3-1-2005

A Statistical Framework for the Analysis of Microarray Probe-Level Data

Zhijin Wu
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Zhijin_Wu@brown.edu

Rafael A. Irizarry
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, rafa@jhu.edu

A Statistical Framework for the Analysis of Microarray Probe-Level Data

Zhijin Wu and Rafael A. Irizarry *

Abstract

Microarrays are an example of the powerful high through-put genomics tools that are revolutionizing the measurement of biological systems. In this and other technologies, a number of critical steps are required to convert the raw measures into the data relied upon by biologists and clinicians. These data manipulations, referred to as *preprocessing*, have enormous influence on the quality of the ultimate measurements and studies that rely upon them. Many researchers have previously demonstrated that the use of modern statistical methodology can substantially improve accuracy and precision of gene expression measurements, relative to ad-hoc procedures introduced by designers and manufacturers of the technology. However, further substantial improvements are possible.

Microarrays are now being used to measure diverse high genomic endpoints including yeast mutant representations, the presence of SNPs, presence of deletions/insertions, and protein binding sites by chromatin immunoprecipitation (known as ChIP-chip). In each case, the genomic units of measurement are relatively short DNA molecules referred to as *probes*. Without appropriate understanding of the bias and variance of these measurements, biological inferences based upon probe analysis will be compromised.

*Zhijin Wu is graduate student and Rafael A. Irizarry is Associate Professor of the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health (E-mail: rafa@jhu.edu). The work of Rafael Irizarry is partially funded by RO1-HG02432P01. The work of Zhijin Wu is partially funded by the Johnson and Johnson Research Foundation.

Standard operating procedure for microarray researchers is to use preprocessed data as the starting point for the statistical analyses that produce reported results. This has prevented many researchers from carefully considering their choice of preprocessing methodology. Furthermore, the fact that the preprocessing step greatly affects the stochastic properties of the final statistical summaries is ignored. In this paper we propose a statistical framework that permits the integration of preprocessing into the standard statistical analysis flow of microarray data. We demonstrate its usefulness by applying the idea in three different applications of the technology.

1 Introduction

Microarray technology measures the quantity of nucleic acid transcripts present in a biological sample referred to as the *target*. To do this we take advantage of hybridization properties of nucleic acid and use complementary molecules attached to a solid surface, referred to as *probes*. The molecules in the target are labeled and a specialized scanner is used to measure the amount of hybridization at each probe and reported as an intensity. A defining characteristic of microarray technology is that it includes thousands of probes on a relatively small surface such as a glass slide. Various manufacturers provide a large assortment of different platforms. Most manufacturers, realizing the effects of optical noise and non-specific binding, include features in their arrays to directly measure these effects. The raw or *probe-level* data are the intensities read for each of these features. In practice, various sources of variation need to be accounted for and these data are heavily manipulated before one obtains the genomic-level measurements that most biologists and clinicians use in their research. This procedure is commonly referred to as *preprocessing*.

The different platforms can be divided into two main classes that are differentiated by the type of data they produce. The *high density oligonucleotide* platforms produce one set of probe-level data per microarray with some probes designed to measure specific binding and others to measure non-specific binding. Affymetrix™GeneChip®arrays are by far the most popular product. The *two-color* platforms produce two sets of probe-level data per microarray (the red and green channels) and local background

noise levels are measured from areas in the glass slide not containing probe. No single company or academic lab dominates this market.

The most popular microarray application of both platforms is measuring genome-wide expression levels. In this application each gene is represented by one or more probes that will hybridize with the RNA transcribed from that gene. In practice, researchers using microarrays for this purpose start out with the probe-level data. However, most microarray products come equipped with software that preprocess these data into higher level measurements where each gene gets assigned one value on each array. This value is presented as the starting point for analyses that eventually lead to the results published in the scientific literature. Examples of these higher level analyses are identifying differentially expressed genes, class discovery and class prediction. In some cases, the data manipulations performed in the preprocessing step turn out to be rather complicated. Three steps typically carried out in preprocessing are:

1. Adjusting probe intensities for optical noise and/or non-specific binding. This task is referred to as *background correction*.
2. Adjusting probe intensities to assure that measurements from different arrays are comparable. This task is referred to as *normalization*.
3. When multiple probes represent a gene, summarizing the observed intensities to attain one number for each gene. We will refer to this step as *summarization*.

We will refer to these as the *three main preprocessing tasks*. For both platform classes, many different approaches have been proposed for each of these three steps resulting in competing preprocessing algorithms. We will describe some of the most popular ones in Section 2. Most of these preprocessing algorithms do not try to estimate the measures of uncertainty that accompany the resulting gene-level expression estimates. For example, normalization routines typically introduce correlation in gene-level measurements, but this correlation is rarely taken into account in the higher level analyses. Notice, that for researchers with the luxury of having a large number of replicate microarrays, this is not necessarily a problem because measures of uncertainty can be estimated from the gene-level data. However, rarely is a scientist in an academic or governmental institution in this position. Thus for most microarray experiments it becomes important to obtain as much information about the stochastic properties of the

final summary statistics from the probe-level data. By posing appropriate models for these data, any manipulation could be described statistically and bottom line results can be better understood.

Microarrays are now being used to measure diverse high genomic endpoints other than gene expression, including yeast mutant representations, the presence of Single Nucleotide Polymorphisms (SNPs), presence of deletions/insertions, and protein binding sites by chromatin immunoprecipitation (known as ChIP-chip). In each case, the genomic units of measurement continue to be the probes. Without appropriate understanding of the bias and variance of these measurements, biological inferences based upon probe analysis will be compromised. In Section 3 we present a general statistical framework, which consists of a stochastic model for probe-level data appropriate for any microarray application and procedures for quantifying answers of scientific interest that permit measuring the statistical properties introduced by the three main preprocessing tasks. In Section 4 we give examples of the usefulness of our proposal in three specific applications of microarray technology: detecting expressed genes, estimation of differential RNA expression, and identification of synthetic lethality and fitness defects in yeast mutants. Data used in the first two examples are from a high-density oligonucleotide platform and data used in the third example are from a two-color platform.

2 Previous work

Various research groups have demonstrated that statistical methodology can provide great improvements over the ad-hoc preprocessing procedures offered as defaults by the companies producing the arrays. The implementation of these methods have resulted in useful preprocessing algorithms which have already provided better bottom-line results for users of microarray expression arrays. Most of these procedures perform all three main preprocessing tasks. However, some approaches follow a step-by-step/modular approach, and others follow a global/unified approach.

In this section we describe the additive-background-multiplicative-error (addimult) model that has been implicitly or explicitly assumed to motivate most of the proce-

dures described here. We will also describe some of the most popular preprocessing methodologies. In the case of the modular approaches we will describe the different tasks in different sub-sections.

2.1 The addimult model

After target RNA samples are prepared, labeled and hybridized with arrays, these are scanned and images are produced and processed to obtain an intensity value for each probe. These intensities represent the amount of hybridization for each probe. However, part of the hybridization is non-specific and the intensities are affected by optical noise. Therefore, the observed intensities need to be adjusted to give accurate measurements of specific hybridization. In this paper, we refer to the part of the observed intensity due to optical noise and non-specific hybridization as *background noise*. Wu et al. (2004) describe experiments useful for understanding background noise behavior that empirically confirm that its effect is additive and its distribution has non-zero mean.

