Journal of Biological Systems

COMPARISON OF DIFFERENT METHODOLOGIES TO IDENTIFY DIFFERENTIALLY EXPRESSED GENES IN TWO-SAMPLE CDNA MICROARRAYS

KATHLEEN MARCHAL[1]*, KRISTOF ENGELEN[1], JOS DE BRABANTER[1], STEIN AERTS[1], BART DE MOOR[1]

[1]*Department of Electrical Engineering, ESAT-SCD, K.U.Leuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium*
kathleen.marchal@esat.kuleuven.ac.be
http://www.esat.kuleuven.ac.be/~dna/BioI/

TORIK AYOUBI[2]

[2]*Flanders Interuniversity Institute of Biotechnology (VIB), 9050, Ghent, Belgium*

PAUL VAN HUMMELEN[3]

[3]*Microarray Facility, Flanders Interuniversity Institute of Biotechnology (VIB), 3000 Leuven, Belgium*

Received (Day Month Year)
Revised (Day Month Year)

This review compares different methods to identify differentially expressed genes in two-sample cDNA arrays. A two-sample experiment is a commonly used design to compare relative mRNA abundance between two different samples. This simple design is customarily used by biologists as a first screening before relying on more complex designs. Statistical techniques are quite well developed for such simple designs. For the identification of differentially expressed genes, four methods were described and compared: a fold test, a t-test (Long *et al.*, 2001), SAM (Tusher *et al.*, 2001) and an ANOVA-based bootstrap method (Kerr and Churchill, 2001). Mutual comparison of these methods clearly illustrates each method's advantages and pitfalls. Our analyses showed that the most reliable predictions are made by the combined use of different methods, each of which is based on a different statistic. The ANOVA-based bootstap method used in this study performed rather poorly in identifying differentially expressed genes.

1. Introduction

Microarray experiments measure the expression levels of many genes simultaneously and can be considered as upscaled Northern-blot analyses. Each spot on an array represents a distinct coding sequence of the genome of interest. A two-sample design

*Corresponding author

aims at identifying genes expressed differentially in one condition versus the other. Most biologists start off with such straightforward experiments to roughly identify the genes involved in the biological system studied. Based on the conclusions drawn, more complex experiments are designed. Therefore, it is of importance to reliably identify the genes that are differentially expressed. To increase reliability, nowadays experiments are designed as such that more replicates of each measurement are available. The availability of replicates allows conclusions about differentially expressed genes to be inferred in a statistically more solid way. However, due to the high experimental cost the number of available replicate measurements remains limited. Novel algorithms designed for the analysis of such two sample experiments are being developed. In this study the performance of four such methods was evaluated. Although these methods can theoretically be used for other types of arrays, we focused on cDNA arrays only. During a cDNA microarray experiment, mRNA of a reference (condition 1) and induced sample (condition 2) is isolated and each labeled with a distinct fluorescent dye. Subsequently, both labeled samples are hybridized simultaneously to the array. Fluorescent signals of both channels are measured and used for further analysis (for reviews on cDNA microarrays see [3,2,18]). A complete analysis flow is not only restricted to identyfying differentially expressed genes but generally comprises a data transformation, a data filtering and a normalization step prior to performing a statistical test. In a cDNA array the rate to which a gene is differentially expressed is usually estimated by the ratio of the expression levels between two the conditions.

2. Data set

The dataset used in this review compares a spontaneous knock-out (KO) and wild-type (WT) mouse. In the spontaneous knock-out mouse, the Hmgi-c gene has been disrupted by a 100kb deletion. Hmgi-c RNA was consequently not transcribed and therefore did not result in a protein. HMGI proteins play a critical role at promoter regions in the correct assembly and stabilization of higher order protein-DNA complexes required for efficient transcriptional activation of genes [6].

From both mice mRNA was extracted, labeled and hybridized on a mouse cDNA microarray containing 4202 cDNA fragments of 0.5 to 2kb. The cDNA fragments were PCR amplified, purified and spotted in duplicate on Type-VII silane coated slides (cat#: RPK0174, Amersham BioSciences, UK) using a Molecular Dynamics Generation III printer with 12 capillary pins (Amersham BioSciences). The duplicate spots were arrayed distant from each other on the left and right side of the slide. For the probes, 5 μg of total RNA was amplified using a modified protocol of *in vitro* transcription as described earlier and labeled during a reverse transcription reaction of the amplified RNA [15] with either Cy3-dCTP (green dye) or Cy5-dCTP (red dye). The probes were mixed and hybridized overnight using an automatic slide processor (Amersham BioSciences). The hybridizations were repeated in the following way: in a first analysis, the test sample (KO) was labeled with the red dye while the corresponding reference (WT) was labeled with the green dye, and in a second analysis the colors were reversed (i.e. color flip experiment). Since every gene was spotted in duplicate, such design resulted in four measurements per gene for each condition tested.

3. Data Preparation

Prior to performing the distinct statistical tests, data were preprocessed as outlined in this paragraph. Background corrected raw measurements were log transformed. The error observed in microarray data is a superposition of a multiplicative and an additive error [16]. Log transforming the data compensates for the multiplicative error but at the expense of an increased additive error at low expression levels. Because observing a higher measurement error at low signal intensities is intuitively plausible and removal of multiplicative errors is essential for most statistical tests, log transformation is advisable [1,11,13].

Genes for which at least one measurement contained a zero value were treated separately. When dividing by zero values or taking the log of a zero value during analysis, zero values result in undefined values. When not treated seperately, the information about such genes is lost. However, in a two-sample experiment zero values in one particular condition might correspond to genes differentially switched off. In this particular example all genes containing a zero value behaved inconsistent meaning that the value of zero was dye-dependent rather than condition-dependent [13]. Genes, consistently switched on in one condition and off in the other were not detected.

Data were normalized in order to remove consistent sources of variation such that, for each gene, the measured value reflects the mere expression level as caused by the condition tested. These consistent sources of variation characteristic for cDNA arrays include array, dye, condition and spot effects. Array effects refer to the differences in hybridization efficiency between different slides. Condition and dye effects reflect differences in respectively mRNA isolation and labeling efficiencies between two distinct samples while spot effects refer to the difference in amount of cDNA spotted on the array. A global normalization procedure was used [21]. Global normalization assumes that only a small fraction of the total number of genes on the array alters its expression level and that symmetry exists in the number of genes that is upregulated versus downregulated. Remark therefore that the assumption of global normalization applies only to microarrays that contain a random set of genes and not to dedicated arrays. Under the assumption of global normalization the average intensity of the test genes should be equal to the average intensities of the reference genes. Based on the hypothesis of global normalization, for the bulk of the genes the $log_2(test/reference)$ ratio should equal 0. Normalizing the data consists of finding the right transformation factor that allows centering the $log_2(test/reference)$ for the bulk of the genes around zero. Linear normalization assumes a linear relationship between the measurements in both conditions (test and reference) and uses a constant transformation factor which can either be the mean or median of the log intensity ratios or a regression factor as determined by linear regression.

