Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays

Sorin Draghici[†,*], Olga Kulaeva[§,†], Bruce Hoff[‡], Anton Petrov[‡],
Soheil Shams[‡], Michael A. Tainsky[§]
†Dept. of Computer Science, Wayne State University,
431 State Hall, Detroit, MI, 48202
‡BioDiscovery Inc., 11150 W. Olympic Blvd. Suite 1170,
Los Angeles, CA 90064
§Karmanos Cancer Institute, Wayne State University
Detroit, MI, 48201

Abstract

Motivation: A crucial step in microarray data analysis is the selection of subsets of interesting genes from the initial set of genes. In many cases, especially when comparing a specific condition to a reference, the genes of interest are those which are differentially expressed. Two common methods for gene selection are: a) selection by fold difference (at least n fold variation) and b) selection by altered ratio (at least n standard deviations away from the mean ratio). **Results:** The novel method proposed here is based on ANOVA and uses replicate spots to estimate an empirical distribution of the noise. The measured intensity range is divided in a number of intervals. A noise distribution is constructed for each such interval. Bootstrapping is used to map the desired confidence levels from the noise distribution corresponding to a given interval to the measured log ratios in that interval. If the method is applied on individual arrays having replicate spots, the method can calculate an overall width of the noise distribution which can be used as an indicator of the array quality. We compared this method with the fold change and unusual ratio method. We also discuss the relationship with an ANOVA model proposed by Churchill et.al. *In silico* experiments were performed while controlling the degree of regulation as well as the amount of noise. Such experiments show the performance of the classical methods can be very unsatisfactory. We also compared the results of the 2-fold method with the results of the noise sampling method using pre and post immortalization cell lines derived from the MDAH041 fibroblasts hybridized on Affymetrix GeneChip arrays. The 2-fold method reported 198 genes as upregulated and 493 genes as downregulated. The noise sampling method reported 98 gene upregulated and 240 genes downregulated at the 99.99% confidence level. The methods agreed on 221 genes downregulated and 66 genes upregulated. Fourteen genes from the subset of genes reported by both methods were all confirmed by Q-RT-PCR. Alternative assays on various subsets of genes on which the two methods disagreed suggested that the noise sampling method is likely to provide fewer false positives which is in good agreement with the results of the analysis. **Contact:** sod@cs.wayne.edu.

1 Introduction

Many microarray experiments aim at comparing gene expression levels in two different specimens. Obvious examples include comparing healthy vs. disease or treated vs. untreated tissues. Sample comparisons may be done using different arrays (e.g. oligonucleotide arrays) or multiple channels on the same array (e.g. cDNA arrays). In all such comparative studies, the essential problem is the identification of the genes that are differentially expressed between the two samples. Although simple in principle, this problem is complex in reality because the measured intensity values are affected by numerous sources of fluctuation and noise[36, 44]. In this context, distinguishing between genes that are truly differentially expressed and genes that are simply affected by noise

*To whom the correspondence should be addressed.
†This author has contributed substantially to this paper.

becomes a challenge.

The noise sampling selection method presented here uses a log-linear statistical model and an ANOVA approach to model the noise characteristic to a given multiple channel array. Subsequently, the noise model constructed from the data is used to find the genes that are differentially expressed for a given confidence level. The method is also applicable to replicate single channel arrays.

The report presents a number of *in silico* experiments that compare the three selection methods above using performance indicators such as specificity, sensitivity, positive predicted value and negative predicted value. Furthermore, the methods were used to study the process of cell immortalization phenomenon in the MDAH041 fibroblasts. To analyze changes in gene expression occurring during immortalization, we have compared gene expression in low-passage (11 PD) vs. high-passage (220 PD) MDAH041 fibroblasts. Low-passage, p53 heterozygous fibroblasts still have wild type (wt) p53 protein and are karyotypically unstable. The heterozygotic state of the p53 gene in pre-immortal cells was confirmed by sequencing. High-passage MDAH041 fibroblasts have lost wt p53 allele and spontaneously developed immortal phenotype [2]. MDAH041 LFS cells contain significant telomerase activity after immortalization [24]. Although in microarray analysis the hTERT gene for the protein of enzymatic subunit of telomerase was not significantly upregulated after immortalization of MDAH041 cells (1.6-fold), we found using Q-RT-PCR analysis that there was a significant increase in hTERT expression (486-fold). Telomerase mRNA (hTERT) has a very low expression level. Our experience has been that genes with low basal expression levels are difficult to quantitate accurately using microarrays alone.

1.1 Existing work

The measurements coming from microarray experiments have multiple sources of variation even after normalization [16, 33, 36]. Even if the mRNA is present in the sample tissue, there is a non-negligible probability (about 5%) that the hybridization of any single spot containing complementary DNA will not reflect the presence of the mRNA [26]. Furthermore, the probability that a single spot will provide a high signal even if the mRNA is not present seems to be even higher (about 10%) [26]. Given this, it is important to distinguish between interesting variation caused by differential gene expression and variation introduced by other causes. Intuitively, the presence of replicate spots on individual arrays should provide information about the quality of the arrays. Array replication can be used together with an ANOVA approach to calculate estimates of the errors introduced by the various sources of variance [23]. Once understood, the structure of the sources of variance can be used to improve the overall precision through a careful experiment design [22, 20]. An increase in the confidence of the results can also be obtained by applying classical randomization techniques such as bootstrapping [13, 21].

Selecting interesting genes can be done in various ways. However, two selection methods are used very widely. The first such selection method involves simply comparing the expression levels in the experiment vs. control. The genes having expression values very different in experiment vs. control are selected. Typically, a difference is considered as significant if it is at least 2 or 3 fold [9, 8, 39, 37, 40, 42]. Sometimes, this selection method is used in parallel on expression estimates provided by several techniques [30].

The second widely used selection method involves selecting the genes for which the ratio of the experiment and control values is a certain distance from the mean experiment/control ratio[38]. A number of other *ad hoc* thresholding and selection procedures have also been used. For instance, Schena et al.[34, 35] only selected genes for which the difference between the duplicate measurements did not exceed half their average. Furthermore, the genes considered as differentially regulated were those genes which exhibited at least a 2-fold change in expression. Although this seems to use the first method, it can be shown[7] that the combination of the duplicate consistency condition and the differentially regulated condition can be expressed in terms of mean and standard deviations (stdevs) and therefore it falls under the scope of the second method. A variation of this method selects those genes for which the absolute difference in the average expression intensities is much larger than the estimated standard error ($\hat{\sigma}$) computed for each gene using array replicates[7].