The component of the observed intensities related to specific binding is also affected by probe properties as well as measurement error. By using the log-scale transformation before analyzing microarray data many investigators have implicitly assumed a multiplicative measurement error model (Dudoit et al., 2002; Newton et al., 2001; Kerr et al., 2000b; Wolfinger et al., 2001). Furthermore, various groups, for example Li and Wong (2001), have demonstrated the existence of strong multiplicative probe effects on the ability to measure specific signal.

Most ad-hoc preprocessing algorithms subtract background and then take the log which arguably implies an addimult model. However, Huber et al. (2002); Cui et al. (2003); Durbin et al. (2002); Irizarry et al. (2003a) have explicitly proposed addimult models and motivated algorithms based on these. A general form of this model is simply

$$Y = B + S, \tag{1}$$

with Y the observed intensity, B the background noise component and S the specific binding component which includes multiplicative effects.

2.2 Task 1: Background adjustment

As mentioned, most microarray manufacturers include features in their products designed to directly measure probe-specific background components. This is done differently in the two different platform types. In the two-color platforms an attempt is made to measure the effect of optical noise by taking intensity measurements from parts of the glass slide with no probe. In this case no attempt is made to measure non-specific binding directly. In the high-density oligonucleotide technologies probes can be specifically designed to directly measure non-specific binding and background noise effects. Below we give more details.

2.2.1 Two-color platforms

The area between the *spots* containing probe in the two-color platforms is large enough that an intensity reading can be made near each spot. Most image processing algorithms provide background measurements, B^G and B^R, taken from these areas for both the red and green scans. Apparently, the manufacturers of image processing software assume that intensities read from the spotted areas result from target specifically binding to probe and target attaching to glass. The values B^G and B^R are considered direct measurements of the effect of the latter. Different software products provide different algorithms that determine whether an image pixel belongs to spotted area or background and summarizes the pixel intensities for each spot. The default adjustment is simply to subtract these values: $Y^G - B^G$ and $Y^R - B^G$. Yang et al. (2002b) demonstrated that the way in which the background estimate is defined can have a large effect on bottom line results. Further, these authors suggest that for some image processing algorithms better results are obtained by ignoring the background measurements.

One obvious problem with this ad-hoc subtraction adjustment is that, in general, $B \geq Y$ for a non-trivial amount of probes. Because ultimately the log ratio is used, the subtraction adjustment yields no usable data when $B \geq Y$. Noting this problem, Kooperberg et al. (2002) proposed a Bayesian approach that uses pixel summary information to define a posterior mean for S that serves as a useful background adjustment. Two advantages are that no missing data is produced and inflated variance for low expressed genes are avoided.

As far as we know, no background adjustment methodology that accounts for non-specific binding exists for this platform.

2.2.2 High density array platforms

In this platform it is common to design a probe with the intention of directly measuring the effect of background noise on each probe designed to measure specific binding. These two probe types are referred to as the mismatch (MM) and perfectmatch (PM) respectively. The naive background adjustment approach first used by the leading manufacturer of these arrays (Affymetrix) was to subtract the Y^{MM} intensity from the Y^{PM} intensity. Irizarry et al. (2003a) demonstrated that the $Y^{PM} - Y^{MM}$ transformation resulted in expression estimates with exaggerated variance and proposed a background adjustment step that ignores the MM intensities. Irizarry et al. (2003a) noticed various problems with the MM probes and developed a PM-only background adjustment. Similar to the procedure defined by Kooperberg et al. (2002) they assumed parametric models for B and S in model (1) and proposed the posterior mean $\mathrm{E}[S|Y^{PM}]$ as a background adjustment that resulted in a large reduction of variance of the gene-level estimates. Wu et al. (2004) noticed that this approach introduced a small amount of bias to attain its large gains in variance reductions. They demonstrated that this was mainly because 1) probe-specific background effects were not modeled and 2) the model used to describe B could be improved. The first problem was addressed by using probe sequence information to predict mean levels of non-specific binding effects. The second problem was addressed by improving on the model used to describe B. Wu et al. (2004) demonstrated that these modifications resulted in improved bottom line results.

2.3 Task 2: Normalization

Most experiments involve the use of multiple arrays. Therefore, it is important to remove obscuring sources of variation of non-biological origin to make data from different arrays comparable. These include different efficiencies of reverse transcription, labeling or hybridization reactions, physical problems with the arrays, reagent batch effects, and laboratory conditions. Normalization is a process for reducing this varia-

tion.

Scatter-plots of probe intensities obtained from *self-self* hybridizations (same RNA hybridized using different labels or on two separate arrays) demonstrate the existence of non-linear dependencies. Various groups (Li and Wong, 2001; Dudoit et al., 2002; Workman et al., 2002; Bolstad et al., 2003; Åstrand, 2003) have noticed these dependencies and proposed methodology for removing them. Bolstad et al. (2003) found that methods that account for these non-linear dependencies outperform methods that do not. Cui et al. (2003) demonstrated that sometimes the non-linear behavior can be explained by the fact that different arrays have different background noise levels. Furthermore, Cui et al. (2003) showed that probe intensities that are properly adjusted for background noise, no longer show these non-liner affects and that log-scale mean level adjustments are appropriate.

2.4 Task 3: Summarization

In some cases, multiple probes are used to represent one gene. In the high-density oligonucleotide platforms it is typical to define *probesets* containing 11 to 20 probes. In two-color platforms it is common to use only one probe per gene, although there are many exceptions. In both platforms, because of the multiplicative measurement error, standard operating procedure is to log the intensities before computing averages.

An important aspect of summarization is to account for the strong probe-effect that has been empirically observed. This *probe-effect* was probably discovered or predicted before the first microarray data worthy of publication were produced by Schena et al. (1995). The system producing these arrays created probes by depositing small drops, containing complementary DNA molecules and referred to as *spots*, on a glass slide. From slide to slide the size, shape, and quality of these spots varied greatly. Thus a comparison of hybridization intensities between two samples hybridized on two different slides would be dominated by an unmeasurable *probe effect*. The solution to this problem presented by Schena et al. (1995) was to hybridize the two samples being compared on the same slide. By using different fluorescent dies (red and green) on the two samples both hybridization intensities could be measured. Because the probe effect appeared to be multiplicative, it was accounted for by computing the ratio of the

two intensities.

The high-density oligonucleotide platforms are industrially produced and this permits the construction of arrays in which each probe is very similar from array to array. However, Li and Wong (2001) observed that in GeneChip arrays different probes within the same probe-set provided substantially different measurements despite the fact that they are designed to measure the same RNA transcripts. Furthermore, they noticed that these *probe effects* were consistent from array to array. This implies that some probes simply have stronger specific binding capabilities. A common approach to modeling the probe-effect is to fit a linear model to the background adjusted, normalized, and log transformed probe intensities (Irizarry et al., 2003a; Kerr et al., 2000b; Wolfinger et al., 2001; Chu et al., 2002). Notice that the procedure proposed by Li and Wong (2001) is an exception.

