02-86

Marchal K., Engelen K., De Brabanter J., De Moor B., ``A guideline for the analysis of two sample microarray data", *Journal of Biological Systems*, vol. 10, no. 4, Dec. 2002, pp. 409-430., Lirias number: 182077.

A GUIDELINE FOR THE ANALYSIS OF TWO SAMPLE MICROARRAY DATA

Marchal K., Engelen K., De Brabanter J., De Moor, B.

May 29, 2002

[1]Department of Electrical Engineering, ESAT-SCD, K.U.Leuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium

1 abstract

This review describes an overview on how to analyze black/white experiments based on recent methodologies. A black/white experiment is a commonly used design to compare relative mRNA abundance between two different samples. The reason to focus on black/white experiments is multiple. Firstly, most biologists start off with such straightforward experiments to have a rough screening for differentially expressed genes in their biological system before relying on a more complex experimental design. Secondly, statistical techniques are better developed for such simple designs. The analysis flow of black/white experiments consists of two major steps: (1) data preprocessing to remove consistent sources of variation and (2) determination of genes that were significantly differentially expressed. For the data preprocessing step we described two approaches, a first one based on a slide by slide normalization/ratio approach and a second one based on ANOVA statistics. For the identification of differentially expressed genes four methods were described: 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).

2 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. The spots typically consist of 60-70 mere oligos or 0.5-2.5 kb cDNA fragments. During a microarray experiment, mRNA of a reference and induced sample 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 more extensive reviews on microarrays we refer to [3, 2, 17]).

Distinct sources of variation consistently influence microarray measurements and circumvent direct comparison of replicate measurements not assessed under exactly similar conditions (i.e. not measured on the same array, not labeled with the same dye etc.) [16, 21]. Preprocessing methods aim at removing these additional sources of variation such that for each gene the measured value reflects the mere expression level as caused by the condition tested. A first set of effects prohibiting direct comparison between measurements are the *condition and dye effects*, reflecting differences in mRNA isolation and labeling efficiencies between samples. These effects result in an overall higher measured intensity for certain conditions as compared to others. For genes expressed in an equal amount in both the reference and test sample, condition and dye effects result in a deviation of the expected ratio $test/reference$ from 1. For statistical testing (e.g. t-test see below) such deviation is undesired. The mathematical transformation that compensates for these effects is

called normalization. A second source of variation is related to the imperfections of the spotting device used to produce the array. Small variations in pin geometry, target volume and target purity cause spot-dependent variations in the amount of cDNA present on the array. Since the observed signal intensity does not only reflect differences in the mRNA population present in the sample but also the amount of spotted cDNA, direct comparison of the absolute expression levels is unreliable. This problem can be alleviated by comparison of the relative expression levels (ratio of the test and reference intensities) instead of the absolute levels. Indeed reference and test have been measured on the same spot and by dividing the measured intensities, spot effects cancel out.

When performing multiple experiments (i.e. by using more arrays), arrays are not necessarily treated identically. Differences in hybridization efficiency can result in global differences in intensities between slides, making measurements derived from different slides mutually incomparable. This effect is generally called the array effect. In this review two distinct approaches, here referred to as the ratio approach and the ANOVA (analysis of variance) approach, to preprocess and identify differentially expressed genes in a black/white experiment will be described. The ratio approach is a multistep procedure comprising log transformation, data filtering, normalization and identification of differentially expressed genes by using a test statistic. The ANOVA approach consists of a log transformation, data filtering, linearization and identification of differentially expressed genes based on bootstrap analysis. The initial preprocessing steps are similar for both approaches and will be discussed in paragraph 3 and 4.