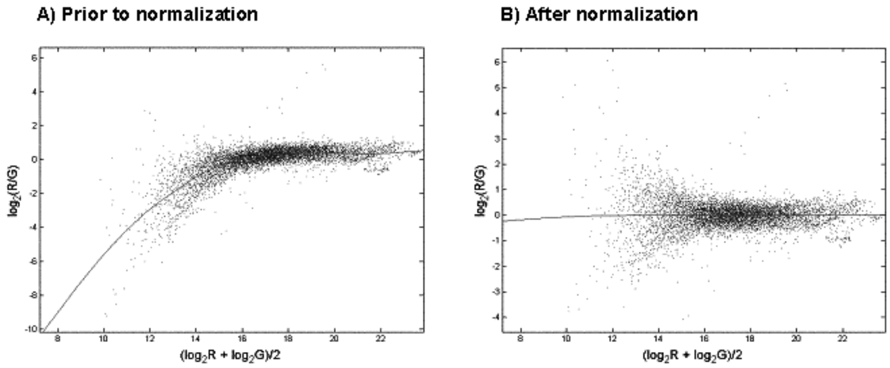

Fig. 1. Illustration of the influence of an intensity-dependent normalization. Panel A: representation of the log-ratio $M = log_2(R/G)$ versus the mean log intensity $A = (log_2(R) + log_2(G))/2$. At low average intensities the ratio becomes negative indicating that the green dye is consistently more intense as compared to the intensity of the red dye. This phenomena is referred to as the non-linear dye effect. Either the sensitivity of the red signal is lower than the one of the green signal or the basal noise level on the green signal is more pronounced. Solid line represent the Lowess fit with f value of 0.02. (R = red; G= green). Panel B: Representation of the ratio $M = log_2(R/G)$ versus the mean log intensity $A = (log_2(R) + log_2(G))/2$ after performing a normalization and linearization based on the Lowess fit. Solid line represent the new Lowess fit [21] with f value of 0.02 on the normalized data. (R = red; G= green).

The actual dependency between the measurements in both conditions on one slide is illustrated in Fig. 1 using a plot of M ($log(R/G)$ versus A (the average expression level in log scale), as suggested by Yang *et al.* [5]. The relationship between dyes clearly depends on the measured intensity. These intensity-dependent dye effects result in non-linearities that are most pronounced at extreme intensities (either high or low). From Fig. 1 it is clear that in a certain range of average intensities A, the log ratio M approximates a certain constant level. In this range a constant normalization factor can be used. However, as the average expression value (A) decreases, the log ratio (M) deviates from a constant level and the use of an intensity-dependent rescaling factor is more appropriate. Therefore a robust scatter plot smoother that performs locally linear fits, Lowess was used Yang *et al.* [5]. The results of this fit can be used to simultaneously linearize (i.e. remove non-linear dye effects) and normalize the data (remove consistent sources of variation due to within slide dye and condition effects)([21,5] (Fig. 1). Since the M versus A plot was used to fit the data, for each gene novel normalized ratio estimates were calculated. From these normalized ratios, new values for the absolute expression levels can be derived. In our approach relative expression levels are approximated by the ratio

$(log_2(ratio) = log_2(condition_1) - log_2(condition_2) = log(\bar{y}_{i1}) - log(\bar{y}_{i2}))$. Using a ratio allows intrinsic compensation for spot effects.

Table 1. Definitions of statistical terms.

Residual
Residuals are the deviations of observed values from their estimated or fitted values. A residual may be regarded as the observed error, in distinction to the actual unknown population error of the model.

Additive error
The absolute error on a measurement is independent of the measured expression level. Consequently, the relative error is inversely proportional to the measured intensity and is high for measurements of low intensity. When replicate measurements are plotted against each other, additive errors result in a constant residual scattering.

Multiplicative error
The absolute error on the measurement increases with the measured intensity. The relative error is constant but the variance between replicate measurements increases with the mean expression value. Multiplicative errors cause signal-dependent variance of residuals.

t-test
A t-test can be defined as a hypothesis test that assumes that the observations are drawn at random from a normal population and that employs a Student t-distributed test statistic for confidence interval estimation. The t-distribution describes the distribution of a normal variable, standardized with the sample variance s^2 as opposed to the population variance σ^2. It is used for hypothesis testing of normally distributed variables when the population variance σ^2 is unknown, in which case the sample variance s^2 is used as an estimator of σ^2.

Paired t-test
The paired t-test is a special case of the two-sample t-tests of hypotheses that occurs when the observations on the two populations of interests are collected in pairs (in a cDNA microarray experiment, measurements of the reference and test for a particular gene, assessed on the same array and the same spot are paired). The difference with an unpaired two-sample t-test is that both variables are presumed to be dependent. This translates into the incorporation of the covariance between both variables in the test statistic. As a result, a positive correlation within the pairs can cause the unpaired two-sample t-test to considerably understate the significance of the data if it is incorrectly applied to paired samples.

Power
The power of a statistical test (computed as 1-β, with β the probability of a type II error) is the probability of rejecting the null hypothesis H_0 when the alternative hypothesis is true. It can be interpreted as the probability of correctly rejecting a false null hypothesis. Power is a very descriptive and concise measure of the sensitivity of a statistical test, i.e. the ability of the test to detect differences.

Correction for multiple testing
When considering a family of tests, the level of significance and power are not the same as those for an individual test. For instance, a significance of $\alpha = 0.01$ for individual gene expression indicates a probability of 1% of finding a ratio similar to the measured ratio under the null hypothesis (no differential expression present). This means that for every 1000 genes tested (a family of 1000 tests), 10 would be expected to pass the test though not differentially expressed. To limit this number of false positives in a multiple test, a correction is needed (e.g. Bonferroni correction).

Heteroscedasticity
The condition of the error variance not being constant over all cases.

4. Identification of differentially expressed genes

When consistent sources of variation have been removed, the different ratio estimates of a particular gene can be combined to find out whether a gene is differentially expressed. In this paragraph distinct methods to perform this analysis are compared.

4.1. *Fold test*

The fold test is a simple selection procedure that makes use of an arbitrary chosen threshold. For each gene an average ratio (arithmetic mean of the *logratio*) is calculated based on the different ratio estimates ($logratio = log(\bar{y}_{i1}) - log(\bar{y}_{i2})$). Average ratios of which the expression ratio exceeds a threshold (usually twofold) are retained. In this particular dataset 110 genes exceeded a two fold threshold. The fold test is based on the intuition that a larger observed fold change can be more confidently interpreted as a stronger response to the environmental signal than smaller observed changes. Note that a fold test discards all information obtained from replicates [1].

4.2. *Other Statistical tests*

A plethora of novel methods to identify differentially expressed genes in a statistically more founded way have recently been proposed provided replicates are available (see Table 2). Distinct classes of models can be discerned, differing from each other in the test statistic used, in the way the null hypothesis is modeled and in their underlying assumptions. For the technical details for each of these methods we refer to the individual references (see Table 2). As examples, we used in this study the method described by Baldi and Long [1] and the SAM method of Tusher *et al.* [20] because to our opinion, though quite advanced, these methods are still most intuitive and straightforward to understand for non-expert users.