2.5 Unified approaches

In the modular approach, the three main preprocessing tasks are divided into a set of sequential steps. A potential disadvantage of the stepwise approach is that each step is independently optimized without considering the effect of previous or subsequent steps. This could lead to sub-optimal bottom-line results. Various investigators have used the addimult model to combine the background adjustment and normalization step into a unified estimation procedure. For example, Durbin et al. (2002), Huber et al. (2002), Geller et al. (2003) and Cui et al. (2003) use addimult models to motivate a transformation of the data that removes the dependence of the variance on the mean intensity levels. However, these procedures do not define and estimate parameters that represent quantities related to a scientific question as we wish to accomplish with our general framework.

Some methods have been proposed to estimate, or test for, differential expression as part of a more general estimation procedure that performs some of the main preprocessing tasks. For example, Kerr et al. (2000b) propose the use of ANOVA models to test for differential expression across different populations in two-color arrays. Their models include parameters to account for the need for normalization. However, the background adjustment step is performed separately. Wolfinger et al. (2001) propose

a similar model that permits some of the effects to be random. This group developed the equivalent approach for high-density oligonucleotide arrays (Chu et al., 2002). In both approaches no background adjustment is performed. Hein et al. (2005) propose a Bayesian model for high-density oligonucleotide arrays that combines background adjustment, summarization, and permits the possibility of estimating more meaningful parameters along with credibility intervals. However the normalization task is not addressed and and probe effects are not considered in the summarization.

In the next section we propose a statistical framework that will permit us to estimate parameters of interest and perform all three main preprocessing tasks in one estimation procedure. The measures of uncertainty will therefore account for the preprocessing.

3 A general statistical framework

The first step in our proposed framework is the definition of a genomic unit of interest or target DNA/RNA molecule of interest. For example, in expression arrays, the unit of interest will be an RNA transcript. Then, for each genomic unit, a set of probes, that will provide specific binding measurements for this target, are identified. Probes that provide information about non-specific binding are also identified. Finally, answers to scientific questions related to these genomic units can be quantified as summaries of the parameters in the following statistical model:

$$Y_{gij}^h = O_{gij}^h + N_{gij}^h + S_{gij}^h, \tag{2}$$

with $g = 1, \ldots, G, i = 1 \ldots, I, j = 1, \ldots, J_g$, and $h = 1, \ldots, H$. Here Y_{gij}^h is the probe intensity read from a probe of type h, for target sequence g, in array i, and probe j. The probe intensity contains three major components: optical noise O, non-specific binding N, and specific binding S. These can be further decomposed into:

$$N_{gij}^h = \exp(\mu_{gij}^h + \xi_{gij}^h) \text{ and} \tag{3}$$

$$S_{gij}^h = \exp(\nu_i^h + \theta_{gi}^h + \phi_{gij}^h + \varepsilon_{gij}^h), \text{ if } S_{gij}^h > 0. \tag{4}$$

The mean level of non-specific binding for the j-th probe of type h related to transcript g is represented by μ_{gj}^h, and a measurement error that explains differences from array

to array is described by ξ^h_{gij}. The fact that the N is strictly positive explains the need for background adjustment. If target molecule g is present, then the specific binding component S_g is formed by an array specific constant ν that explains the need for normalization, a log-scale probe effect ϕ, measurement error ε, and a quantity proportional to the amount of transcript $\exp\{\theta\}$. The index h denotes the type of probe. For example, in GeneChip arrays $h = 1, 2$ will correspond to PM or MM and in two-color arrays to Red or $Green$.

The distribution of stochastic components in (3) will depend on the platform and application. However, we conjecture that ξ follows a normal distribution in all microarray technologies. Using an experiment designed specifically to motivate a stochastic model for background noise, Wu et al. (2004) demonstrate this is the case for GeneChip arrays. Below we present evidence that the log-normal assumption applies to two-color platforms as well. If we remove outliers, the distribution of ε also appears to follow a normal distribution in many different types of data.

Notice that most parameters in model (3) are not identifiable. However, the platform designs impose certain parameter constrains that allow the parameters to be identifiable. For example, in GeneChip arrays we will assume that the probe-effects ϕ^h_{gij} does not depend on array i and that ν^h_i does not depend on the probe-type h. In two-color platforms we will assume that ϕ^h_{gij} does not depend on probe-type h. Other application specific assumptions that make the model more parsimonious will be demonstrated by example in Section 4.

The choice of which components in the model are random and which are fixed will also vary from application to application. In some applications we may assume ϕ_{gi} follows a normal distribution that does not depend on i or g as done by Wolfinger et al. (2001). In cases where we assume the variance of ε depends on g, then assuming this variance follows, for example, a gamma distribution across g will add power to the analysis. For gene-level data, these types of hierarchical models have greatly improved results in practice, see for example, Lonnstedt and Speed (2002), Smyth (2004), Gottardo et al. (2003), Pan et al. (2003), Kendziorski et al. (2003). Finally, in some applications it will be useful to model θ_{gi} with parametric models as described in Section 4.3.

Notice that some of the models motivating the unified preprocessing algorithms described in Section 2 are special cases of model (3). An example is the model proposed by Durbin et al. (2002) for two-color platforms. To obtain their model from ours we need to assume N is 0 and that O is normally distributed. Instead of estimating θ, Durbin et al. (2002) derive a transformation t for which the variance of $\Delta = t(Y^R) - t(Y^G)$ does not depend on the expectation of S^R and S^G. The difference Δ is used as a measure of relative expression on the two samples. Huber et al. (2002) follow a similar approach. Unlike Durbin et al. (2002) they explicitly include ϕ and ν in their model. As in Durbin et al. (2002) they assume N is 0 and consider O to be normally distributed. Because their procedure was originally developed for two-color arrays ϕ is absorbed into θ. Using an ad-hoc robust version of maximum likelihood estimation the parameters are estimated to derive a transformation similar to the one proposed by Durbin et al. (2002). The model described by Kerr et al. (2000b) is also a special case of ours. They assume Y has been background adjusted and, therefore, that O and N are 0. They incorporate the estimation of differential expression with the normalization step by permitting θ_{gi}^h to be constant for measurements from the same population. Hein et al. (2005) use our model as well but impose further assumptions on the distribution of the parameters. The ν and ϕ parameters are not accounted for though.

The approaches described by Durbin et al. (2002) and Huber et al. (2002) assume $O + N$ to be IID normal for each hybridization. As mentioned, empirical evidence suggests that this assumption is incorrect and that a log-normality assumption is more appropriate. This incorrect assumption has a relatively large impact on the accuracy of expression level estimate. Figure 1 compares the resulting expression estimates obtained from using the generalized-log (*glog*) and *VSN* procedures proposed by Durbin et al. (2002) and Huber et al. (2002) respectively to the *GCRMA* procedure which uses a log-normal assumption (Wu et al., 2004). We also compare these to a procedure that does no background correction at all. The figure shows averaged log (base 2) expression estimates plotted against known log (base 2) concentration levels for data from an assessment experiment (described in more detail in Section 4.1.3). Appropriate background adjustment will yield a straight line and, according to our model, no

background adjustment will yield flat local slopes for low concentrations. Notice that the procedures using the normality assumption are almost equivalent to not correcting for background.