Although the methods above are the most widespread, more complex methods have also been used. One such method is the one way analysis of variance (ANOVA) [1, 4, 15]. Another approach is to assume normal distributions and use replicates and a maximum likelihood approach to calculate the probability of a particular gene being expressed. Subsequently, only those genes for which all replicates indicate that the gene is expressed are selected [26].

Other statistically based methods for the selection of differentially regulated genes include model based maximum likelihood estimation approaches [6, 32]. Sapir and Churchill [32] present a robust algorithm

for estimating the posterior probability of differential expression based on an orthogonal linear regression of the signals obtained from the two channels. Two hierarchical models (Gamma-Gamma and Gamma-Gamma-Bernoulli) for the two channel intensities have been proposed [**?**]. One advantage of such an approach is that the models constructed take into consideration the fact that the posterior probability of change depends on the absolute intensity level at which the gene is expressed[31]. Such intensity dependency reduces to defining some curves in the green-red plane corresponding to the two channels and selecting as differentially regulated the genes that fall outside the equiconfidence curves. A similar approach constructs an explicit statistical model for the variance vs. mean dependency. Subsequently, this model is used to stabilize the variance and calibrate and normalize the data [12, 19]. It is interesting to note that, for large intensities, the normalization proposed corresponds to the classical log transformation. However, at low intensities, a slightly more complex log transform was shown to perform better [12].

Other related techniques include using conditional density error models [27], mixed models [41], a limit fold change model [28], gene shaving techniques [14] and many others. It has also been noted that a median of ratios may be more useful than the ratio of medians as it produces a more Gaussian like distribution [5]. Another interesting approach suggests analyzing the data at the distribution level as opposed to spot level [17]. The issue of the number of replicates necessary have also been discussed in a number of papers [3, 26, 25, 29]. A number of articles compared various methods for the selection of differentially regulated genes[10, 18, 29].

2 Algorithms

2.1 The noise sampling selection method

Microarray experiments may involve multiple arrays to compare multiple samples. To account for the multiple sources of variation in a microarray experiment, Kerr and Churchill used an analysis of variance (ANOVA) approach [22, 23] and proposed the following model:

$$\begin{aligned}\log\left(y_{ijkg}\right) &= \mu + A_i + D_j + V_k + G_g + \\ &\quad (AG)_{ig} + (VG)_{kg} + \epsilon_{ijkg} \end{aligned} \quad (1)$$

In this model, μ is the overall mean signal of the array, A_i is the effect of the i^{th} array, D_j represents the effect of the j^{th} dye, V_k represents the effect of the k^{th} variety[1], G_g represents the variation of the g^{th} gene, $(AG)_{ig}$ is the effect of a particular spot on a given array, $(VG)_{kg}$ represents the interaction between the k^{th} variety and the g^{th} gene and ϵ_{ijkg} represents the error term for array i, dye j, variety k and gene g. The error is assumed to be independent and of zero mean.

We have taken another approach and the purpose of our design is two fold. Firstly, we wanted to to control the noise on a single slide and be able to use this knowledge in order to distinguish between genes that are truly differentially expressed and genes that may only appear to be so due to the measurement noise. Secondly, we wanted to be able to give the laboratory researcher some feedback about the quality of an individual array. Such feedback is essential for improving the laboratory protocols towards obtaining an improvement in the overall quality of the process. An accurate estimation of the noise on each slide may serve as such feedback allowing the researcher to understand important factors in the overall process. Our approach assumes the existence of replicate spots on the array. In order to achieve the two purposes above, we have modified the Kerr-Churchill model as follows:

$$\log R(gs) = \mu + G(g) + \epsilon(g,s) \quad (2)$$

where $\log R(gs)$ is the measured log ratio for gene g and spot s, μ is the average log ratio over the whole array, $G(g)$ is a term for the differential regulation of gene g and $\epsilon(g,s)$ is a zero-mean noise term. In the model above, one can calculate the following estimates:

$$\hat{\mu} = \frac{1}{n \cdot m} \sum_{g,s} \log\left(R\left(g,s\right)\right) \quad (3)$$

$$\widehat{G\left(g\right)} = \frac{1}{m} \sum_{g} \log\left(R\left(g,s\right)\right) - \hat{\mu} \quad (4)$$

where $\hat{\mu}$ is the estimate of the average log ratio μ, $\widehat{G(g)}$ is the estimate of the effect of gene g, m is the number of replicates and n is the number of genes. Using the estimates above, one can now calculate an estimate of the noise as follows:

$$\widehat{\epsilon(g,s)} = \log\left(R\left(g,s\right)\right) - \hat{\mu} - \widehat{G(g)} \quad (5)$$

Note that no particular shape (such as Gaussian) is assumed for the noise distribution or for the distribution of the gene expression values, which makes this approach very general.

[1] In this context a variety is a condition such as healthy or disease.

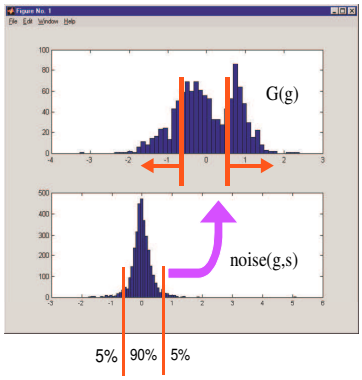

Figure 1: Using the empirically calculated noise distribution to find differentially expressed genes. Note that the width of the confidence interval will be different between the up and downregulated parts of the distribution if the noise is skewed.

The approach permits the calculation of an empirical distribution of the noise. In order to find the genes that should be reported as differentially regulated at a given confidence level, it is sufficient to calculate the deviation (from the mean) of the empirical distribution that corresponds to the given confidence level. In order to avoid the use of any particular statistical model, this is done by numerically integrating the area under the graph of the empirical distribution. The deviation thus calculated can then be mapped to the distribution of the gene expression values (see Fig. 1).

The mapping from the distribution of the noise to the distribution of the log ratios can be done by a simple scaling factor of $\frac{1}{\sqrt{m-1}}$ where m is the number of replicates if the distribution is assumed to be normal (see Supplementary Materials). However, this assumption is not necessary. A more reliable but also more computationally intensive way of mapping the confidence levels from the noise distribution to the log ratio distribution is bootstrapping. This method has the advantage that it does not make any assumptions about the characteristics of the data and takes into account any dependencies. In order to use this method, we construct artificial replicates by choosing m random samples from the noise pool and adding them to the average log ratio $\widehat{G(g)}$. This will provide a data set with the same characteristics as the original data. A number M of such data sets are created and used to calculate confidence boundaries for the log ratios.