4.2.1. *t-test*

A t-test (see Table 1) is more appropriate to make statistical inference about the differential expression of a gene than a simple fold test since it does not only take into account how much a gene is differentially expressed, but also the consistency of the individual measurements, used to assess the average differential expression level. The non-paired t-test evaluates if the average expression level of a gene in the test condition is significantly different from its average expression level in the reference condition. The H_0 hypothesis states that the expression level of the test and reference are equal. The formula to compute the test statistic is depicted in Table 2. To calculate the within sample variance of a regular non-paired t-test, the four observations of the test are used to estimate the mean expression level of the gene in the test condition. In the same way the four measurements of the

Table 2. Overview of recently described methods to determine differentially expressed genes across two conditions.

Method	**Assumptions**	**Test statistic**	**Error restrictions**	**Distribution** H_0: $\mu_1 = \mu_2$	**Additional modifications**
Independent Samples t-test for equality of means [a]	Observations are independent Observations for each group are a sample from a population with a normal distribution Unequal sample variance Unequal sample size	$t_i = \frac{\bar{y}_{i1}-\bar{y}_{i2}}{\sqrt{\frac{s_{i1}^2}{n_1}+\frac{s_{i2}^2}{n_2}}}$ $df = \frac{(\frac{s_{i1}^2}{n_1}+\frac{s_{i2}^2}{n_2})^2}{\frac{\frac{s_{i1}}{n_1}^2}{n_1-1}+\frac{\frac{s_{i2}}{n_2}^2}{n_2-1}}$	Errors normally distributed	Parametrized : Student t-distribution	Emperical Bayesian estimate of variance
Paired Samples t-test [b]	Each pair of measurements is independent of other pairs Differences are from a normal distribution Unequal sample variance Equal sample size $n_1 = n_2 = n$	$t_i = \frac{\bar{y}_{i1}-\bar{y}_{i2}}{\sqrt{\frac{s_{i1}^2+s_{i2}^2-2cov(y_{i1},y_{i2})}{n}}}$ $df = n-1$	Errors normally distributed	Parametrized : Student t-distribution	
Weigthed least squares [cd]	Unequal sample variance Unequal sample size	$t_i = \frac{\bar{y}_{i1}-\bar{y}_{i2}}{\sqrt{\frac{s_{i1}^2}{n_1}\cdot\frac{n_1-1}{n_1}+\frac{s_{i2}^2}{n_2}\cdot\frac{n_2-1}{n_2}}}$	Unequal error variances acceptable	Parametrized: Standard normal distribution	Weighted least squares
Mixture model approach [d]	Unequal sample variance Unequal sample size	$t_i = \frac{\bar{y}_{i1}-\bar{y}_{i2}}{\sqrt{\frac{s_{i1}^2}{n_1}+\frac{s_{i2}^2}{n_2}}}$	Errors equal variance (iid) and symmetrically distributed	test statistic used in a likelihood ratio test	distributions estimated by Normal mixture models
SAM [e]	Equal sample variance: use of 'pooled' variance s_{ip}^2 Unequal sample size	$t_i = \frac{\bar{y}_{i1}-\bar{y}_{i2}}{s_0+s_{ip}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	Errors equal variance (iid)	No explicit H_0 distribution but use of order statistics	Addition of s_0 to ensure that the distribution of t_i is independent of the level of gene expression

Note: Each of the methods uses variations of a mean and variance normalized test statistic t_i. The methods differ from each other in the way the corresponding significance level is calculated. A first class of methods makes use of simple t-test statistic. For each gene i the test statistic t_i is calculated. $\bar{y}_{i1}$: average expression level of the n_1 replicates of gene i in the first condition, s_{i1}: within variance of this group of replicates, $\bar{y}_{i2}$, s_{i2}: similar but for the second condition. Based on the calculated t_i value, a preset significance level and the degrees of freedom the corresponding p-value is calculated. The p-value expresses the probability of finding a certain value of the test statistics t_i by coincidence assuming that both genes were not differentially expressed (H_0 hypothesis). As a H_0 distribution, a parametrized (Student t-distribution) is used for small sample sizes. Due to the small sample size and the corresponding low degrees of freedom, t-tests have a low power. Non-parametric alternatives to the t-test and the paired t-test respectively are the Wilcoxon Rank Sum test and the Wilcoxon Signed rank test. For a sufficiently large sample size, the test statistic t_i used by Thomas *et al.* [19] and that of the regular t-test may be considered equal. For small sample sizes Thomas *et al.* [19] make use of the maximum likelihood estimator of the variance. The advantage of the model of Thomas *et al.* [19] is that it does not assume a constant variance of the error term. However, to calculate their significance, Thomas *et al.* [19] make use of a normal distribution for H_0, which might be too strong an assumption viewing the small sample size. The second class of models estimates the distribution of H_0 directly by permutation analysis (a comparable method is used by Kerr *et al.* [11]). The mixed model described by Pan *et al.* [14] make use of complex estimation procedures to determine the distribution of H_0 while the method of Tusher *et al.* [20] uses order statistics. In contrast to the other approaches, the SAM method assumes that the variances are equally distributed and therefore uses the pooled variance $s_{ip}^2 = \frac{(n_1-1)s_{i1}^2+(n_2-1)s_{i2}^2}{n_1+n_2-2}$ as an estimator of $\sigma_{i1}^2 = \sigma_{i2}^2 = \sigma_i^2$. [a](Baldi and Long, 2001), [b](Dudoit, Yang *et al.*, 2000), [c](Thomas, Olson *et al.*, 2001), [d](Pan, 2001), [e](Tusher, Tibshirani *et al.*, 2001) reference are considered as a single group. The standard deviations (s_{i1}, s_{i2}) are computed based on the deviation of the different measurements of a group from their respective group means ($\bar{y}_{i1}, \bar{y}_{i2}$) (Table 2). Of course when the within variance is calculated in such a way it intrinsically contains the consistent variations due to array and spot effects (the absolute expression values instead of the ratios are used to calculate an estimate of the average differential expression level). This problem can be overcome by using a paired t-test. Indeed, in a cDNA array the reference and test measurements for the same gene, assessed on the same array and the same spot can be treated as paired observations. In Table 2 is outlined how a paired t-test (Table 1) for cDNAs is calculated. For computation of the variance, a pair of observations is considered as a new variable ($log(\bar{y}_{i1}) - log(\bar{y}_{i2})$)). The within group variation, as calculated by a paired t-test evaluates the deviation of this new variable from the mean of that variable (i.e. the variation between the $log(\bar{y}_{i1}) - log(\bar{y}_{i2})$))). As such a paired t-test, in contrast to a regular non-paired t-test intrinsically compensates for the variation over spots and arrays. The lower within group variation increases the power (Table 1) of a paired t-test as compared to a regular t-test. Note that when performed on the log transformed data, the t-test approach can be considered as the counterpart of the fold test (calculating $log(\bar{y}_{i1}) - log(\bar{y}_{i2})) = log(\bar{y}_{i1}/\bar{y}_{i2})$). The theoretical advantage of a (paired) t-test is that smaller fold changes are considered significant for genes whose expression levels are measured with great accuracy (high consistency) and large fold changes are considered non-significant if expression levels were not measured accurately (low consistency). Using the paired t-test of Baldi and Long [1] on our dataset resulted in 186 genes with an individual p-value lower than 0.01 (106 genes with a p-value lower than 0.005). Usually a t-test is combined with a correction for multiple testing (see Table 1). The implementation of Baldi and Long (Cyber-T) uses a Bonferroni correction [1]. Only 3 genes in our dataset passed the significance test after correction for multiple testing (assuming an experiment wide false positive rate of 0.25). Therefore, the single step adjusted p-values, as implemented in the Cyber-T software are seemingly too conservative, decreasing the power of the statistical test (ability to detect real positives). Moreover, the choice of the Bonferroni correction factor is quite arbitrary. To handle these pitfalls, other corrections for multiple testing have been proposed recently [5]. Long *et al.* [12] provide other extensions to their implementation of the t-test such as the Bayesian t-test, a methodology developed to cope with the low number of replicates. For more information on this topic we refer to Long *et al.* [12].