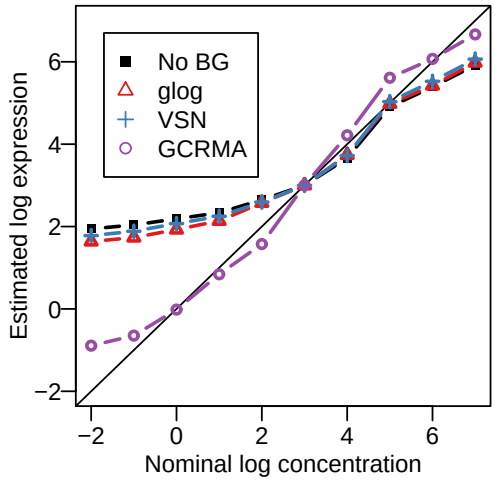

Figure 1: Log (base 2) expression estimates plotted against nominal log (base 2) concentration in picoMolar, computed with background adjustment described in the text. To make the curves comparable, the lines are shifted so that they have the same expression at log concentration 8 picoMolar (3 in log base 2).

Although the proposal of using a log-normal distribution for the background noise provides great practical improvements, the major advantage of our statistical framework is that it will permit us to describe final results of scientific interest with rigorous statistical statements. We will be able to quantify scientific questions as an exercise of optimizing estimation of a set of model parameters. With the proper model in place, fitting the model will produce useful estimates along with uncertainty measures that take into account the effects of the three main preprocessing tasks. Fitting model (3) in practice will sometimes be challenging. However, in cases where rigorous statistical procedures are impossible to convert into deliverable useful algorithms, ad-hoc versions are possible. In these cases we can still use the model assumptions to describe the statistical characteristics of the resulting data summaries.

4 Applications

In this section we describe how our framework can be adapted to give useful solutions to three important practical problems: detecting expressed genes, estimation of differential RNA expression, and identification of synthetic lethality and fitness defects in yeast mutants. In each section we briefly describe the scientific problem, the way our framework will be implemented, a dataset used to assess the performance of our approach, and results comparing our approach to standard ones.

4.1 Detecting expressed genes

4.1.1 Scientific problem

For any given target sample, it is not likely that transcripts from all genes are present. Determining which transcripts are present may be of scientific interest. The Affymetrix default software (MAS 5.0) includes an algorithm for the detection of expressed genes using GeneChip arrays. The results are summarized as *detection calls* that can take the values absent (A), marginal (M), and present (P). In our framework this can be viewed as testing the hypothesis

$$\mathrm{E}[S_{gij}^{PM}] = 0 \text{ for all } j = 1, \ldots, J_g,$$

for each gene g on each array i. Because the variability of the μ_{gij}^{PM} across g, j has been demonstrated to be very large (Wu et al., 2004), this problem is not trivial. The solution offered by Affymetrix implicitly assumes that $\mu_{gij}^{PM} = \mu_{gij}^{MM}$ and that $S_{gij}^{MM} = 0$ for all probes (Liu et al., 2002). Under these assumptions $E[Y_{gij}^{PM} - Y_{gij}^{MM}] = 0$ under the null hypothesis. A Wilcoxon test on $R_g = (Y_{gj}^{PM} - Y_{gj}^{MM})/(Y_{gj}^{PM} + Y_{gj}^{MM})$ is performed on the J_g observations to obtain a p-value[1]. The default behavior of MAS 5.0 is to assign a P, M, or A call to a p-value smaller than 0.4, between 0.4 and 0.6, and bigger than 0.6 respectively. Liu et al. (2002) demonstrate that the algorithm works relatively well in practice. However, in this section we demonstrate that our framework can be used to improve this model and to save money!

[1]MAS 5.0 software tests the null hypothesis that median(R_g)=τ versus alternative hypothesis median(R_g) $> \tau$ for a positive constant τ as opposed to 0. The default is $\tau = 0.015$.

4.1.2 Our solution

There are at least two problems with the assumptions made by Liu et al. (2002). Empirical results show strong evidence that $S_{gij}^{MM} > 0$ for many probes (Irizarry et al., 2003a) and that $\mu_{gij}^{PM} \neq \mu_{gij}^{MM}$ (Naef and Magnasco, 2003; Wu et al., 2004). Notice that if we can not use the MM probes then we need to have probe-specific information about μ_{gij}^{PM}. Wu et al. (2004) describe methodology for estimating μ_{gij}^{h} using probe sequence information. This value is estimated with enough precision to consider it known. If we treat μ_{gij}^{h} as constant, then testing the null hypothesis without using the MM probes is straight forward. Notice that GeneChip arrays include one MM for each PM, thus PM-only arrays can represent twice as many genes at the same price or represent the same genes at half-price. We therefore refer to our approach as the *half-price* procedure. Notice that the commercial arrays created by Nimblegen (Singh-Gasson et al., 1999) do not include MM probes. The half-price procedure will permit users of these arrays to perform detection calls.

4.1.3 Assessment data

To compare the two approaches we used Affymetrix's spike-in experiment (Irizarry et al., 2003b; Cope et al., 2004). In this experiment transcripts from 16 genes were artificially added or *spiked-in* to a complex cRNA target at 14 different concentrations ranging from 0 to 1024 picoMolar. Fourteen different mixtures were formed by varying the concentrations following a Latin-square design. Replicates of these mixtures were formed and hybridized to 59 GeneChip arrays of the same type. The 16 spiked-in genes were known not to be present in the original cRNA target thus if their spike-in concentration was 0, then the correct detection call is A. For all other concentrations the correct call is obviously P.

4.1.4 Results

Figure 2 demonstrates that the half-price procedure outperforms Affymetrix's default. This plot shows the p-values obtained across all 59 hybridizations for the different concentration groups. Notice that, under default parameters, the Affymetrix's detection algorithm was perfect at calling genes absent when they were truly absent, but

performed poorly when the genes were truly present. The fact that perfect results were obtained for the absent genes using a cut-off p-value of 0.04 suggests that the MAS 5.0 algorithm is the result of over-training on these genes. Our algorithm performs slightly worse at absent calls but much better at present calls.

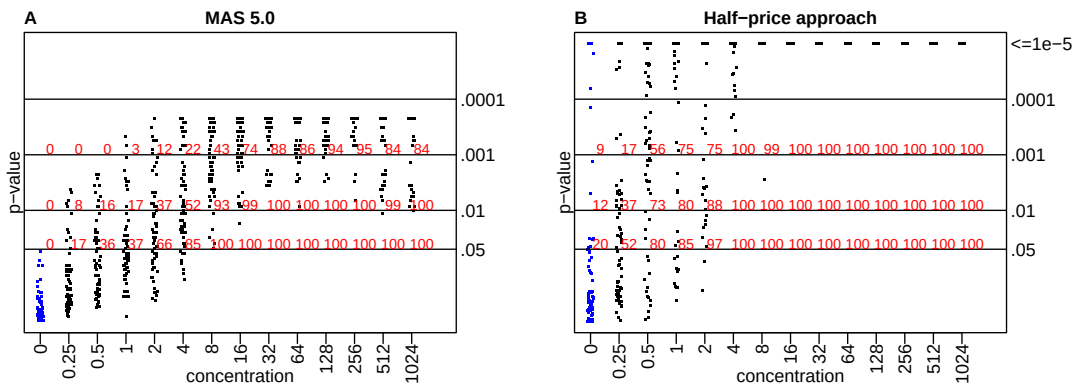

Figure 2: A) MAS 5.0 p-values, obtained from testing H_0: Gene g is absent, plotted against the spiked-in nominal concentration (picoMolar). The numbers above each horizontal line are percentages of genes with p-value exceeding these levels, i.e, the percentage of genes that would be called *present* if such a cut-off is used. B) Same as A) but using p-values from our half-price approach.