3 Mathematical transformation of the raw data: need for a log transformation

A log transformation of the data is the initial step in the preprocessing data analysis flow. The necessity of this transformation is clear from Fig 1. In Fig. 1A the expression levels of all genes measured in the test sample were plotted against the corresponding measurements in the reference sample. By assuming that only a restricted number of genes alters its expression level, measurements of the reference and the test sample can for most genes be considered as replicates. The residual scattering as observed in Fig. 1A therefore reflects the measurement error. When considering untransformed raw data (with raw data we refer to background corrected intensity values), the increase of the residual scattering with increasing signal intensities clearly reflects the multiplicative effects (see Table 1). Multiplicative errors cause signal-dependent variance of residual scattering. This is deteriorating for most statistical tests as will be further illustrated by e.g. the underlying assumption of ANOVA models. Removal of multiplicative errors by transforming the data is therefore essential. The effects of a logarithmic data transformation are shown in Fig. 1B: residuals were constant over a long range of high signal intensities. The log transformation has increased however, the scattering of the residuals at low expression levels. This shows the presence of an additional additive error in the original data. Conclusively, the error in the data is a superposition of a multiplicative error and an additive error and log transforming the data will compensate for the multiplicative error but will increase the additive error at low expression levels. Though, an increase of the measurement error with decreasing signal intensities as present in the transformed data is intuitively plausible. Indeed low expression levels are generally considered less reliable than high levels. In most cases log transformation of raw data is advisable [1, 12].

Using a log transformation of the data has an additional advantage. By log transforming the data, statistical relevant differences in expression level are calculated based on the difference in expression level between the two channels ($log(test) - log(reference)$) (see below statistical testings). This comes down to taking the log of the ratio $test/reference$ which allows bringing levels of under- and overexpression to the same scale: values of underexpression are no longer squashed between 0 and 1.

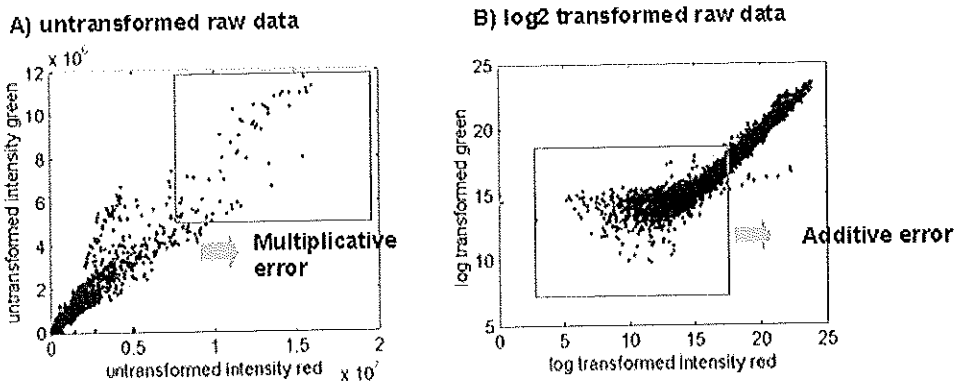


Figure 1: Illustration of the influence of log2 transformation on the multiplicative and additive errors. panel A: representation of untransformed raw data. X axis: intensity measured in the red channel, Y axis: intensity measured in the green channel. panel B: representation of log2 transformed raw data. X axis: intensity measured in the red channel (log2 value), Y axis: intensity measured in the green channel (log2 value). Assuming that only a small number of the genes will alter their expression level under the different conditions tested, for most genes the measurement in the green channel can be considered as a replica of the measurement in the red channel.

4 Filtering data

Often approaches have been used to remove unreliable measurements (e.g. discarding all expression levels that are below two standard deviations of the background intensity). Such filtering procedures depend merely on the choice of an arbitrary threshold. Since in our experiments the Cy5 and Cy3 channel displayed different sensitivities in the low expression level range (non-linear dye effect, see 6.), the approaches illustrated in this overview do avoid the use of such an arbitrary threshold [8].

Zero values result in undefined values (e.g. when dividing by zero values or taking the log of a zero value) and therefore are automatically discarded for further analysis. However, in a black/white experiment consistent zero values in one particular condition might correspond to genes differentially switched off. Therefore instead of rejecting genes with zero values, genes for which a least one measurement contained a zero value were treated separately. Genes containing a zero value often resulted in inconsistencies. Inconsistency indicates that genes seemingly underexpressed (when labeled in red) on one array are overexpressed (when labeled in green) on the other array. This points towards a strange dye effect: seemingly, if no mRNA is present neither in the reference sample nor in the test sample, the green dye sticks aspecifically to the spotted cDNA resulting in a high intensity signal. Use of a color flip therefore is mandatory to remove such false positives. This has also been observed in other studies.

5 Ratio approach

The ratio approach uses the $log_2(ratio) = (log_2(test) - log_2(reference))$ as an estimate of the relative expression. Using ratios (relative expression levels) instead of absolute expression levels allows intrinsic compensation for spot effects. After the initial data transformation and filtering steps, the ratio approach comprises data normalization and identification of differentially expressed genes by using a test statistic.