4.2.2. *SAM*

SAM (Significance Analysis of Microarrays) is another method for the analysis of paired or unpaired black/white experiments [20]. Instead of calculating a $t(i)$-value, SAM calculates for each gene a modified $t(i)$ value, called relative difference and referred to as $d(i)$ in the original article (see Table 2). The difference between $t(i)$

and $d(i)$ calculated by SAM is the constant term s_0, used to compensate for the dependency of the distribution of $d(i)$ on the measured expression level. After calculating for each gene the corresponding $d(i)$ value, genes are ranked according to their $d(i)$ value. The higher the $d(i)$ value (in absolute value), the more likely that the gene will be differentially expressed. Instead of calculating a p-value using a student t-distribution, genes called differentially expressed are identified by performing a permutation analysis. New random datasets are generated by permuting the original data. In such permutated datasets none of the genes is differentially expressed. The $d(i)$ values in these randomized datasets are calculated, ranked and subsequently used to infer the expected differences i.e. the $d(i)$ value that can be expected if a gene is not differentially expressed. By using a scatterplot (Fig. 2), ranked $d(i)$ values of the experimental dataset are compared to ranked expected $d(i)$ values.

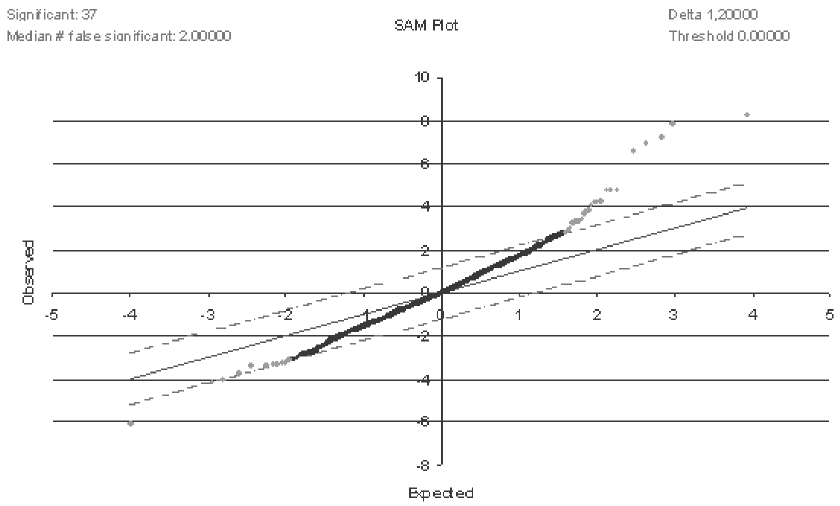

Fig. 2. Result of a SAM analysis on the preprocessed dataset (see Section 3). The following parameter settings were used: paired test, permutation analysis: 1000 iterations, delta value = 1.2, threshold = 0.

The delta value, a user-specified parameter determines the number of significantly expressed genes, it expresses how much the measured $d(i)$ value should exceed the expected $d(i)$ value in order to consider a gene significantly expressed (delta measured as a displacement of the $d(i)$ value from the $d(i) = d_{Expected}(i)$ line). The number of false positives can be estimated as the number of genes present in the permuted dataset for which the $d(i)$ value exceeds the lowest $d(i)$ value that was considered significant based on a given setting of the delta slider. Permutation analysis overcomes the need of a high number of replicates and is used as an alternative to correction for multiple testing. Using a paired test and a value for the delta slider of 0.93, 106 genes were considered as differentially expressed with a median number of false positives of 7. The setting of the delta slider allows choosing a tradeoff between the number of false positives (type I error) and the number of false negatives (type II error). The lower the number of false positives, the more stringent the test and the lower the number of genes withheld as significant. The SAM software outputs a listing of the number of genes withheld and the possible number of false positives for each different value of the deltaslider.

4.2.3. *ANOVA-based bootstrapping*

Another method used in this study to identify differentially expressed genes is based on the use of a bootstrap confidence intervals. Bootstrapping allows creating confidence intervals based on an estimate of the experimental noise distribution of the dataset. This experimental noise can be estimated by using an ANOVA-based approach [10]. ANOVA can be viewed as a special case of multiple linear regression where the explanatory variables are entirely qualitative. ANOVA models the measured expression level of each gene as a linear combination of the explanatory variables that reflect, in the context of this study, the major sources of variation in a microarray experiment. Several explanatory variables representing the condition, dye and array effects and combinations (2, 3 and 4 level combinations) of these effects are taken into account in the models. One of the combined effects, the Gene-Condition (GC) effect, reflects the expression of a gene merely depending on the tested condition (i.e. the condition-specific expression). Since this is the effect in which biologists are interested it is referred to as the *factor of interest*. Similarly the difference between the GC effects of two conditions reflects the differential expression and is called the *contrast of interest*. Of the other combined effects only those having a physical meaning in the process to be modeled are retained. Reliable use of an ANOVA model therefore requires a good insight into this process. The residuals of the fit can be considered as estimates of the experimental noise, likewise the fitted values are estimates of the measurement values devoid of the noise. Remark also that ANOVA can not only, estimate the experimental noise in the dataset but also inherently performs a multidimensional linear data normalization.

