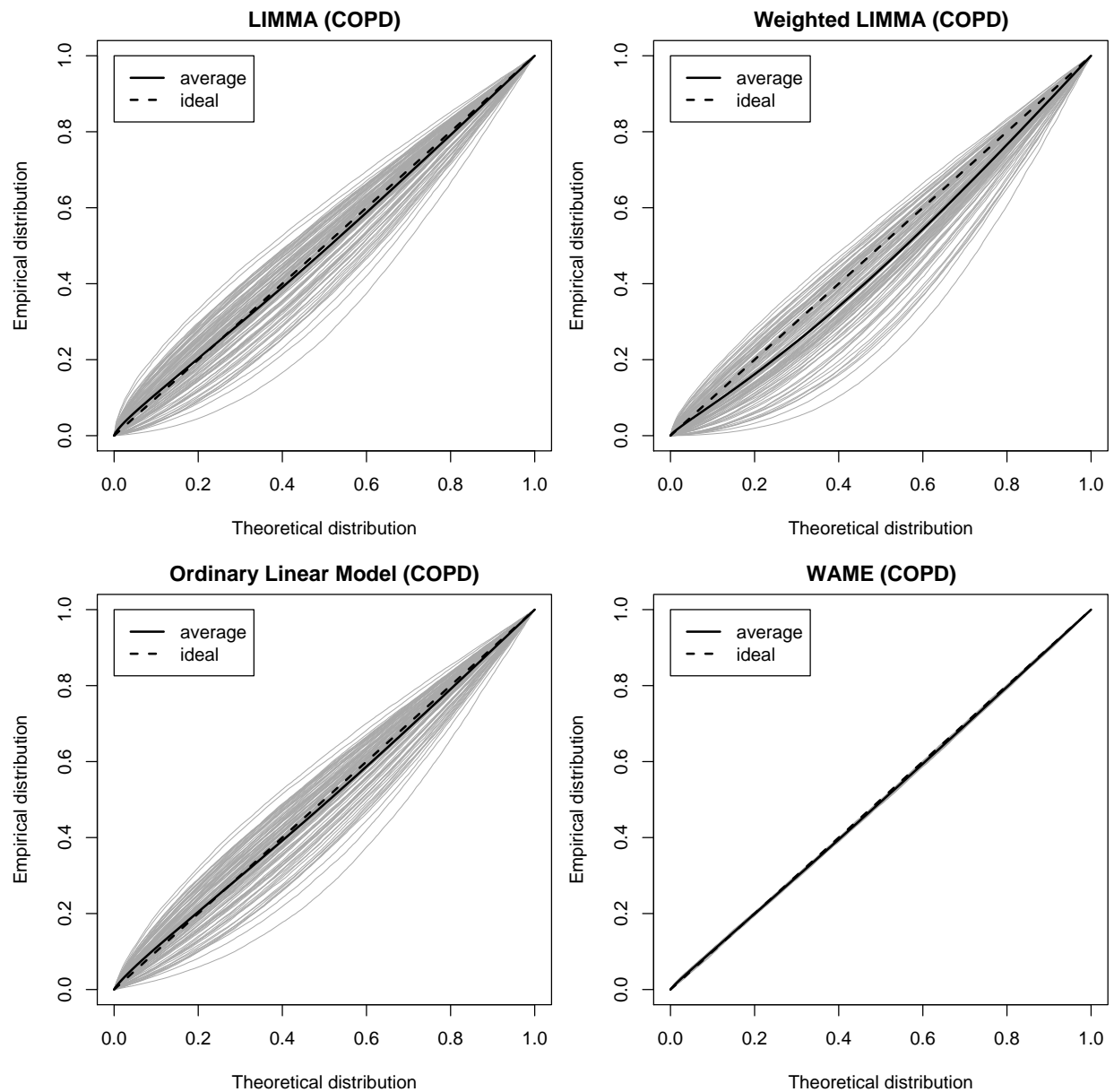


Figure 5 - Average empirical p-value distribution for WAME under regulation

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.