Another issue that must be taken into consideration is the fact that the noise varies with the intensity: low intensity spots tend to have a lot more noise than high intensity ones. Confidence limits that are calculated without taking this factor into consideration draw boundaries parallel to the $y = x$ line on a scatterplot. This is not appropriate since for the genes expressed at low intensities, the average noise underestimates the real noise while for the genes expressed at high intensities, the average noise overestimates the real noise. In order to correct for this, the intensity range can be divided into bins and noise distributions can be constructed for each such bin. Thus, genes at low intensities will be affected by a correction using a larger noise variance which gene expressed at high intensity levels will user a smaller noise variance. In effect, this reduces to numerically approximating the non-linear confidence curves used in [**?**] by piece-wise linear functions. Alternatively, the noise sampling method can be combined with one of the methods described in [11] or [**?**]. Investigating such combined approaches is beyond the scope of the current paper.

The width of the empirical noise distribution (e.g. the area corresponding to 90% of the noise samples in Fig.1) will be inversely proportional with the quality of the slide (a cleaner slide means less noise, which means a narrower distribution). It is worth noting that while this approach was designed for use on individual arrays, it is not limited to such use. Thus, the model in Eq. 2 can be used to deal with replicate arrays, as well as Affymetrix arrays. In this case, the noise term $\hat{\epsilon}$ will simply stand for the experimental noise over the whole multi-array experiment.

3 Implementation

3.1 *In silico* experiments

The main problem when comparing such methods is the lack of the absolute truth. In most cases, it is a major undertaking to confirm such conclusions with biological assays for tens or hundreds of genes. In order to alleviate this problem, we have simulated some data sets in which we know the subsets of genes that are upregulated, downregulated and unregulated, respectively. Thus, having the absolute truth in this data sets allows us to assess the behavior of the various analysis approaches by controlling the amount of noise as well as the differential regulation.

The data sets were constructed as follows. A set of 1000 genes was considered as having (log) expression levels drawn from a normal distribution[6, 26, 43]. We considered the expression values to be affected

by a multiplicative noise (additive noise in log scale). The log of the noise values were drawn from another normal distribution with zero mean and different standard deviation (stdev). We considered each gene is printed in triplicate on the hypothetical array. Each replicate of a gene was affected by random noise drawn from a distribution with the same parameters. For this experiment, we divided the set of 1000 genes into three groups: upregulated (20% of the genes), downregulated (20%) and unregulated (60%). We considered that the upregulation is an effect that would add some values to the log of the expression value for each upregulated gene[23, 20, 21]. The upregulation values were drawn from another normal distribution with different parameters. We consider the downregulation to behave in a similar way except that the log expression values are reduced as opposed to increased. The data can be expressed as follows: $CG_i = values + CNOISE_i$ and $EG_i = values + ENOISE_i + REG$ where CG_i are the values measured for replicate i on the control slide, EG_i are the values measured for replicate i on the experiment slide, $values$ are the intrinsic expression values (in the control), $CNOISE_i$ are the noise values affecting replicate i on the control slide, $ENOISE_i$ are the noise values affecting replicate i on the experiment slide and REG is the effect of some genes being differentially regulated. The values of the differentially regulated genes were taken to be: $UPREG = N(meanreg, stdevreg)$ and $DOWNREG = -N(meanreg, stdevreg)$ where $N(m, s)$ is the normal distribution with mean m and stdev s. Using these values, the regulation vector was constructed as $REG = [ZERO; UPREG; DOWNREG]$ where $ZERO$ is a matrix of zero elements corresponding to the genes which are not regulated.

The basic data set used $N(0, 2)$ for $CNOISE_i$ and $ENOISE_i$, $N(6, 0.8)$ for $UPREG$ and $DOWNREG$ and $N(0, 1)$ for the noise. In all sets the expression values had zero mean. The basic data set is designed to model a clean experiment in which some genes are clearly regulated. The composition of the data for these parameter settings are shown in Fig. 2. A complete set of experiments was also done with a mean regulation of 2 stdevs (data not shown).

The following methods are reported top to bottom: 1) noise sampling and selection with 99.95 confidence (here forth 99.95), 2) selecting genes with a ratio of 2 stdevs away from the mean ratio (2 stdev), 3) selecting genes with at least 4 fold difference between the control and experiment (4 fold) and 4) 99.9 confidence (99.9). We chose to test the noise sampling methods at two different confidence levels (99.9 and 99.95) in order to check the stability of the method, ie. to see whether a small change in the chosen confidence level would produce dramatically different results. The sets of genes reported as up and downregulated were compared with the set of genes which were truly differentially expressed. The performance is calculated in terms of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), positive predicted value (PPV), negative predicted value (NPV), specificity and sensitivity.

Fig. 3 presents the results for a noise/signal ratio of 0.5. The noise sampling method gives almost perfect results (0 FN and only 2 FP) for a confidence value of 99.95% as well as 99.9%. For this noise level, the performance of the 4 fold selection method is comparable providing only 5-6 FP and 0 FN. The 2 stdev selection method gives 0 FP but **finds only 14 and 11 out of the 200 upregulated and 200 downregulated genes**, respectively. This translates into sensitivities of only 0.07 and 0.055 for up and downregulated genes respectively. When the noise/signal ratio increases to 0.75, the number of FP remains low for noise sampling both at 99.95% and 99.9% as well as for the 2 stdev method (data not shown). The FP starts to increase for the 4 fold method but the increase is not proportional to the noise increase (for a noise increase of 50%, the FP increases by approximately 600% from 5-6 to 35-32). The sensitivity of the 2 stdev method continues to remain low (which is to be expected since increasing the noise cannot increase the number of TP detected by this method).

As the noise was increased, the number of FP remained low for the noise sampling method and started to increase dramatically for the 4 fold method. The sensitivity of the 2 stdev method continued to remain low (as expected because this method gives a constant proportion of genes independent of the amount of noise or the true amount of regulation). In the noisiest case, the 4 fold method gave 139 (up) and 112 (down) FPs for 185 TPs reported in both cases. This corresponds to a PPV of approx. 60%. In other words, approximately **4 of out of every 10 genes reported as differentially expressed by the 4 fold method will not truly be so**.