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 error of the fitted 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 s_2. It is used for hypothesis testing of normally distributed variables when the population variance s_2 is unknown, in which case the sample variance s_2 is used as an estimator of s_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. Bonferonni correction).

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

5.1 Normalization

Normalization methods as described in this paragraph aim at removing consistent condition and dye effects (see above). Although the use of spikes (control spots, external control) and housekeeping genes (genes expected not to alter their expression level under the conditions tested) have been described, global normalization is customarily used [20]. 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. Under this assumption 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. Regardless of the procedure used, all normalized log-ratios therefore will be centered around zero. The assumption of global normalization applies only to microarrays that contain a random set of genes and not to dedicated arrays. In Table 2 different procedures to perform global normalization are summarized. By performing normalization of the log-ratios, array effects are intrinsically compensated.

Table 2: Overview of recently described methods for normalization

Normalisation	Formula	Assumptions
Linear fit [4]	$\log(R) = \beta \log(G)$	Constant linear relationship [a] between red and green dye
Lowess fit [20]	$\log R/G \longrightarrow \log(R/G) - c_j(A)$ $M = \log R - \log G$ $A = \frac{\log R + \log G}{2}$ $c_j(A)$ is the lowess fit of M vs A	Nonlinear Intensity dependent relationship between red and green dye

[a] `http://afgc.stanford.edu/~finkel/talk.htm`

Linear normalization assumes a linear relationship between the measurements in both conditions (test and reference). A common choice for the constant transformation factor is the mean or median of the log intensity ratios for a given gene set. Chen *et al.* [4] use an iterative method to estimate the constant normalization factor. Alternatively the constant normalization factor can be determined by linear regression of the Cy5 signal versus the Cy3 signal. The regression factor determines the rescaling factor that should be used to transform the measurements in one channel in order to obtain an average ratio of 0 (in log scale) (i.e. shifting the center of the distribution of the log-ratios to zero).

As shown in Fig. 2, most often assuming a linear relationship between the measurements in both conditions is an oversimplification. The relationship between dyes depends on the measured intensity and therefore is not linear. These nonlinearities are most pronounced at extreme intensities (either high or low). As suggested by Yang *et al.* [5]., intensity-dependent patterns are better visualized using a plot of M ($log(test/reference)$) versus A (the average expression level in log scale) [5]. From Fig. 2 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 an intensity-dependent rescaling factor needs to be calculated. Yang *et al.* described the use of a robust scatter plot smoother, Lowess that performs locally linear fits. The results of this fit can be used to simultaneously linearize and normalize the data ([20, 5] see Table 2). After normalization using Lowess there is a clear compensation for the intensity-dependent effects i.e. linearization of the data (Fig. 2). Note, however that Lowess does not cope with the additive error at low intensities. Another drawback of the Lowess fit is that it depends on the choice of a parameter (span parameter). When chosen too small, data will be overfitted leading to decreasing signals of differentially expressed genes [9]. Shift log normalization was proposed by Kerr et al. 2001 as an alternative to cope with intensity-dependent dye effects. Prior to normalization, a scale

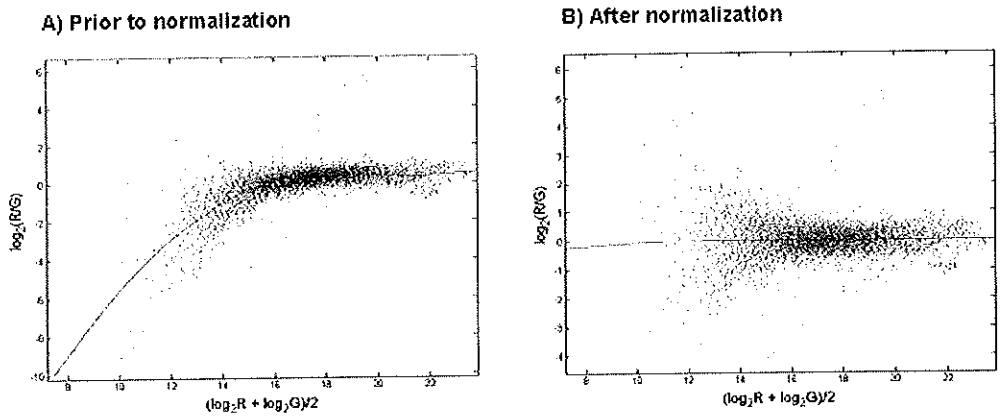


Figure 2: Illustration of the influence of an intensity-dependent normalization Panel A: representation of the log-ratio $M = log2(R/G)$ versus the mean log intensity $A = (log2(R) + log2(G))/2$. At low average intensities the ratio becomes negative indicating that the green dye is consistently more 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 $M = log2(R/G)$ versus the mean log intensity $A = (log2(R) + log2(G))/2$ after performing a normalization and linearization based on the Lowess fit. Solid line represent the new Lowess fit with f value of 0.02 on the normalized data. (R = red; G= green).