4.2 Estimating differential expression

4.2.1 Scientific problem

In this application we typically have two classes of samples (e.g. experimental and control) and in many cases we have various replicates. We are interested in measuring differential expression for each gene. Currently, the standard approach is to first preprocess the probe-level data and then use statistical procedures developed for gene-level data (Schena et al., 1996; Kerr et al., 2000b,a; Lee et al., 2000; Newton et al., 2001; Wolfinger et al., 2001; Tusher et al., 2001; Dudoit et al., 2002; Chu et al., 2002; Yang et al., 2002a; Lonnstedt and Speed, 2002) without consideration of the preprocessing algorithm. In this example we will use data from GeneChip arrays.

4.2.2 Our solution

In this context, we quantify differential expression by defining $\theta_{gi} \equiv \beta_{0,g} + \beta_{1,g}X_i$, with $X_i = 1$ if array i was hybridized to the experimental target, and $X_i = 0$ otherwise. The parameter of interest will be $\beta_{1,g}$. For this application we only use MM probes to estimate μ_{gij} and find that assuming $S^{MM}_{gij} = 0$ does not have the adverse consequences that it has in the detection call application. To reduce the number of parameters needed to represent the probe-specific mean levels we use probe sequence information as described by Wu et al. (2004). In summary, the μ^h_{gij} and ϕ^h_{gij} are assumed to be linear functions of indicator variables denoting what base (G, C, T, or A) is in each position of the probe. We assume the base effect is a smooth function of position and use splines with 5 degrees of freedom to model these functions. These assumptions reduce the number of parameters from hundreds of thousands to less than 20. See Wu et al. (2004) for more details. Other minor assumptions about across and within array correlations are described in the Appendix.

With the specifics of the model in place, we will be ready to estimate $\beta_{1,g}$. A possibility is to obtain the MLE along with a standard error for this estimate. Alternatively, we can pose a Bayesian model and obtain posterior distributions of the $\beta_{1,g}$. However, because neither of these solutions are computationally practical we used generalized estimating equations instead. Details of the implementation are in the Appendix.

Notice that in this framework expression level measurements for each array are never calculated. Instead the parameter of interest is calculated along with a measure of uncertainty that includes the effects of background adjustment, normalization, and summarization. We refer to the procedure that leads to an estimate $\hat{beta}_{1,g}$ and its standard error, as the *unified* approach.

4.2.3 Assessment data

To demonstrate the utility of the unified procedure we use data from the spike-in experiment described in Section 4.1.3. Recall that the RNA samples in this experiment were the same in all hybridizations except for the spiked-in genes. The spiked-in genes varied in concentration, within and across arrays. This implies that we can find comparisons of arrays for which only 16 genes are expected to be differentially expressed.

Furthermore, for various comparisons we had three technical triplicates in each group. We choose comparisons of two triplicates for which the expected fold-changes, for most of the spiked-in genes, was 2.

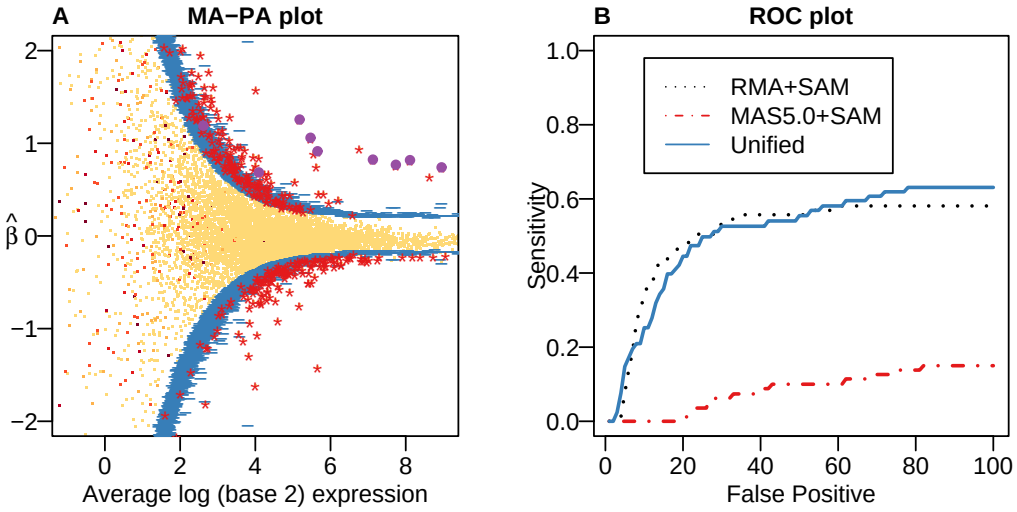

Figure 3: A) $\hat{\beta}_{1,g}$ plotted against $(\hat{\beta}_{1,g} + \hat{\beta}_{0,g})/2$. The color of each dot represents the p-value. Yellow represents low p-value (present genes), red presents large p-value (absent genes). The blue bars mark $\phi(.995)\hat{SE}$ and $\phi(.005)\hat{SE}$, where $\phi(.)$ is the cummulative density function of Normal(0,1). Spiked-in genes are labeled as big purple points. B) Averaged receiver operating curves from 14 comparisons of 2-condition with 3-replicates each from the GeneChip spike-in experiment.

4.2.4 Results

Figure 3A shows $\hat{\beta}_{1,g}$ plotted against the average log expression level (taken across the six arrays) for each gene. Notice that this provides similar information to an MA-plot. The blue bars denote point-wise critical values for rejecting the hypothesis that $\beta_{1,g} = 0$ at the 0.01 level. These critical values are computed using the fact that our estimates are asymptotically normal. Non-spiked in genes, which are known not to be differentially expressed, exceeding these bounds are shown with red stars. Spiked-in genes are shown with large purple dots. We add detection call information (described in Section 4.1) for all other points. Yellow represents low p-value (present gene), red

represents large p-value (absent gene). In this case the null hypothesis was that the genes were absent in all six hybridizations. A common approach used by biologist is to filter genes with Affymetrix produced absent calls and then compute fold change estimates. Figure 2 demonstrates that this will result in many false negatives. We propose looking at both fold change estimates and p-values in one plot such as Figure 3A. Because we are adding P/A call information to an MA-plot we refer to this as an MA-PA plot.

Figure 3A demonstrates that a procedure calling genes differentially expressed when they are outside the critical value bounds, performs rather well. Figures 3B compares our results to those obtained with the Significance Analysis of Microarray (SAM) procedure (Tusher et al., 2001). SAM is arguable the most popular procedure for detecting differentially expressed genes among biologist. This method requires expression-level data, thus we demonstrate results obtained using two popular preprocessing algorithms: RMA and MAS 5.0 (Affymetrix's default). Figure 3B shows average ROC curves for the three procedures obtained from 14 *three versus three* comparisons. To imitate real data we excluded comparisons with unrealistically large nominal fold changes and with high nominal concentrations. Specifically, only comparisons with nominal fold changes of 2 and nominal concentrations smaller than 4 picoMolar were included. The ROC curve demonstrates that our procedure performs similar to RMA/SAM in terms of detecting differentially expressed genes and much better than MAS 5.0/SAM. Figure 4 shows log fold change estimates obtained with the three procedures and demonstrates that our unfied approach provides estimates with less bias.