This set of experiments is summarized in Fig. 4 for the upregulated genes (the results on the downregulated genes are very similar). The methods considered compare as follows. The 2 stdev selection method is characterized by a high PPV and specificity but a very low sensitivity. The specific value for the sensitivity depends on the prevalence of the differentially regulated genes so for real data, this value will be different every time. Since this method uses a threshold positioned at a fixed distance (in stdevs)

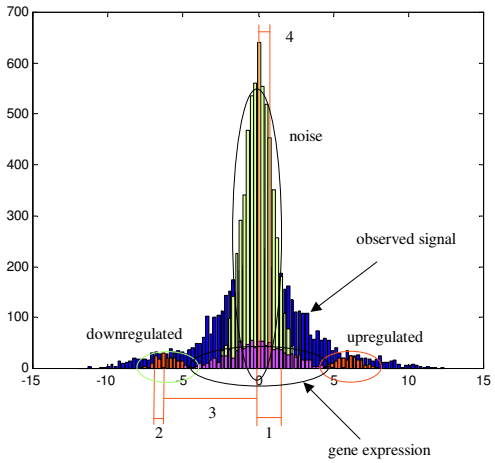

Figure 2: The composition of the observed signal (log scale). The smaller zero mean distribution in the foreground (magenta) is the distribution from which the gene expression samples were drawn. The taller zero mean distribution (yellow) is the noise distribution. The small distributions to the right and to the left (red) are the distributions which were used to modify the differentially regulated genes. The large distribution in the background (blue) is the observed overall distribution of the log ratios (log ratios of expression + noise + differential regulation). The parameters are as follows: 1 - stdev of the real gene expression distribution, 2 - stdev of the regulation distribution, 3 - mean regulation and 4 - noise stdev.

Dataset: random1 Inoise = 0.5std **Method:** 99.95 replicates

	TRUE up	TRUE not up			TRUE down	TRUE not down		
up	200	0	PPV	1	down	200	2	PPV 0.990099
not up	0	800	NPV	1	not down	0	798	NPV 1
	Sensitivity 1	Specificity 1				Sensitivity 1	Specificity 0.9975	

Dataset: random1 Inoise = 0.5std **Method:** 2 std replicates z-score

	TRUE up	TRUE not up			TRUE down	TRUE not down		
up	14	0	PPV	1	down	11	0	PPV 1
not up	186	800	NPV	0.811359	not down	189	800	NPV 0.808898
	Sensitivity 0.07	Specificity 1				Sensitivity 0.055	Specificity 1	

Dataset: random1 Inoise = 0.5std **Method:** 4 fold replicates

	TRUE up	TRUE not up			TRUE down	TRUE not down		
up	200	6	PPV	0.970874	down	200	5	PPV 0.97561
not up	0	794	NPV	1	not down	0	795	NPV 1
	Sensitivity 1	Specificity 0.9925				Sensitivity 1	Specificity 0.99375	

Dataset: random1 Inoise = 0.5std **Method:** 99.9 replicates

	TRUE up	TRUE not up			TRUE down	TRUE not down		
up	200	0	PPV	1	down	200	2	PPV 0.990099
not up	0	800	NPV	1	not down	0	798	NPV 1
	Sensitivity 1	Specificity 1				Sensitivity 1	Specificity 0.9975	

Figure 3: TP, TN, FP and FN for a mean regulation of 3 stdevs and noise/signal ratio of 0.5.

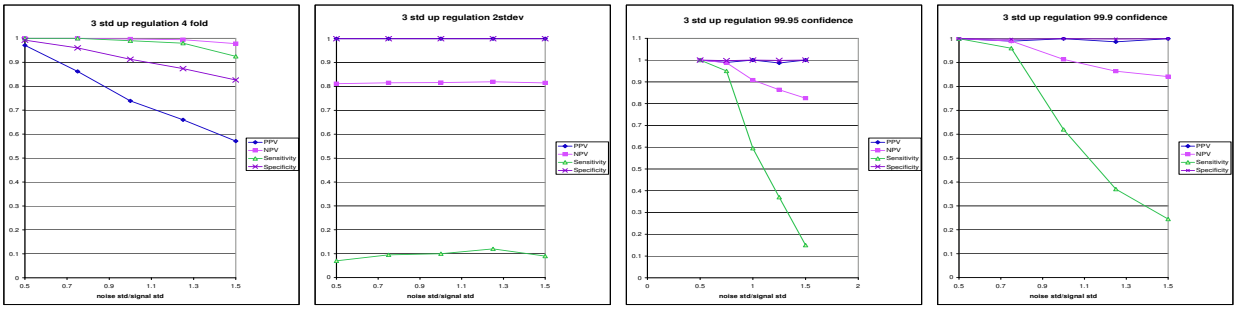

Figure 4: The performance of the three methods on the upregulated genes with a mean regulation of three stdevs (left to right: 4 fold, 2 stdev, 99.95 confidence and 99.9 confidence).

from the mean, these results remain almost constant across the whole noise range. The 4 fold selection method gives good results for a clean data set (0.5 noise/signal) but the performance decreases with the amount of noise. The performance measure with the most accentuated decline is the PPV which reflects a dramatic increase in the number of false positives. For a noisy array, this method is almost completely ineffective inasmuch about half of the genes reported are FPs. The noise sampling method provides very good results at low noise. When the noise increases, the performance measure affected most drastically is the sensitivity. This is because, as the noise increases, fewer differentially regulated genes can be confidently distinguished from non-regulated genes that are simply affected by the noise.

A comparison between the 3 methods show unequivocally that the two classical methods provide less than satisfactory results. The n fold method tends to provide many false positives. In contrast, the noise sampling method reports very few false positives across the whole noise range. This is crucially important in microarrays where differentially expressed genes need to be confirmed with other assays which are very time consuming and expensive. The other classical method, 2 stdev, provides a very low sensitivity. Note that the lowest sensitivity provided by the noise sampling method (for the noisiest data set) is still better than the best sensitivity of the 2 stdev method.

We also studied the performances of the noise sampling method as a function of the user chosen confidence level. Fig. 5 presents the PPV, NPV, specificity and sensitivity surfaces plotted for various confidence and noise levels. The data show that the specificity, PPV and NPV of the method are very high and almost constant across the range of noise and confidence values. The most variation is shown by the sensitivity surface which has lower values for the high confidence and high noise area of the domain. However, the absolute minimum of the surface is around 10% which is comparable to the best sensitivity of the 2 stdev method on these data.