transformation of the raw fluorescent intensities is performed such that the relationship between the color channels becomes linear with additive errors that are independent of the absolute signal intensity. The idea behind this approach is that an additive error (i.e. independent of the absolute signal intensity), present in the measured intensities, creates a non-linear trend after log-transformation. Compensating for this additive error should then result in a linear relationship between both the color channels. Advantage of this method is that results are not dependent on the choice of a smoothing parameter. We tried this procedure but results were not satisfactory (data not shown)??.

The procedure as described above is a slide-dependent Lowess normalization. This means that all genes on the slide are used to calculate the intensity-dependent fit. Other approaches have been described that subdivide a slide in individual print tip groups that are separately normalized [20]. These approaches theoretically perform better in removing position-dependent within slide variations. The drawback, however, is that the number of measurements to calculate the fit is reduced, a pitfall that can be overcome by the use of ANOVA (see further).

5.2 Identification of differentially expressed genes

When preprocessed properly, consistent sources of variation have been removed and 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 described.

The fold test is a non-statistical selection procedure that makes use of an arbitrary chosen threshold. For each gene an average ratio is calculated based on the different ratio estimates ($logratio = log(test) - log(reference)$). Average ratios of which the expression ratio exceeds a threshold (usually twofold) are retained. 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. This approach is an extreme oversimplification of

the problem. A fold test indeed discards all information obtained from replicates [1].

A plethora of novel methods to calculate a test statistic and the corresponding significance level have recently been proposed provided replicates are available (see Table 3). Distinct classes of models can be discerned, differing from each other in the way the test statistic is calculated, the null hypothesis is modeled and in their underlying assumptions (Table 3). For an exhaustive comparison between the individual performances of each of these methods we refer to Pan [14] and for the technical details we refer to the individual references (see Table 3). As examples we used the method described by Baldi and Long [1] and the SAM method of Tusher *et al.* [19] because to our opinion, though quite advanced, these methods are still most intuitive and straightforward to understand for non-expert users.

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 3. 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 reference are considered as a single group. The within group variances (s_{i1}, s_{i2}) are computed based on the deviation of the different measurements of a group from their respective group means (y_{i1}, y_{i2}) (Table 3). 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 3 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 ($log(test) - log(reference)$). 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 $log(test/reference)$). 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 of a paired t-test as compared to a regular t-test. Therefore, whenever possible use of a paired t-test is highly recommended. Note that when performed on the log transformed data, the t-test approach can be considered as the counterpart of the ratio-based fold test (calculating $log(test) - log(reference) = log(test/reference)$). 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). 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 Bonferonni correction [1]. The single step adjusted p-values, as implemented in the Cyber-T software are too conservative, decreasing the power of the statistical test (ability to detect real positives). Moreover, the choice of the Bonferonni correction factor is quite arbitrary. To handle these pitfalls, other corrections for multiple testing have been proposed recently [5]. The number of replicates is usually too small for a t-test to be reliable. Long et al. 2001 suggested the use of a Bayesian t-test to cope with the low number of replicates. The population variance, used in the t-test is estimated by a posterior variance consisting of a contribution of the measured variance and a prior variance. The introduction of a prior variance avoids the need for many replicates. Indeed, based on the assumption that the variance of a gene depends on its average signal intensity only (i.e. variance is intensity dependent), it can be estimated based on the measurements of other genes with similar expression levels. This is exactly how the prior variance is calculated. The influence of the prior variance becomes more pronounced as the number of replicates decreases. This Bayesian estimate is an intriguing and intuitively easy comprehensible approach. However, it applies to t-tests on absolute levels only