The use of ANOVA requires two major assumptions to be satisfied. At first the data should adequately be described by the linear ANOVA model. Secondly, observations should be normally distributed with constant within group variances equal for all groups. If both these assumptions are satisfied, the major advantage of using ANOVA for normalization consists of its ability to assess the different sources of variation across the entire experiment (i.e. the entire set of arrays) instead of treating each slide separately. In contrast to the slide by slide approach, all measurements are combined during statistical inference. Satisfying both requirements, mentioned above, results in the model errors (as estimated by the residuals of the fit) being independently and normally distributed random variables with zero mean and constant variance. The behavior of the residuals can be observed by visual inspection of the residual plots (Fig. 3). If the data can not be fitted by a linear model (not satisfying the first assumption), residual plots show a non-linear behavior which can best be observed by plotting the residuals against the estimated values for the individual combinations of the major effects (i.e. dye, array and condition effects). Not satisfying the second assumption results in heteroscedasticity (Table 1), indicated by an observed wedge-shaped trend in the residual plot. When both assumptions are satisfied and the residual distribution shows only slight deviations from normality (so that the actual errors, estimated by the residuals can be assumed to be normally distributed) significantly differentially expressed genes can be identified by constructing confidence intervals on the difference in GC effect. These confidence intervals are then based on normal assumptions. If the distribution of the residuals shows serious deviations from normality, confidence interval construction can still be done, but bootstrap analysis should be used as an alternative. In bootstrap analysis, similar to the permutation analysis of SAM, no explicit assumption on the distribution of the errors is made, but confidence intervals are estimated based on novel *in silico* generated datasets. The only assumption is that the errors are identically and independently distributed i.e. assuming a constant error variance (*iid*). Fitting the ANOVA model results in a set of residuals and estimated values $\hat{y}$. By adding a residual, randomly sampled-with-replacement from the available set of residuals to the estimated expression values, thousands of novel bootstrapped datasets can be generated. In each of the novel dataset the difference in GC effect between two conditions is calculated, as a measure for the differential expression. Based on these thousands of estimates of the difference in GC effect, a bootstrap confidence interval can be calculated [9].

Different ANOVA models to describe microarray experiments, originally proposed by Kerr *et al.* were tested on our dataset. Each of these differs in the number of additional combined effects included [11]. The model that performed best consisted of an adaptation on our part of the original models (Table 3). The model includes array, dye, condition and gene effects. Combined effects include GC effect and spot effects. Spot effects are modeled by assuming a relationship between spots on the same array and a relationship between all left and right spots.

As mentioned in the section data preparation, microarray data show a strong non-linear behavior. This behavior prohibits readily using linear ANOVA models on non-linearized data. The influence of using a linear model on non-linearized microarray data is illustrated by Table 3A and Fig. 3A. In the ANOVA table represented in Table 3, the SS-value describes for each effect its contribution to the global variation in the experiment. Using our ANOVA model, without prior linearization resulted in residuals being far from normally distributed and showing an apparent slight heteroscedasticity (a non-constant variance of the residuals) at low expression levels (data not shown). By plotting the residuals against the estimated values for the individual combinations of major effects (see Fig. 3A), it was clear that the observed heteroscedasticity did not only result from non-constant variance

Table 3. ANOVA Results of the ANOVA model before and after linearization by Lowess

Source	SS	df	MS	SS	df	MS
	without Lowess normalization			**Lowess normalization**		
G-effects	1333176.7	3784	35.2	1333176.7	3784	35.2
C-effects	4536.6	1	4536.6	0	1	0
A-effects	22432.4	1	22432.4	22432.4	1	22432.4
D-effects	4822.5	1	4822.5	0	1	0
AG-effects	9052.6	3784	2.4	9052.6	3784	2.4
RG-effects	2164	3785	0.6	2164	3785	0.6
GC-effects	1313.6	3784	0.3	1239.8	3784	0.3
Error	7993.1	15139	0.3	5656	15139	0.4
Corrected Total	185492.2	30279	6.1	173722.2	30279	5.7

ANOVA model: $I_{ijklm} = \mu + G_i + C_j + A_k + D_l + R_{m(i)} + (AG)_{ki} + (GC)_{ij} + \epsilon_{ijklm}$
μ: overall mean of the expression levels, A: array effect, D: dye effect, G: gene effect, C: condition effect, GC: effect of interest, R: replicate effect, AG: combined effect representing a spot effect, i: number of genes, j: number of conditions, k: number of arrays, l: number of dies, m: number of replicates. ANOVA tables: represent for each effect in the corresponding ANOVA model its contribution to the total variance (SS = sum of squares error). The residual SS, represented by *Error* is the variation in the dataset that could not be explained by any of the effects. The total variation in the dataset represented by *Corrected Total*. Df: degree of freedom, MS: mean square error. Left part: Data were partially preprocessed as in Section 3 but no normalization by Lowess was performed. Right Part: Data were completely preprocessed as in Section 3.

in the dataset (presence of additive error in the low expression range) but that non-linear effects occurred in the data. To minimize the influence of the non-linearity, non-linear dye effects were removed by performing a Lowess fit prior to ANOVA. The results of the ANOVA model on Lowess modified data are depicted in (Table. 3, Panel B). The effect of the Lowess normalization is reflected by the zero contribution of the dye and condition effects in the ANOVA table. When residuals were plotted as shown in Fig. 3B for each array and dye separately after Lowess fit it is clear that non-linear tendencies have sufficiently been removed. It should be noted, however, that this was most often not the case. We regularly observed in other datasets that non-linearities not only originated from non-linear dye effects but also, though less pronounced, from non-linear array and/or condition effects that either compensated or exacerbated the corresponding dye effects. In such case the use of ANOVA to estimate the error is not advisable. Because in this dataset only a slight heteroscedasticity was observed, the studentized residuals of the model were used for bootstrap analysis. As such we could identify 163 genes as potentially differentially expressed based on a 95 % and 71 genes based on a 99 % bootstrap confidence interval. Remark that by performing a Lowess normalization prior to applying ANOVA, we only use the ANOVA model to estimate the experimental noise and do not make use of its ability to normalize data. Secondly, for the experimental design used in our study the difference in GC effect, after performing a Lowess fit approximates the log of the ratio. Therefore, as applied in this study, ANOVA is used as an alternative method to estimate statistically differentially expressed genes that are approximated by their log ratio.

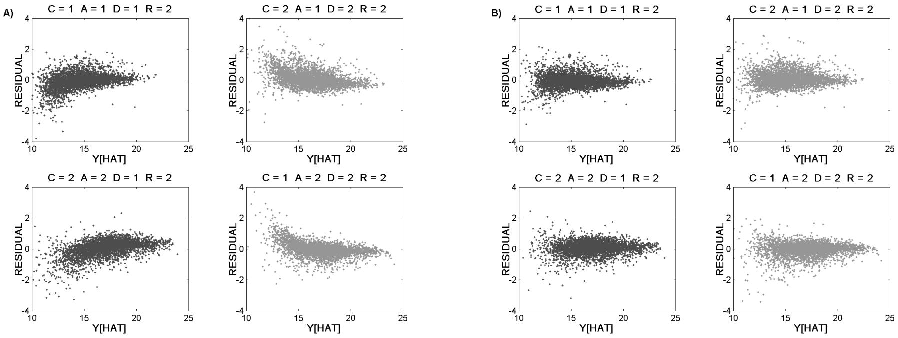

Fig. 3. Influence of non-linear effects in the data on the residual plots of the ANOVA model. Panel A: Residuals for the application of ANOVA model on the partially preprocessed data plotted separately for each array-dye combination. Data were log transformed, genes containing at least 1 zero value were removed. Panel B: Residuals for the application of ANOVA model on the data preprocessed data as outlined in section 3, plotted separately for each array-dye combination. An additional pretransformation by Lowess allowed removal of strong non-linear dye effects. A: array, R:replica, D:dye, C: condition