3.2 Gene-chip experiments

The MDAH041 cell line was derived from primary fibroblasts obtained by skin biopsy from patient with LFS. Characterization and immortalization of these cells in vitro was as described previously[2]. Microarray experiments were performed using the Affymetrix HG-U95A array. Two mRNA preparations from immortal cells (high passage) were compared with two mRNA preparations from pre-immortal cells (low passage) using four HG-U95A chips. The purity and quality was checked for each RNA prep by electrophoresis and all RNAs were good, showing no degradation. Also, we routinely checked each sample for the 5'/3' ratios. None of the samples showed degradation. For the control vs. experiment comparisons, all four possible pairings between the two controls and the two experiments were considered.

Methods are only as good as their available implementations. We wanted to compare the fold change method as currently deployed by the users of the MAS software with the noise sampling method as deployed by GeneSight users. A sampling among our collaborators indicated that virtually all MAS users involve the gene calls (Absent/Marginal/Present) and change calls (Increased/Decreased/Not Changed/Marginally Incresed/Marginally Decreased) in their analysis. Furthermore, we knew from the simulations that the fold change method is prone to many false positives. We considered that by using the calls, the fold change method would be able to eliminate more false positives and therefore be as competitive as possible.

The upregulated genes were selected as follows. A first filter selected the genes that have a P call in

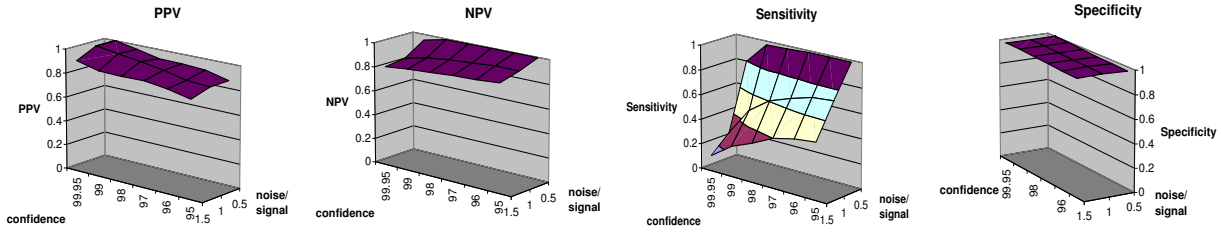

Figure 5: PPV, NPV, specificity and sensitivity as functions of confidence and noise ratio for the weakly regulated data set (mean regulation equal two stdevs).

all experiments (high passage). No call condition (A/P/M) was placed on the gene calls in controls (low passage). A second filter subsequently selected the genes with a difference call of I (increased). Finally, we sub-selected the genes with at least a two fold change between the control and experiments. This method reported 198 genes as upregulated. The downregulated genes were selected as those genes having a difference call of decreased (D) in all pairwise comparisons as well as a fold change of at least two fold. The noise sampling method was used as implemented in GeneSight 5.0 (BioDiscovery Inc.). The intensities obtained from each chip were normalized by dividing by the mean intensity. Four ratios were formed by taking all possible combinations of experiments and controls. In this case, the replication is done at the chip level and the noise model corresponds to the overall noise in the system and cannot be used to evaluate the individual quality of the arrays. We looked for genes differentially regulated with 99.99% confidence.

A comparison between the results of the 2-fold change and noise sampling methods is presented in Table 1. The comparison shows that at this level of confidence, the noise sampling method selects roughly half of the genes selected by the 2 fold method. Note that both methods report more genes as downregulated than upregulated. Although we are keenly aware of the importance of the correction for multiple experiments, no such corrections were done in this case in order to provide a more accurate comparison of the gene selection methods.

In order to confirm the results of the data analysis, we randomly selected 14 genes from the subset of genes reported as differentially regulated by both methods and proceeded to do Q-RT-PCR on them to confirm their differential regulation. All 14 genes were confirmed as differentially regulated by Q-RT-PCR. We considered a gene as confirmed if the fold change indicated by Q-RT-PCR was at least 2 fold. The comparative results between the microarray and Q-RT-PCR are presented in Fig. 6.

We also randomly selected a few genes from the subsets reported as differentially regulated by only one of the methods. Table 2 presents the comparative results obtained on eight such genes. The first two genes were reported as up-regulated by DMT while GS indicated they were below the chosen 99.99% confidence level and did not report them. The Q-RT-PCR showed that they were not differentially regulated (a fold change of less than 2 was considered unreliable in Q-RT-PCR). Therefore these genes are false positives for the DMT software and true negatives for the GS. The next set of two genes were reported as down-regulated by DMT while GS showed they were below the chosen confidence level. Q-RT-PCR showed that these genes were in fact downregulated. Thus, this group constitutes true positives for the DMT and false negatives for GS. The next two genes were reported by GS while not selected by DMT. Note that these genes were rejected by DMT based on the additional call information (A/M/P) since the fold change by itself would have indicated they are up-regulated. Both genes failed to be confirmed by Q-RT-PCR and are therefore true negatives for DMT and false positives for GS. Finally, the last group of two genes were reported by GS to be downregulated while DMT failed to identify them. Both genes were confirmed by Q-RT-PCR thus being true positives for GS and false negatives for DMT.

The results of these experiments showed that all genes reported by both methods were indeed differentially regulated as reported. However, neither method proved to be fail safe. Both methods failed to find some true positives found by the other method and both methods provided some false positives. However, the GS had the advantage that approximately 88% of the genes reported were also reported by DMT. Extrapolating from the fact that 100% of the genes reported by both packages were confirmed by Q-RT-PCR, we feel that the number of false positives reported by GS is low. This is in good agreement with the results of the *in silico* experiments in which the true number of false positives could be calculated

041LP vs 041HP	2-fold	Noise sampling @ 99.99 confidence	Common
Down regulated	493	240	212
Up regulated	198	98	66

Table 1: A comparison between the results provided by the 2-fold change and noise sampling methods.