Table 3: 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	$t_i = \frac{\bar{y}_{i1} - \bar{y}_{i2}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}}}$ $df = \frac{(\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2})^2}{\frac{s_{i1}^2}{n_1}^2 \frac{s_{i2}^2}{n_2}^2}$ $\frac{n_1}{n_1-1} + \frac{n_2}{n_2-1}$	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 [c, d]	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	Null distribution directly estimated	H_0estimated by EM
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

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.* [18] and that of the regular t-test may be considered equal. For small sample sizes Thomas *et al.* [18] make use of the maximum likelihood estimator of the variance. The advantage of the model of Thomas *et al.* [18] is that it does not assume a constant variance of the error term. However, to calculate their significance, Thomas *et al.* [18] 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.* [12]). 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.* [19] uses order statistics. In contrast to the other approaches, the SAM method assumes that the variances $s_{ip}^2 = \frac{(n_1-1)s_{i1}^2 + (n_2-1)s_{i2}^2}{n_1+n_2-2}$ are equally distributed and therefore uses the pooled variance as an estimator of $\sigma_{i1} = \sigma_{i2} = \sigma i$. [a] (Baldi and Long, 2001).[b] (Baldi and Long, 2001),[c] (Thomas, Olson *et al.*, 2001),[d] (Pan, 2001), [e] (Tusher, Tibshirani *et al.*, 2001)

and the extension to paired t-tests on ratios is less trivial. Indeed, the variation on the ratio is the result of the variation on the absolute measurements in each channel separately used to calculate the ratio, information that is discarded in the ratio approach. For more information on this topic we refer to Long *et al.* [13].

SAM (Significance Analysis of Microarrays) is another method for the analysis of paired or unpaired black/white experiments [19]. Instead of calculating a $t(i)$-value, SAM calculates for each gene a modified $t(i)$ value, called relative difference en referred to as $d(i)$ (see Table 3). 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. 3), ranked $d(i)$ values of the experimental dataset are compared to ranked expected $d(i)$ values. 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. 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.

6 ANOVA (analysis of variance)

ANOVA can be used as an alternative of the ratio approach. It theoretically avoids the use of ratios and the need for a high number of replicates [11]. 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 (see above) and combinations (2, 3 and 4 level combinations) of these effects are taken into account in the models (see Fig. 4). 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.

ANOVA requires at first that the data are adequately described by the linear ANOVA model and secondly that the observations are normally distributed with constant within group variances equal for all groups. If both these assumptions are satisfied, the major advantage of the ANOVA approach over the slide by slide/ratio approach consists of its ability to assess the different sources of variation across the entire experiment (i.e. the entire set of arrays). In contrast to the slide by slide/ratio approach, all measurements are combined during statistical inference. If

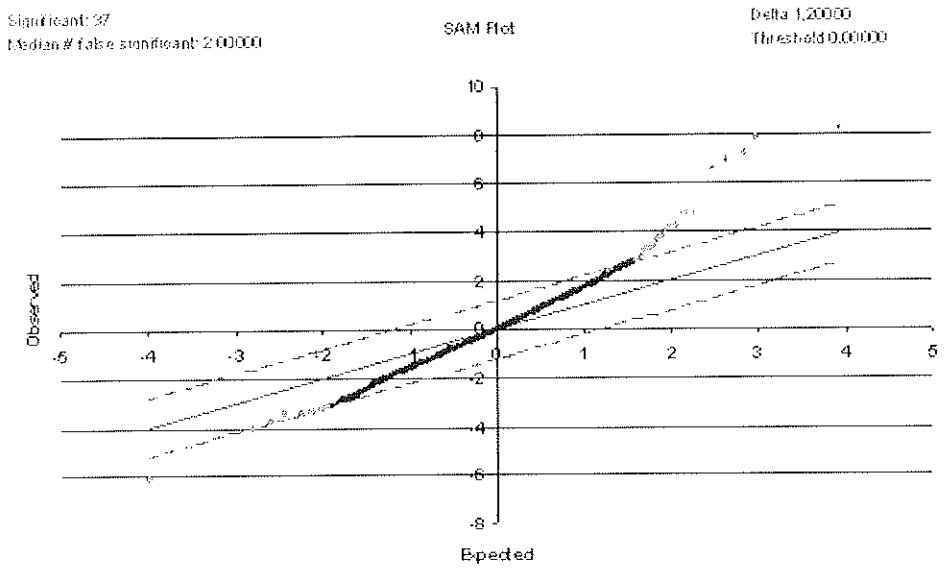
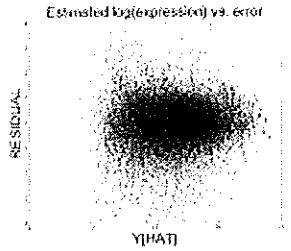