5. Comparison of the different methods tested

The output of the fold test, the t-test, SAM and the ANOVA-bootstrap method were compared. All methods were performed on the data preprocessed as outlined in section 3. In the discussion a distinction will be made between the models that make an explicit assumption on the distribution of the H_0 hypothesis (t-test) and those that make use of bootstrap-based procedures (ANOVA, SAM). Since the number of genes called significantly differentially expressed depends on the specific parameter setting of each method (p-value chosen as threshold, deltaslider, ANOVA significance level), parameter settings were chosen such that each method predicted approximately the same number of genes as being significant (Table 4). The t-test as described by Baldi and Long [1] was used without Bonferroni correction i.e. the calculated p-values were used to rank the genes. Genes behaving similarly under the different methods were grouped. Of the 3785 genes, 246 genes were detected by at least one of the methods tested. Results are summarized in Table 4.

In all statistical tests the ratio was used as an estimator of the differential expression (in ANOVA differential expression is estimated as a difference in GC effects, which is basically a rescaled ratio). Each gene was characterized by its average expression ratio and its p-value (as determined by the t-test). A lower p-value reflects a low variation between the replicate measurements for the ratio

Table 4. Overview of the different methods tested to detect statistically differentially expressed genes.

test	number of genes called significant
1. ANOVA 95	163
2. ANOVA 99	99
3. SAM	106
4. t-test	106
5. fold test	110

Parameters for each method were chosen as such that each method withheld approximately the same number of genes.

estimate of that gene. Therefore the p-value can be considered as a measure of the consistency of a particular measurement. This means that the better the specific characteristics of genes belonging to a group (higher differential expression level and more consistent measurements), the more reliable the prediction on the genes within that group were. Comparing the gene characteristics of the different groups allowed us to make conclusions on the performance of the different methods tested. Only 8 genes were detected by all methods (see Fig. 4, group 1) pointing towards a rather low degree of agreement between the different methods in the prediction of the differentially expressed genes. Based on the results in Table 5 the following conclusions about the performance of the different methods could be made:

Genes that were called differentially expressed merely based on a fold test showed a huge variation across the different replicate measurements. As can be seen in Table 5, in group 8 and 9 the high p-values reflected a low consistency. These genes would have been rejected by tests that take into account explicitly the within group variation (such as a t-test or SAM). Indeed, the choice of a constant arbitrary threshold implicitly assumes that the variance among replicates is the same for every gene. This is, however, not the case since the variation on the ratio, as estimator of the differential expression depends on the variation of the absolute signals that constitute the factors of the ratio. Low absolute expression values in one of the two channels (test or reference) results in unstable, often artificially high ratios. As such a fixed threshold of 2 gave rise to a high number of false positives especially in the low expression range where the signal to noise ratio is low. However, as the intensities in both the channels increase, the ratios become theoretically a more reliable estimate of the differential expression. In this region a fixed threshold of 2 might have been too stringent. Different variants of the fold test have been described hitherto that are based on additional series of filtering steps e.g. a filtering step removing all genes below a certain signal to noise level. Though likely to give better results than the fold test as described here, these fold tests make use of arbitrarily defined thresholds and are not statistically founded.

From this perspective the t-test is a better alternative to the fold test. It does not only focus on the extent to which a gene is differentially expressed but also takes

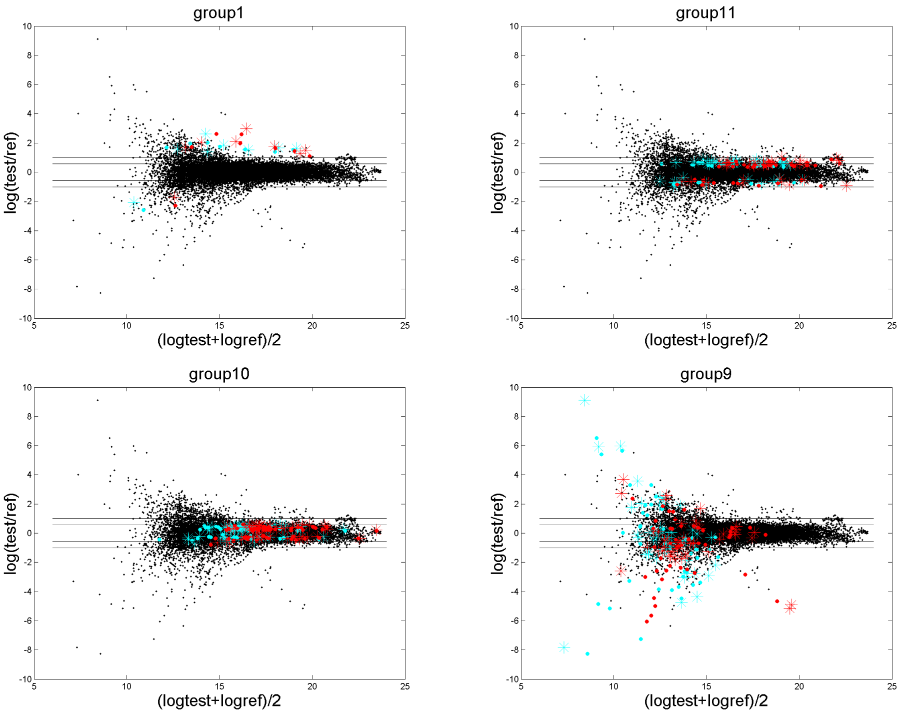

Fig. 4. Detailed representation of the genes corresponding to the groups 1, 9, 10, and 11. X-axis: average gene expression level in log scale. Y-axis: log of the ratio. Black dots: normalized expression level of all 3785 genes. Dots colored otherwise indicate the expression levels of the genes showing the profile of the corresponding groups. Each gene is represented by its 4 replicate measurements. For genes of which the expression measurements are behaving consistently, the 4 replicate measurements are represented by spots located on the same horizontal level in the plots (i.e. having a similar ratio). Cyaan: expression levels as measured on the first array, red: expression levels as measured on the second array. +: right spots; *: left spots. Dashed lines indicate the 1.5 fold and 2 fold over- and underexpression levels respectively. Group 1: genes detected by all methods. Group 10: genes only detected by the t-test. These measurements were very consistent but probably too close to zero (not differentially expressed) to be biologically relevant. Group 11: genes only detected by t-test and SAM: measurements were consistent and sufficiently different from zero (not differentially expressed) to be detected by bootstrap-based method such as SAM. Group 9: genes detected by the fold test and ANOVA-based methods only. Due to their high average expression value these genes were considered as being significant but the consistency of these genes was remarkably low. Most of the datapoints were located in the region of low average intensity. In this range the ratio becomes a poor estimator of the differential expression. Due to the heteroscedasticity in the data, the bootstrap-based confidence intervals systematically underestimated the variation at low average intensity levels and failed to reject these potentially false positives.