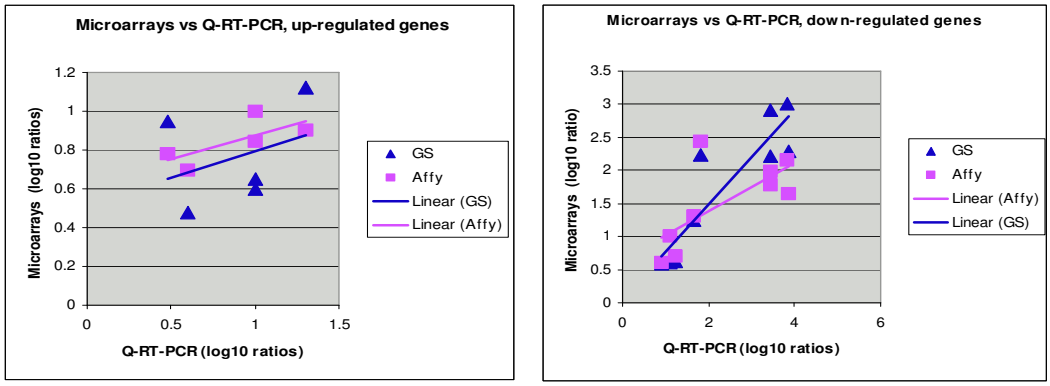

Figure 6: A comparison between the fold change obtained with microarrays and Q-RT-PCR for 5 upregulated and 9 downregulated genes. The graphs show the normalized values obtained using the different methods as well as the linear regression lines. The small difference between Affy and GS values is due to the different normalization procedures.

exactly.

4 Discussion

Our study demonstrates that two commonly used methods for the selection of differentially regulated genes have important disadvantages in terms of PPV, NPV, specificity and sensitivity. The PPV of the fold selection method degrades rapidly with the amount of noise in the data and can reach values as low as 55% (i.e. almost half of the genes reported are false positives). The unusual ratio method (selecting genes n standard deviations away from the mean log ratio) always report a constant fraction of the genes as differentially regulated. If the proportion of the genes truly differentially regulated (prevalence) is any different from this arbitrary proportion, the method will provide either a very low sensitivity or a large number of false positives.

In contrast with the classical methods, the noise sampling selection method (NS) is able to produce consistently very few false positives which translate directly into a PPV of close to 100%. Furthermore, *in silico* experiments showed the NS has a sensitivity better than that of the unusual ratio method. Even the lowest sensitivity provided by the NS (about 10% for the noisiest data set) is higher than the average sensitivity provided by the unusual ratio. While the specific values depend on the data, the unusual ratio method will always select a constant fraction of the genes without providing a specific confidence level while the NS method will provide an adaptive threshold suitable for the level of noise in the data.

The NS method presented here is a particular application of the analysis of variance (ANOVA) method. Another ANOVA variation was proposed by Kerr-Churchill (KC) [22, 23]. It can be shown that, in some very special cases (see Supplementary materials), the KC model (Eq. 1) and the NS method (Eq. 2) produce exactly the same set of genes for a given significance level. However, there are several important differences between the KC and NC approaches: i) KC requires a specific experiment design while NS can be applied in various designs (including single slide analysis); ii) KC is more complex and provides information about more factors and factor interactions while NS focuses exclusively on the differential gene expression; iii) KC obtains its final conclusions based on an F-test while NS constructs an empirical noise distribution and involves a bootstrapping processa and iv) The KC model is parametric while our model is distribution-free (non-parametric).

Gene name	Probe	Calls	Fold change DMT	Fold change GeneSight	Fold change Q-RT-PCR
Selected by DMT only: up-regulated					
SMAD5	1013_at	P,I	3.4	3.8	1.2
ERF	1242_at	P,I	2.6	2	1.8
Selected by DMT only: down-regulated					
INAE	1140_at	P,D	2.6	2.7	14
TSG-6	1372_at	P,D	5.9	2.5	11
Selected by GeneSight only: up-regulated					
ELF2G	1272_at	P, 1 in 4 NC	4.8	6	1.1
JAK1	1457_at	P, 1 in 4 NC	5.8	10	1.9
Selected by GeneSight only: down-regulated					
BCR	34679_at	A, 1 in 4 NC	9.4	13	9.4
KIAA1065 PR	36121_at	A, 1 in 4 NC	16	$-\infty$	4

Table 2: A comparison between microarray and Q-RT-PCR results on some genes on which the two packages disagreed. The calls are as follows: P - present, A - absent, M - marginal, I - increased, D - decreased, NC - not changed. The DMT filter did not consider a gene if it was NC in any of the four arrays.

In essence, all methods for gene selection are similar in that they provide a threshold for making a selection on the gene expression distribution. The advantages of the noise sampling method with respect to the classical methods of fold change and altered ratio are that i) the threshold provided is directly associated with a given confidence level; ii) the threshold changes automatically in accordance with the level of noise such that the confidence remains constant and iii) the threshold is calculated as a function of the intensity thus providing a non-linear adaptive contour of constant confidence.

5 Acknowledgments

We thank T. Twomey and S. Land (Applied Genomics Facility, Wayne State University) for the technical assistance in preparation of Affymetrix microarrays and J. Ferguson and S. Dryden (KCI, Wayne State University) for assistance with Q-RT-PCR assays and the Genomics Core of the Karmanos Cancer Institute, for support through the cancer center grant P30CA022453.

References

[1] A. Aharoni, L. C. P. Keizer, H. J. Bouwneester, Z. Sun, et al. Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays. *The Plant Cell*, 12:647–661, May 1975.

[2] F. Bischoff, S. Yim, S. Pathak, G. Grant, M. Siciliano, B. Giovanella, L. Strong, and M. Tainsky. Spontateous abnormalities in normal fibroblasts from patient with Li-Fraumeni cancer syndrome: aneuploidy and immortalization. *Cancer Research*, 24(50):7979–7984, 1990.

[3] M. A. Black and R. W. Doerge. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics*, 18(12):1609–1616, Dec 2002.

[4] A. Brazma and J. Vilo. Gene expression data analysis. *Federation of European Biochemical Societies Letters*, 480(23893):17–24, 2000.

[5] J. P. Brody, B. A. Williams, B. J. Wold, and S. R. Quake. Significance and statistical errors in the analysis of dna microarray data. *Proc. Natl. Acad. Sci.*, 99(20):12975–12978, Oct 2002.

[6] Y. Chen, E. R. Dougherty, and M. L. Bittner. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, 2(4):364–374, 1997.

[7] J.-M. Claverie. Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics*, 8(10):1821–1832, 1999. Available online at http://biosun01.biostat.jhsph.edu/ gparmigi/688/claverie1999.pdf.

[8] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

[9] J. L. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14(4):457–460, 1996.