Figure 4 demonstrats that for low nominal conentrations the unified has more variance. However, model based standard error estimates account for this fact. Figure 5 plots our model-based standard error estimate against observed average log intensity for each gene. We also plot sample standard deviations of $\hat{\beta}_{1,g}$ for various strata of the average log intensity. The model-based standard errors are very close to the sample standard errors. Notice the strong dependence of both standard error estimates on the average log intensity. This dependence is predicted by our model. Equation (5) in the Appendix shows that the standard error is proportional to the inverse of $\mathrm{E}[S]$.

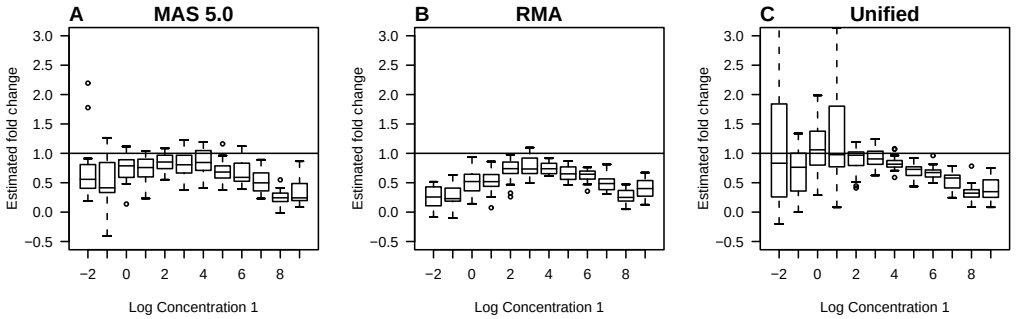

Figure 4: A) Estimated differential expression for genes with nominal fold change of 2 obtained with MAS 5.0. The x-axis shows the lower of the two nominal concentrations involved in the comparison. B) As A) but using RMA. C) As A) but with our estimate $\hat{\beta}_{1,g}$.

This provides strong evidence against the claim made by many biologist that the high variation observed for low abundance genes is a biological reality. Our calculations show that the high variation is due to the statistical manipulations needed to correct for background.

4.3 Identification of synthetic lethality and fitness defects in yeast mutants

4.3.1 Scientific problem

The Yeast Deletion strain collection was created by an international consortium of yeast geneticists (Giaever et al., 2002) and is an invaluable resource for genetics research. For each of the 6000+ genes in the yeast genome, a *mutant* yeast strain was created missing that gene. Some genes are essential and thus the mutants are not viable. Two unique DNA tags were incorporated into the genome of each mutant strain. Recently, two-channel microarray technology has been developed containing the necessary probes to detect the tags (Yuan et al., ????). Thus, Microarray hybridization can be used to measure the *representation* of each mutant in a complex mixture of many different mutants.

A new collection of mutant yeast that are missing two genes is being created. Of

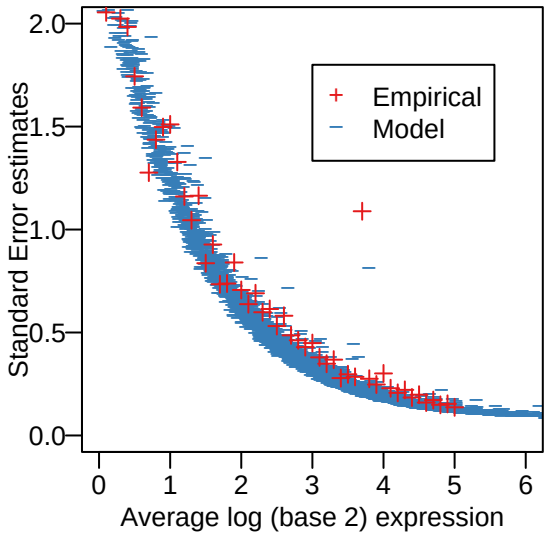

Figure 5: Model based (blue bars) and empirical standard deviations (red crosses) of $\hat{\beta}_{1,g}$ as described in the text.

interest is to find pairs of non-essential genes for which removing both causes lethality or *fitness to grow* defects. In a typical hybridization, various tags will be missing in the experimental target, these represent dead yeast, and present in the control target, these represent live yeast. Mutants with fitness defects will be under-represented in the experimental target. The tasks is to identify these tags using the microarray data.

4.3.2 Our solution

Because of financial constrains (for each of the 4000+ non-essential genes we need a hybridization), we will typically have only one array $I = 1$ per *query gene*. As mentioned, two tags are used to represent each gene; thus we have two probes per mutant, i.e. $J_i = 2$ for all i. Because the yeast mutants are either dead or alive, we will model θ_g^h with a two component mixture distribution. One component will represent the dead mutant, i.e. $S_{gj}^h = 0$ for $h = R, G$, the other will represent the live mutants. Figure 6A plots a density estimate of log intensities for both R and G channels and clearly shows both alive and dead components. This figure motivates the assumption that θ_g^h follows a normal distribution for the alive mutants. Furthermore,

the figures also supports our claim that ξ^h_{gij} is normally distributed.

Once the model is fitted under these assumptions, we are ready to provide useful summaries. To quantify the evidence for a gene being dead in the experimental target and alive in the control, we compute a likelihood ratio comparing a model where Y^R and Y^G come from different mixture components to a model where they come from the same. For mutants that appear to be alive in both cultures we can estimate the difference in representation $\log(\theta^G_g) - \log(\theta^R_g)$.

4.3.3 Assessment data

One mixture of yeast DNA was split into two halves, and into each half DNA from a few selected mutants were spiked in with known concentration ratio. The concentrations were chosen so that 1) some mutants were not represented in the experimental pool and represented in the control and 2) some mutants had known fold changes in representation when comparing both samples. The spike-in material was introduced into the hybridization mixture in three different concentration groups (high, medium, and low). See Peyser et al. (2005) for more details.

4.3.4 Results

In figure 6B we show the log likelihood ratios of mutants that had the same representation (imitating alive/alive or dead/dead) or were spiked-in only in one sample (imitating dead/alive) plotted against the naive log-ratio statistic. This Figure shows that the log likelihood ratio statistic clearly discriminates the dead/alive mutants from the rest. Various of these genes would not have been detected had we used the log ratio.

In figure 7 we show box-plots of the MLE of $\log(\theta^G_g) - \log(\theta^R_g)$ for the genes that were spiked in to be differentially represented stratified by concentration groups. In this figure we also show estimates obtained using two standard preprocessing procedures. The first is what we refer to as the default procedure which background corrects using the direct estimates of background noise and normalizes by the log ratio medians. The second is the approach proposed by Dudoit et al. (2002). The figure demonstrates our estimation procedure offers an improvement in accuracy and preci-

sion over the other two. As in the previous example, the uncertainty introduced by the background adjustment and normalization can be included with our result.