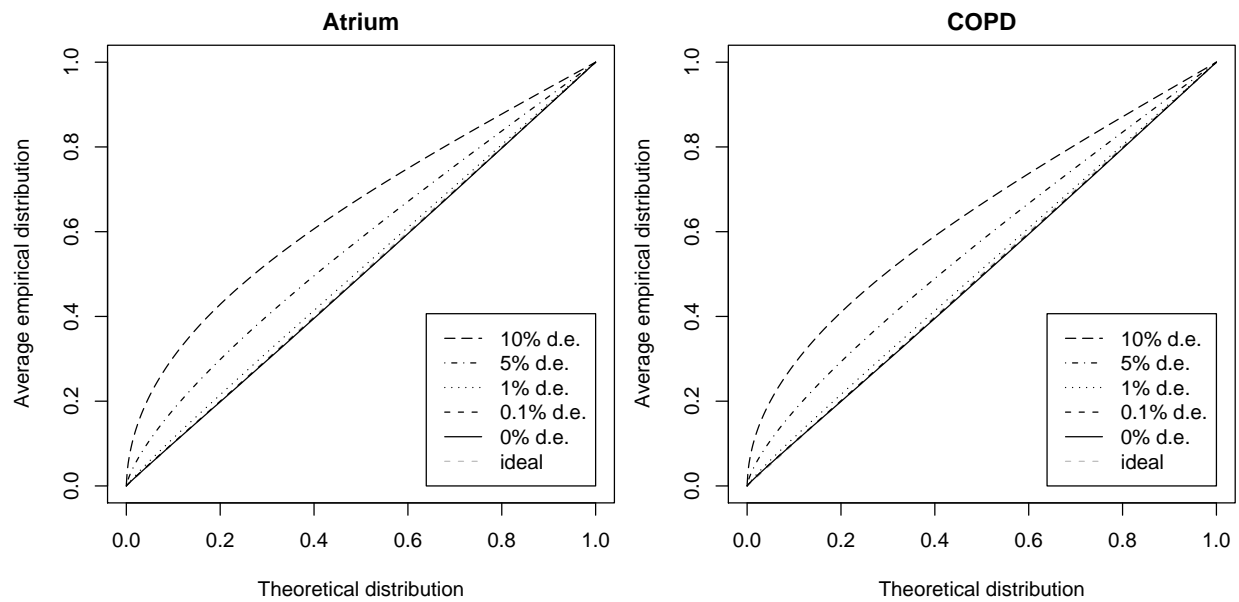
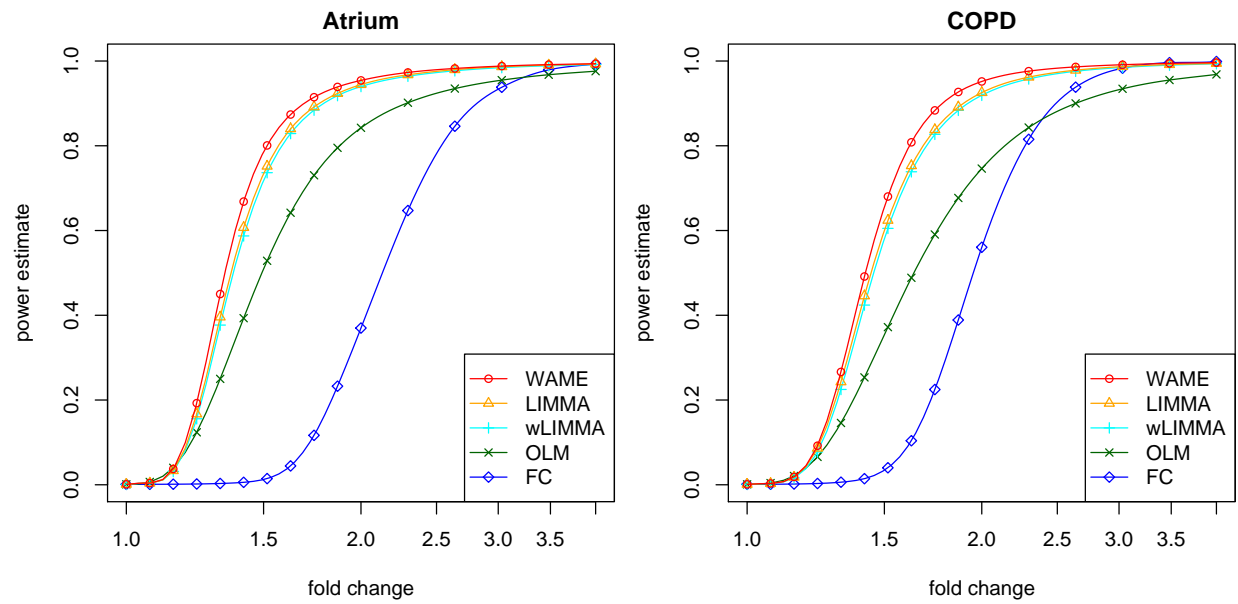


Figure 6 - Estimated power

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