[10] S. Draghici. Statistical intelligence: effective analysis of high-density microarray data. *Drug Discovery Today*, 7(11):S55–S63, 2002.

[11] S. Dudoit, Y. H. Yang, M. Callow, and T. Speed. Statistical models for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578, University of California, Berkeley, 2000. Available at www.stat.berkeley.edu/tech-reports/index.html.

[12] B. Durbin, J. Hardin, D. Hawkins, and D. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18(Suppl. 1):S105–s110, Jul 2002.

[13] J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783–791, 1985.

[14] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. Chan, D. Botstein, and P. Brown. 'Gene shaving' as a method for indentifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1(2):1–21, 2000.

[15] A. A. Hill, C. P. Hunter, B. T. Tsung, G. Tucker-Kellogg, and E. L. Brown. Genomic analysis of gene expression in *C. elegans*. *Science*, 290:809–812, 2000.

[16] T. M. Houts. Improved 2-color Exponential normalization for microarray analyses employing cyanine dyes. In S. Lin, editor, *Procedings*

of CAMDA 2000, "Critical Assesment of Techniques for Microarray Data Mining", December 18-19, Durham, NC, 2000. Duke University Medical Center.

[17] D. C. Hoyle, M. Rattray, R. Jupp, and A. Brass. Making sense of microarray data distributions. *Bioinformatics*, 18(4):576–584, Apr 2002.

[18] X. Huang and W. Pan. Comparing three methods for variance estimation with duplicated high-density oligonucleotide arrays. *Funct. Integr. Genomics*, 2(3):126–133, Aug 2002.

[19] W. Huber, A. Von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(Suppl. 1):S96–S104, Jul 2002.

[20] M. K. Kerr and G. A. Churchill. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Science USA*, 98(16):8961–8965, July 2001. www.jax.org/research/churchill/pubs/index.html.

[21] M. K. Kerr and G. A. Churchill. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Science USA*, 98(16):8961–8965, July 2001. www.jax.org/research/churchill/pubs/index.html.

[22] M. K. Kerr and G. A. Churchill. Statistical design and the analysis of gene expression microarray data. *Genetical Research*, 77(2):123–128, Apr 2001. www.jax.org/research/churchill/pubs/index.html.

[23] M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6):819–837, 2000.

[24] T.-A. W. S. O. Y. L. C. S. J. W. S. Lauren S Gollahon, Eliyahu Kraus and M. A. Tainsky. Telomerase activity during spontaneous immortalization of Li-Fraumeni syndrome skin fibroblasts. *Oncogene*, 17(6):709–717, Aug 1998.

[25] M. L. Lee and G. A. Whitmore. Power and sample size for dna microarray studies. *Stat. Med.*, 21(23):3543–3570, Dec 2002.

[26] M.-L. T. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci.*, 97(18):9834–9839, 2000.

[27] B. Love, D. R. Rank, S. G. Penn, D. A. Jenkins, and R. S. Thomas. A conditional density error model for the statistical analysis of microarray data. *Bioinformatics*, 18(8):1064–1072, Aug 2002.

[28] D. M. Mutch, A. Berger, R. Mansourian, A. Rytz, and M. A. Roberts. The limit fold change model: A practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics*, 3(1):17, Jun 2002. This article is available from: http://www.biomedcentral.com/1471-2105/3/17.

[29] W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4):546–554, Apr 2002.

[30] C. S. Richmond, J. D. Glasner, R. Mau, H. Jin, and F. R. Blattner. Genome-wide expression profiling in *Escherichia coli* K-12. *Nucleic Acids Research*, 27(19):3821–3835, 1999.

[31] C. J. Roberts, B. Nelson, M. J. Marton, R. Stoughton, M. R. Meyer, H. A. Bennett, Y. D. He, H. Dia, W. L. Walker, T. R. Hughes, M. Tyers, C. Boone, and S. H. Friend. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*, 287(5454):873–880, Feb 2000.

[32] M. Sapir and G. A. Churchill. Estimating the posterior probability of differential gene expression from microarray data. Technical Report http://www.jax.org/research/churchill/pubs/, Jackson Labs, Bar Harbor, ME, 2000.

[33] E. E. Schadt, L. Cheng, C. Su, and W. H. Wong. Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, 80(2):192–202, Oct 2000.

[34] M. Schena, D. Shalon, R. Davis, and P. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[35] M. Schena, D. Shalon, R. Heller, A. Chai, P. Brown, and R. Davis. Parallel human genome

analysis: microarray-based expression monitoring of 1000 genes. *Proc. National Academy of Science USA*, 93:10614–10519, 1996.

[36] J. Schuchhardt, D. Beule, E. Wolski, and H. Eickhoff. Normalization strategies for cDNA microarrays. *Nucleic Acids Research*, 28(10):e47i–e47v, May 2000.

[37] P. Sudarsanam, V. R. Iyer, P. O. Brown, and F. Winston. Whole-genome expression analysis of snf/swi mutants of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci.*, 97(7):3364–3369, 2000.

[38] H. Tao, C. Bausch, C. Richmond, F. R. Blattner, and T. Conway. Functional genomics: Expression analysis of *Escherichia coli* growing on minimal and rich media. *Journal of Bacteriology*, 181(20):6425–6440, 1999.

[39] J. J. M. ter Linde, H. Liang, R. W. Davis, H. Y. Steensma, J. P. V. Dijken, and J. T. Pronk. Genome-wide transcriptional analysis of aerobic and anaerobic chemostat cultures of *Saccharomyces cerevisiae*. *Journal of Bacteriology*, 181(24):7409–7413, 1999.

[40] A. Wellmann, C. Thieblemont, S. Pittaluga, A. Sakai, et al. Detection of differentially expressed genes in lymphomas using cDNA arrays: identification of *clusterin* as a new diagnostic marker for anaplastic large-cell lymphomas. *Blood*, 96(2):398–404, 2000.

[41] L. Wernisch, S. L. Kendall, S. Soneji, A. Wietzorrek, T. Parish, J. Hinds, P. D. Butcher, and N. G. Stoker. Analysis of whole-genome microarray replicates using mixed models. *Bioinformatics*, 19(1):53–61, Jan 2003.

[42] K. P. White, S. A. Rifkin, P. Hurban, and D. S. Hogness. Microarray analysis of Drosophila development during metamorphosis. *Science*, 286:2179–2184, 1999.

[43] B. L. Wiens. When log-normal and gamma models give different results: A case study. *The American Statistician*, 53(2):89–93, 1999.