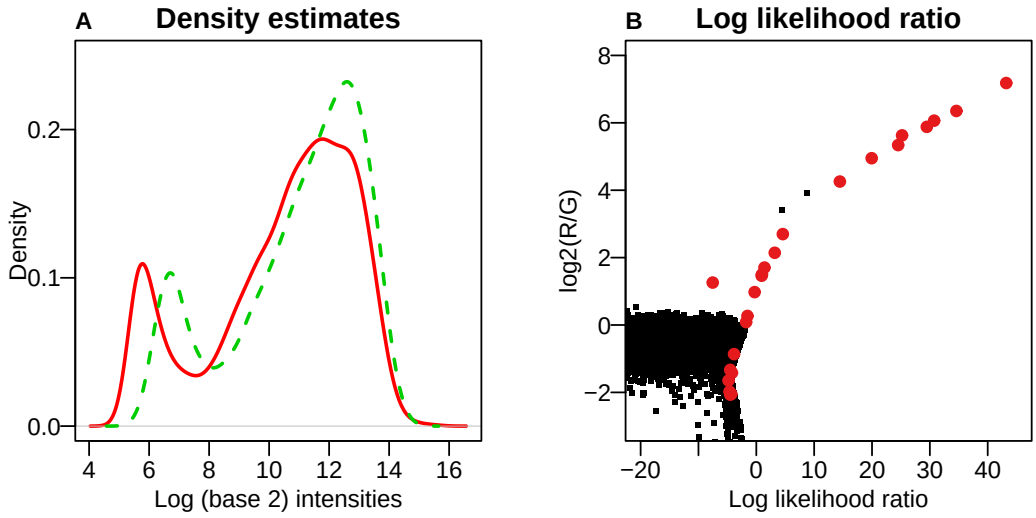

Figure 6: A) Density estimates of log intensities from the Green (dashed green line) and Red channels (solid red line). B) Log intensity ratios plotted against the log likelihood ratios described in the text.

5 Discussion

We have presented a general statistical framework useful for the analysis of microarray data. We believe it is general enough for it to be relevant in any microarray application, and targeted enough to be useful in practice. We have demonstrated the usefulness of our proposal with three examples from three very different applications and two different platforms. These examples are not intended to be final solutions to the specific problems we presented but rather examples of the usefulness of the proposed framework. An immense amount of useful work has been published in the statistics literature for both preprocessing and higher-level analyses of microarray data. Our hope is that our work will serve as a useful infrastructure that will permit the integration of these two bodies of work.

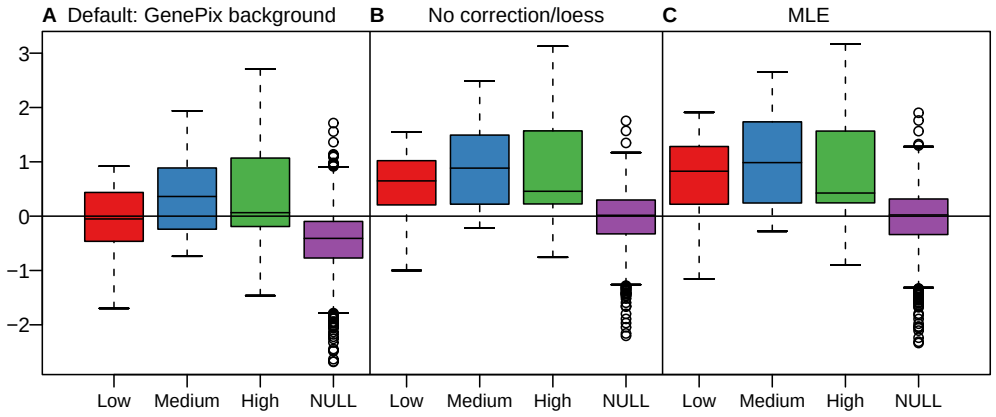

Figure 7: A) Box-plot of log fold change estimates using the default preprocessing algorithm for the low, medium, and high concentration groups. The fourth box-plot shows the log fold change estimates of genes that were not spiked-in. B) As A) but with a popular alternative preprocessing algorithm. C) As A) but with our model-based estimate.

Appendix

5.1 Generalized Estimating Equations for GeneChip spike-in experiment

To define the model we let $\mathbf{Y}_{gj} = (Y_{g1j}, Y_{g2j}, \ldots, Y_{gij}, \ldots, Y_{gIj})'$ denote the vector of PM intensities for probe j across the samples $i = 1, \ldots, I$. Similarly $\mathbf{N}_{gj}$, $\mathbf{S}_{gj}$, $\boldsymbol{\xi}_{gi}$, $\boldsymbol{\varepsilon}_{gj}$ denote the vectors for probe j across samples corresponding to the definition in model (3), $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_i, \ldots, \nu_I)'$ and $\boldsymbol{\phi}_{gj} = (\phi_{gj}, \phi_{gj}, ...)'$. We ignore the variance in optical noise and, as explained in the next section, adjust for it by subtracting the minimal intensity on each array. We write the optical-noise-adjusted intensities as

$$\begin{aligned} \mathbf{Y}_{gj} &= \mathbf{N}_{gj} + \mathbf{S}_{gj} \\ &= \exp\{\boldsymbol{\mu}_{gj} + \boldsymbol{\xi}_{gj}\} + \exp\{\boldsymbol{\nu} + \boldsymbol{\phi}_{gj} + \mathbf{X}^T\boldsymbol{\theta} + \boldsymbol{\varepsilon}_{gj}\} \end{aligned}$$

Here $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ is the vector of the log scale expression in the two conditions, $\mathbf{X}$ is the design matrix.

We compute plug-in estimators for $\boldsymbol{\mu}$, ϕ_{gj} and $\boldsymbol{\nu}$ using data from the entire array as described in the next section. We use probe sequence information to predict μ_{gj}. However the probe effects are not completely accounted for by that linear function base effect. Therefore, considering the same probes are used across arrays, we allow the measurement error to be correlated: var($\boldsymbol{\xi}_{gj}$) $= \Sigma^N$, where $\Sigma^N_{ii} = \sigma^2_N$ and $\Sigma^N_{ii'} = \rho_N \sigma^2_N$ for $i \neq i'$. var($\boldsymbol{\varepsilon}_{gj}$) $= \Sigma^S$, where $\Sigma^S_{ii} = \sigma^2_S$ and $\Sigma^S_{ii'} = \rho_S \sigma^2_S$ for $i \neq i'$. The N and S notation denote the non-specific and specific components. We assume $\boldsymbol{\xi}$ and $\boldsymbol{\varepsilon}$ follow normal distribution and the mean of variance of Y_{gj} is determined accordingly.

We then estimate $\boldsymbol{\theta}$ for each gene j using the following generalized estimating equation:

$$A_{gj}(\theta) = \frac{\partial \mathrm{E}_\theta[Y_{gj}]}{\partial \theta} V_0^{-1}.$$

The asymptotic variance of $\hat{\theta}$ is

$$D^{-1} \Omega D^{-1\prime},$$

where

$$D = \mathrm{E}\left\{ A_{gj}(\theta_0) \frac{\partial \mathrm{E}_{\theta_0}[Y_{gj}]}{\partial \theta'} \right\}$$
$$\Omega = \mathrm{E}\left\{ A_{gj}(\theta_0) \mathrm{var}_{\theta_0}(Y_{gj}) A_{gj}(\theta_0)' \right\}$$

.

We estimate D and Ω with

$$\hat{D} = \frac{1}{J} \sum A_{gj}(\hat{\theta}) \frac{\partial \mathrm{E}_{\hat{\theta}}[Y_{gj}]}{\partial \theta'},$$

$$\hat{\Omega} = \frac{1}{J} A_{gj}(\hat{\theta}) \mathrm{var}_{\hat{\theta}}(Y_{gj}) A_{gj}(\hat{\theta})'$$

.