into account, the variation across the different measurements used to determine this average differential expression level. Indeed, genes that are retained by the t-test will per definition behave consistently (only genes with a p-value $<$ 0.005

are withheld in our tests). However, what is often observed is that the lower the signal, the more consistent genes tend to behave. This could be observed in our dataset in Fig. 4 group 10 that represent all genes retrieved by the t-test only. Although behaving consistently, these genes were almost not differentially expressed (relative expression value in log scale close to zero). The observed consistency might have been merely coincidence. These genes were indeed rejected by the resampling-based methods (SAM and ANOVA followed by bootstrap) and are probably from a biological point of view not really relevant (Table 5 cluster 10). Therefore using a t-test alone will probably result in the retrieval of consistently behaving but not differentially expressed genes. On the other hand, the t-test apparently missed a number of real differentially expressed genes in our data set. Indeed, from Table 5, group 3 and group 4 contained genes that were rejected based on the t-test but not by the resampling-based methods. These genes exceeded a 2 fold expression level. It is, however, not straightforward to judge the relevance of these genes. Although not very consistent, their replicate measurements had the same tendency either being considerably over- or underexpressed. Due to the restricted number of available measurements, the power of the t-test was probably too low to retain these genes. In contrast, the SAM method is less stringent because it makes no explicit assumptions on the H_0 distribution. Therefore, these genes though missed by the t-test, are still considered significant by SAM. Another interesting set of genes were those detected by both a t-test and resampling-based approaches. These genes were grouped in groups 11 and 5 and were only marginally but reliably down(up)regulated (Table 5). These genes probably undergo subtle changes in expression level, that are just exceeding what can be expected by coincidence (in contrast to the genes detected by the t-test only) and are probably from biological point of view most interesting.

The behavior of the ANOVA model is illustrated by Fig. 4, depicting group 9, (group 7 and group 8 suffer from the same problem but are not shown). The ANOVA-based bootstrapping approach assumes a constant confidence interval identical for all genes. The size of the confidence interval is estimated based on a fixed residual distribution of the model fit on the complete dataset. As mentioned previously, if either one of the channels measures a signal close to zero (reflected by a low average expression level) the ratio becomes an unreliable estimator of the differential expression. Indeed, in our data set, gene expression values close to zero in either one of the channels often resulted in relative high but inconsistent expression ratios (p-values of the t-test range from 0.04-0.55). For these genes the constant confidence interval was a serious underestimation of the variation on these measurements. Therefore groups 7, 8 and 9 (Table 5) most likely contained only false positives. On the other hand, for genes that were only slightly differentially expressed, the constant confidence intervals based on the constant residual variance were probably too stringent to retain these genes. This resulted in a failure of the ANOVA-based bootstrapping test to detect genes with more subtle alteration in expression level such as those present in cluster 11. For all these reasons, predictions made by ANOVA-bootstrap-based bootstrapping therefore were most unreliable.

Finally, in group 6 and 12, genes were grouped that were only detected by SAM (Table 5). These genes all behaved rather consistent (64 % of the genes have p-value lower than 0.01, Table 5) and deserve further investigation. A more detailed description on the biological relevance of the genes retrieved by our analysis in the process studied will be described elsewhere.

6. Conclusions

In this review, different approaches to identify differentially expressed genes were compared. Prior to testing each method, data were preprocessed. Ratios were used as estimates of the rate to which genes are differentially expressed. We were interested in finding out which of the methods tested performed best when only a restricted number of replicates was available since this is the situation, most often encountered in real life. It should be noted that because of the specificities of cDNA arrays (cDNA inherent sources of variation, specific preprocessing procedures), the conclusions resulting from this study can not readily be extrapolated to the analysis of other array types. From our observations the following conclusions could be made. Each of the methods differs in the required assumptions on the variance of the data and on the distribution of the residuals under the H_0 hypothesis. Therefore, the method for which the underlying assumptions are best satisfied will give the most reliable results, i.e. as it is often the case with statistical methods the reliability of the methods is dataset-dependent. The t-test could certainly be used as a more statistically founded alternative of the fold test. However, it had the tendency to retrieve many consistently behaving ratio estimates too close to 0 to be called differentially expressed. Moreover, the t-test has a rather low power because of the restricted number of replicates. Of all methods tested on our dataset, SAM clearly outperformed the other methods because the underlying assumptions were probably best satisfied.

The ANOVA-based bootstrap method clearly underperformed in identyfying differentially expressed genes. Nevertheless, from a theoretical point of view, ANOVA is most powerful to analyze microarray data. The simultaneous use of all measurements, not only to estimate the experimental noise, but also to normalize the data is a major advantage. Moreover, ANOVA can be extended to more complex designs and can take into account the specifications of each experimental setup. However, the non-linear tendencies in the data prohibit the use of ANOVA for data normalization. This problem could be partially alleviated by performing a Lowess normalization prior to the application of the ANOVA model. Even using it for measurement error estimation and bootstrap analysis only, ANOVA seemed to fail. The assumption of a constant residual variance is obviously an oversimplification viewing the non-linear trends in the data and the additivity of the error in the low expression range. At this stage this oversimplification renders the use of bootstrapping for reliable identification of differentially expressed genes impos-

Table 5. Overview of the performance of the different methods

	#genes	#under-expressed	range	#over-expressed	range	range p-value	fold test	ANOVA 99%	ANOVA 95%	SAM	t-test
reliable genes											
group 1	8	1	> -2	7	>2	<0.005	+	+	+	+	+
group 2	3	0	> -2	3	>2	<0.005	+	+	-	+	+
group 5	6	1	(-1.71)	5	(1.93;1.94)	<0.005	-	-	-	+	+
group 11	36	25	(-1.85;-1.43)	11	(1.37;1.80)	<0.005	-	-	-	+	+
genes of mediocre reliability											
group 3	18	10	>-2	8	>2	0.005-0.01: 4 0.01-0.05: 13 > 0.05:1	+	+	+	+	-
group 4	3	1	>-2	2	>2	0.005-0.01: / 0.01-0.05: 3 > 0.05:/	+	+	-	+	-
group 6	6	4	(-1.97;-1.71)	2	(1.82;1.85)	0.005-0.01: 3 0.01-0.05: 3 > 0.05: /	-	+	-	+	-
group 12	28	4	(-1.76;-1.60)	24	(1.50;1.72)	0.005-0.01: 19 0.01-0.05: 7 > 0.05:/	-	-	-	+	-
genes of low reliability											
group 10	51	19	(-1.51;-1.09)	32	(1.14;1.38)	<0.005	-	-	-	-	+
group 7	41	21	(-0.50;-0.55)	20	(1.81;1.97)	0.005-0.01:/ 0.01-0.05: 5 >0.05:36	-	+	-	-	-
group 8	33	22	>-2	11	>2	0.005-0.01:/ 0.01-0.05: 2 > 0.05:31	+	+	-	-	-
group 9	45	30	>-2	15	>2	0.005-0.01:/ 0.01-0.05: 1 > 0.05:44	+	+	+	-	-