[44] S. E. Wildsmith, G. E. Archer, A. J. Winkley, P. W. Lane, and P. J. Bugelski. Maximizing of signal derived from cDNA microarrays. *BioTechniques*, 30:202–208, 2000.

6 Supplementary material

6.1 The mapping factor

If normality is assumed, the mapping coefficient from the noise distribution histogram to the ratio histogram for normal distributions can be calculated as follows. If we use $Var(\epsilon) = \sigma^2(\epsilon)$ to denote the variance of the noise, then the variance of the noise estimate $Var(\hat{\epsilon}) = \sigma^2(\hat{\epsilon})$ can be expressed as:

$$\begin{aligned} Var(\hat{\epsilon}) &= Var(\epsilon_i - \overline{\epsilon}) = Var\left(\epsilon_i - \frac{1}{m}\sum_{j=1}^{m}\epsilon_j\right) \\ &= Var\left(\epsilon_i - \frac{1}{m}\epsilon_i - \frac{1}{m}\sum_{j\neq i}^{m}\epsilon_j\right) = \\ &= Var\left(\frac{m-1}{m}\epsilon_i - \frac{1}{m}\sum_{j\neq i}\epsilon_j\right) \end{aligned} \quad (6)$$

Now, since the two terms are independent, we can write:

$$\begin{aligned} Var(\hat{\epsilon}) &= Var\left(\frac{m-1}{m}\epsilon_i\right) + Var\left(\frac{1}{m}\sum_{j\neq i}\epsilon_j\right) \\ &= \frac{(m-1)^2}{m^2}Var(\epsilon_i) + \sum_{j\neq i}Var\left(\frac{1}{m}\epsilon_j\right) \\ &= \frac{(m-1)^2}{m^2}\sigma^2(\epsilon) + \frac{m-1}{m^2}\sigma^2(\epsilon) \\ &= \frac{m^2 - 2\cdot m + 1 + m - 1}{m^2}\sigma^2(\epsilon) \\ &= \frac{m-1}{m}\sigma^2(\epsilon) \end{aligned} \quad (7)$$

Therefore, given the variance of the empirical noise distribution, the variance of the real noise distribution can be obtained as:

$$\sigma^2(\epsilon) = \frac{m}{m-1}\sigma^2(\hat{\epsilon}) \quad (8)$$

However, each data point in the distribution of the log-ratios corresponds to the mean of m data points in the noise distribution. Therefore, the variance obtained from the noise distribution needs to be divided by m to yield the variance of the log-ratio distribution caused by the noise:

$$\sigma^2(logR(g,s)) = \frac{1}{m}\sigma^2(\epsilon) = \frac{1}{m-1}\sigma^2(\hat{\epsilon}) \quad (9)$$

which can be written:

$$\sigma(log(R(g,s)) = \frac{1}{\sqrt{m-1}}\cdot\sigma(\hat{\epsilon}) \quad (10)$$

This shows that the threshold calculated on the estimated noise histogram for a given confidence level needs to be scaled by a factor of $1/\sqrt{m-1}$ (where m is the number of replicates) before applying it on the histogram of the log-ratio values.

6.2 A special case

The noise sampling method presented here is a particular application of the analysis of variance (ANOVA) method. Another ANOVA variation was proposed by Kerr-Churchill [22, ?]. It is informative to compare the two approaches. Let us revisit Eq. 1:

$$\begin{aligned}\log(y_{ijkg}) &= \mu + A_i + D_j + V_k + G_g \\ &\quad +(AG)_{ig} + (VG)_{kg} + \epsilon_{ijkg} \quad (11)\end{aligned}$$

In this equation, i indexes the array, j indexes the dye, k indexes the mRNA and g indexes the genes. Let us assume that there are two arrays in a flip dye experiment[2] ($i = 1,2$, $j = 1,2$, $k = 1,2$, $g = 1..n$). The flip dye experiment will provide the following $4n$ values: y_{111g}, y_{122g}, y_{212g} and y_{221g}. According to the Kerr-Churchill approach, the values for μ, A_i, D_j, V_k, G_g, VG_{kg} and AG_{ig} are obtained using a least-square fit. In consequence, there will be $4n$ residuals ϵ_{ijkg}. Subsequently, the mean differential expression can be calculated as:

$$\overline{dE_g} = \frac{(y_{111g} - y_{122g}) + (y_{221g} - y_{212g})}{2} - bias \quad (12)$$

and the $4n$ residuals can be used to add error bars. In contrast, the approach proposed here goes straight to ratios which correspond to difference of log values, as follows:

$$y_{111g} = \mu + A_1 + D_1 + V_1 + G_g + VG_{1g} + AG_{1g} + \epsilon_{111g} \quad (13)$$

$$y_{122g} = \mu + A_1 + D_2 + V_2 + G_g + VG_{2g} + AG_{1g} + \epsilon_{122g} \quad (14)$$

which can be subtracted to yield:

$$\begin{aligned}y_{111g} - y_{122g} &= (D_1 - D_2) + (V_1 - V_2) + \\ &\quad (VG_{1g} - VG_{2g}) + (\epsilon_{122g} - \epsilon_{122g}) \quad (15)\end{aligned}$$

Taking $(D_1 - D_2) + (V_1 - V_2) = \mu$, $(VG_{1g} - VG_{2g}) = G(g)$ and $(\epsilon_{122g} - \epsilon_{122g}) = \epsilon(g,s)$ we obtain Eq. 2. In conclusion, the approach presented here yields only $2n$ residuals (as opposed to $4n$ residuals when ANOVA is applied on individual channels) and no estimates for μ, A_i, G_g and AG_{ig}. However, the approach does provide the differential expression terms for all genes ($VG_{1g} - VG_{2g}$) as well as noise estimates for calculating confidence intervals. Furthermore, the approach presented here is suitable for use on individual arrays which has two important consequences. Firstly, the application of this technique does not require repeating experiments[3] and thus is advantageous when either the number of arrays and/or the amount of mRNA are limited. Secondly, because the noise estimates come from a single array, they can be used to construct an overall noise distribution which will characterize the experimental quality of the array and thus provide the experimental biologist with a very useful feedback.

[2] A flip dye experiment involves two arrays. In the first array, the experimental mRNA is labeled with one dye, typically cy3, and the control mRNA is labeled with a second dye, typically cy5. In the second array, the colors are reversed: cy5 for experiment and cy3 for control.

[3] However, repeating experiments is very strongly recommended.