Figure 3: Result of a SAM analysis on the preprocessed dataset. For this representation the data were log transformed, genes containing at least 1 zero value were removed and data were Lowess normalized. The following parameter settings were used: paired test, permutation analysis: 1000 iterations, delta value = 1.2, threshold = 0.

both requirements mentioned above are satisfied, the model errors (as estimated by the residuals of the fit) should be 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. 4). If both assumptions are satisfied, the residual plots of the fit should be structureless. If the data can not be fit 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 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 $\hat{y}$. 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 [10]. A workable implementation of ANOVA is, however, not as straightforward as it might seem at a first glimpse. So far, interactions are included in or excluded from the models on a somehow arbitrary basis. Secondly, the assumption of a constant residual variance is obviously an oversimplification viewing the nonlinear trends in the data and the additivity of the error in the low expression range

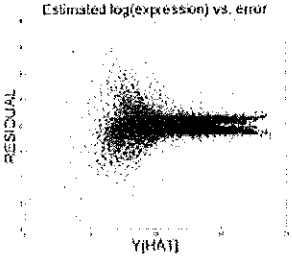
Model 1: $I_{ijl} = \mu + G_i + C_j + A_l + D_l + (GC)_{ij} + \varepsilon_{ijl}$

Source	SS	df	MS
G-effects	133126.7	3784	35.2
C-effects	4536.6	1	4536.6
A-effects	22432.4	1	22432.4
D-effects	4022.5	1	4022.5
GC-effects	1313.6	3784	0.3
Error	19210.4	22708	0.8
Corrected Total	185492.2	30279	6.1



Model 2: $I_{ijlm} = \mu + G_i + C_j + A_l + D_l + R_{m(l)} + (GC)_{ij} + \varepsilon_{ijlm}$

Source	SS	df	MS
G-effects	133126.7	3784	35.2
C-effects	4536.6	1	4536.6
A-effects	22432.4	1	22432.4
D-effects	4022.5	1	4022.5
RAG-effects	15005.5	7570	2
GC-effects	1313.6	3784	0.3
Error	4204.9	15138	0.3
Corrected Total	185492.2	30279	6.1



Model 3: $I_{ijlm} = \mu + G_i + C_j + A_l + D_l + R_{m(l)} + (AG)_{il} + (GC)_{ij} + \varepsilon_{ijlm}$

Source	SS	df	MS
G-effects	133126.7	3784	35.2
C-effects	4536.6	1	4536.6
A-effects	22432.4	1	22432.4
D-effects	4022.5	1	4022.5
AG-effects	9052.6	3784	2.4
RG-effects	2164.7	3785	0.6
GC-effects	1313.6	3784	0.3
Error	7993.1	15139	0.5
Corrected Total	185492.2	30279	6.1

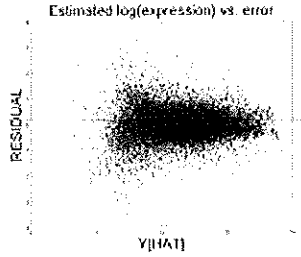


Figure 4: Results of three different ANOVA models tested on the partially preprocessed data. Data were partially preprocessed: data were log transformed, genes containing at least 1 zero value were removed, but no normalization by Lowess was performed. ANOVA models used: M: overall mean of the expression 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. Corresponding residual plots: represent for each ANOVA model the plot of the model errors (residuals) versus the estimated expression measurements ($\hat{y}_l$) values. If the assumptions underlying an ANOVA model are satisfied residual plots should be structureless. The possible causes for the observed heteroscedasticity observed in the residual plots of model 2 and 3 are explained in the text.

(for a more detailed overview on this topic we refer to Internal report [6].

7 Software availability

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 excel [19]. 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).

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] Engelen K., Marchal K., De Brabanter J., De Moor B. Critical assessment of ANOVA for microarray analysis. Internal Report 02.87, Department of Electrical Engineering, Catholic University of Leuven.

[7] 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.

[8] 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.

[9] 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)

[10] 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.

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

[12] 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.

[13] 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.

[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] 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.

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

[18] 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.

[19] 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.

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

[21] 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.