Note: Genes were grouped as follows: a binary profile was assigned to each gene indicating whether the gene was detected (+) or not (-) by the methods tested and genes with the same binary profile were grouped. Each group of genes is characterized by a p-value range, reflecting the consistency of its replicates and the range of over- or underexpression. Based on these characteristics the performance of the distinct methods was evaluated. # genes: number of genes within a group, #number overexpressed: number of overexpressed genes within a group, #number underexpressed: number of underexpressed genes within a group, range: range of differential expression: determined as the maximal and minimal levels of over(under)expression of the individual genes belonging to that group, levels of over(under) expression are expressed as fold overexpression (positive values) or fold underexpression (negative values). range p-value: determined as the maximal and minimal p-values of the individual genes belonging to that group.

sible. Performing different transformations prior to ANOVA can help to alleviate the problem of heteroscedasticity. In order to allow bootstrap analysis despite the unequal variance in residuals, Kerr *et al.* proposed an adapted bootstrap procedure [8]. Instead of choosing for all genes a constant error, the residual was considered either gene-specific or at least intensity-specific. This approach, however, does not work when heteroscedasticity is due to a hyperposition of non-linear trends in residuals for separate combinations of major effects (Fig A. 3, e.g. all genes measured with dye 1 and array2), as was often observed in our test examples and as has similarly been noticed for other datasets [8]. Currently we are investigating different possibilities to improve the ANOVA-based technique. Because of the overall low agreement between the different methods on the predictions, combining the predictions made by the different methods gives the most reliable results and - at least partly - overcomes the specific problems of each method.

7. Software used

A userfriendly publicly available implementation of a t-test, t-test adapted for paired samples, t-test for samples with 0-level in one channel and Bayesian t-test with correction for multiple testing is available in Cyber-T software http://genomics.biochem.uci.edu/genex/cybert/ [1].

The SAM software was downloaded from http://www-stat.stanford.edu/∼tibs/SAM/ and used as a plug in in Excel [20]. The ANOVA models were implemented in Matlab 6.1 (the MathWork Inc., Natick, Mass) and are available on request (kathleen.marchal@esat.kuleuven.ac.be), httpwww.esat.kuleuven.ac.be/∼dna/BioI/.

Acknowledgments

K. Marchal is a post-doctoral researcher of the FWO; K. Engelen is research assistant of the IWT; Prof. B. De Moor is professor at the KULeuven, P. Van Hummelen is research manager of the microarray facility at VIB. This work is partially supported by: 1. IWT project: STWW-Genprom 980396; 2. Research Council KULeuven: GOA Mefisto-666; 3. FWO projects: G.0115.01; 4. DWTC (IUAP IV-02 (1996-2001) and IUAP V-22 (2002-2006); 5.IDO (IOTA Oncology, Genetic networks); 6. Flanders Interuniversity Institute of Biotechnology (VIB). The authors thank K. Coddens, R. Maes and K. Seeuws from the VIB-microarray facility for their excellent technical help and F. De Smet and G. Thijs for the useful remarks.

References

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

[2] Blohm D. H. and Guiseppi-Elie A., New developments in microarray technology, *Curr Opin Biotechnol.* **12** (2001) pp. 41-47.

[3] Brown P. O. and Botstein D., Exploring the new world of the genome with DNA microarrays, *Nat. Genet.* **21** (1999) pp. 33-37.

[4] Chen Y., Dougherty E. R. and Bittner M., Ratio-based decisions and the quantitative analysis of cDNA microarray images, *J. Biomed. Opt.* **2** (1997) pp. 364-374.

[5] Dudoit, S., Y. H. Yang, M. J. Callow and T. P. Speed., Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Technical Report #578, Stanford University. (2000) pp. 1-38.

[6] Jansen E., Petit M. M. R., Schoenmakers E. F. P. M., Ayoubi T. and Van de Ven W. J. M., High mobility group protein HMGI-C: a molecular target in solid tumor formation, *Gene Ther. Mol. Biol.* **3** (1999) pp. 387-395.

[7] Kadota K., Miki R., Bono H., Shimizu K., Okazaki Y. and Hayashizaki Y., Preprocessing implementation for microarray (PRIM): an efficient method for processing cDNA microarray data, *Physiol Genomics* **4** (2001) pp. 183-188.

[8] Kerr M. K., Afshari C. A., Bennett L., Bushel P., Martinez J., Walker N. J. and Churchill G. A., Statistical analysis of a gene expression microarray experiment with replication, *Statistica Sinica* in press (2001)

[9] Kerr M. K. and Churchill G. A., Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments, *Proc.Natl.Acad.Sci.U.S.A.* **98** (2001) pp. 8961-8965.

[10] Kerr M. K. and Churchill G. A., Experimental design for gene expression microarrays,*Biostatistics* **2** (2001) pp. 183-201.

[11] Kerr M. K., Martin M. and Churchill G. A., Analysis of variance for gene expression microarray data, *J Comput.Biol.* **7** (2000) pp. 819-837.

[12] Long A. D., Mangalam H. J., Chan B. Y., Tolleri L., Hatfield G. W. and Baldi P., Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12, *J Biol Chem.* **276** (2001) pp. 19937-19944.

[13] Marchal, K., Engelen, K., De Brabanter, J., De Moor, B. A guideline for the analysis of two sample microarray data. Technical Report 02.86, Department of electrical engineering, Catholic University of Leuven. (2002).

[14] Pan, W., A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments, Technical Report 2001-028, Division of Biostatistics, University of Minnesota. (2001).

[15] Puskas L. G., Zvara A., Hackler L. J. and Van hummelen P., RNA amplification results in reproducible microarray data with slight ratio biases, *BioTechniques* in press (2002).

[16] Rocke D. M. and Durbin B., A model for Measurement Error for Gene Expression Arrays, *J Comp Biol.* **8** (2001) pp. 557-569.

[17] Schuchhardt J., Beule D., Malik A., Wolski E., Eickhoff H., Lehrach H. and Herzel H., Normalization strategies for cDNA microarrays, *Nucleic Acids Res.* **28** (2000) pp. E47.

[18] Southern E. M., DNA microarrays. History and overview,*Methods Mol. Biol.* **170** (2001) pp. 1-15.

[19] Thomas J. G., Olson J. M., Tapscott S. J. and Zhao L. P., An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles, *Genome Res.* **11** (2001) pp. 1227-1236.

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

[21] Yang Y. H., Dudoit S., Luu P., Lin D. M., Peng V., Ngai J. and Speed T. P., Nor-

malization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res.* **30** (2002) pp. e15.

[22] Yue H., Eastman P. S., Wang B. B., Minor J., Doctolero M. H., Nuttall R. L., Stack R., Becker J. W., Montgomery J. R., Vainer M. and Johnston R., An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression, *Nucleic Acids Res.* **29** (2001) pp. E41-E41.