Although the number of probes in a probe set is not very large, Figure 5 shows that the estimated variance based on this asymptotic result fits the observed variance quite well.

An interesting application of these results is that we can illustrate the relationship between the asymptotic variance and the magnitude of θ. We consider the simplest null case where $\boldsymbol{\theta} = (\theta, \theta, \ldots, \theta)'$, $\nu_i = 0$ and all probes in this probeset has the same

probe effect. We consider a simple k-control/k-treatment comparison, therefore $\mathbf{X}^T$ is the matrix

$$\begin{bmatrix} 1 & 1 & ... & 0 & 0 & ... \\ 0 & 0 & ... & 1 & 1 & ... \end{bmatrix}.$$

To simplify notations we use $\gamma_1 \equiv E[N_{gij}]$, $\gamma_2 \equiv E[S_{gij}]$, $V \equiv \text{var}(Y_{gij})$, $W \equiv \mathbf{cov}(Y_{gij}, Y_{gi'j})$. The normal assumption about ϵ implies $\gamma_2 = e^{\phi+\theta+\sigma_S^2/2}$ and $\frac{\partial \gamma_2}{\partial \theta} = \gamma_2$. Therefore

$$A = \frac{\gamma_2}{V}\mathbf{X}^T,\ D = \frac{k\gamma_2^2}{V}\mathbf{I}_{2\times 2},\ \Omega = \frac{\gamma_2^2}{V^2}\begin{bmatrix} k[V+(k-1)W] & k^2W \\ k^2W & k[V+(k-1)W] \end{bmatrix}.$$

The asymptotic variance of $\hat{\theta}$ is then

$$D^{-1}\Omega D^{-1\prime} \propto \gamma_2^{-2}\begin{bmatrix} k[V+(k-1)W] & k^2W \\ k^2W & k[V+(k-1)W] \end{bmatrix}$$

and the variance of $\hat{\theta}_1 - \hat{\theta}_2 \propto (V-W)/\gamma_2^2$. Using the normal assumption again, we have $V = \gamma_1^2(e^{\sigma_N^2}-1) + \gamma_2^2(e^{\sigma_S^2}-1)$ and $W = \gamma_1^2(e^{\rho_N\sigma_N^2}-1) + \gamma_2^2(e^{\rho_S\sigma_S^2}-1)$. This implies

$$\text{var}(\hat{\theta}_1 - \hat{\theta}_2) \propto \frac{\gamma_1^2(e^{\sigma_N^2} - e^{\rho_N\sigma_N^2}) + \gamma_2^2(e^{\sigma_S^2} - e^{\rho_S\sigma_S^2})}{\gamma_2^2}, \tag{5}$$

which predicts that the variance of estimated differential expression converges to a constant as expression levels increases (γ_2 increases) and is approximately proportional to $1/S^2$ when S is small.

5.2 Ad-hoc plug-in estimates

For our example, we assume O_{gij} is constant and form an estimate $\hat{O}$, using the minimum observed intensity on each array. We do this because the variance of O_{gij} is negligible compared to the variance of N_{gij} (Wu et al., 2004). To estimate μ_{gij}, probe *affinities* α_{gj} are computed using probe sequence as described in Wu and Irizarry (2004). We assume that μ_{gij} is a smooth function h of these affinities, i.e. $\mu_{gj} = h(\alpha_{gj})$, and estimate μ_{gij} through estimating h. Specifically, a loess curve is fit to the $\log(Y^{MM} - \hat{O})$ versus α^{MM} scatter plot to obtain $\hat{h}$. The μ_{gij} are then estimated as $\hat{h}_i(\alpha_{gj}^{PM})$. The residuals from the loess fit are used to estimate the variance of ξ,

σ^2_N. To estimate the correlation coefficient ρ_N, we identify a subset of probes with $\log(Y^{PM} - \hat{O})$ less than the corresponding $\hat{\mu}$. The target mRNAs of these probes are likely to be absent and $\log(Y^{PM} - \hat{O}) \approx N$. We obtain sample variance of each probe across arrays and use $\hat{\sigma}^2_{N0}$ to denote the mean of these variances. ρ_N is calculated as $(\hat{\sigma}^2_N - \hat{\sigma}^2_{N0})/\hat{\sigma}^2_N$.

To estimate Σ^S we first identify a subset of probe sets with high expression level such that $\log(PM) \approx \log(S)$. Within each probe set we estimate the sample variance of $\log(PM)$ and use the mean as the estimate of σ^2_S. To estimate ρ_S we use a similar approach as for ρ_N: from a subset of probes with strong signals, we obtain sample variance of each probe across arrays and set $\hat{\sigma}^2_{S0}$ as the mean of those variances. ρ_S is calculated as $(\hat{\sigma}^2_S - \hat{\sigma}^2_{S0})/\hat{\sigma}^2_S$.

References

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

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

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

Cope, L., Irizarry, R., Jaffee, H., Wu, Z., and Speed, T. (2004). A benchmark for Affymetrix Genechip expression measures. *Bioinformatics* **20**, 323–331.

Cui, X., Kerr, M. K., and Churchill, G. A. (2003). Transformations for cDNA microarray data. *Statistical Applications in Genetics and Molecular Biology* **2**(1), Article 4.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**(1), 111–139.

Durbin, B. P., Hardin, J. S., Hawkins, D. M., and Rocke, D. M. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* **18**(Suppl. 1), S105–S110.

Geller, S. C., Gregg, J. P., Hagerman, P., and Rocke, D. M. (2003). Transformation and normalization of oligonucleotide microarray data. *Bioinformatics* **19**(14), 1817–1823.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A. P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K. D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kotter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S., Revuelta, J., Roberts, C., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C. Y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W., and Johnston, M. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**(6896), 387–91.

Gottardo, R., Pannucci, J. A., Kuske, C. R., and Brettin, T. (2003). Statistical analysis of microarray data: a bayesian approach. *Biostatistics* **4**, 597–620.

Hein, A.-M., Richardson, S., Causton, H. C., Ambler, G. K., and Green, P. J. (2005). BGX: a fully bayesian gene expression index for Affymetrix GeneChip data. *Biostatistics* **(in press)**.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **1**, 1:9.

Irizarry, R. A., B. Hobbs, F. C., Beaxer-Barclay, Y., Antonellis, K., Scherf, U., and

Speed, T. (2003a). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003b). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research* **31**.

Kendziorski, C. M., Newton, M. A., Lan, H., and Gould, M. (2003). On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**(24), 3899–3914.

Kerr, M., Afshari, C., Bennett, L., Bushel, P., Martinez, J., Walker, N., and Churchill, G. (2000a). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* 203.

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

Kooperberg, C., Fazzio, T. G., Delrow, J. J., and Tsukiyama, T. (2002). Improved background correction for spotted DNA microarrays. *Journal of Computational Biology* **9**(1), 55–66.

Lee, M.-L. T., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences U S A* **97**(18), 9834–9839.

Li, C. and Wong, W. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science U S A* **98**, 31–36.

Liu, W., Mei, R., Di, X., Ryder, T. B., Hubbell, E., Dee, S., Webster, T. A., Harrington, C. A., h. Ho, M., Baid, J., and Smeekens, S. P. (2002). Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* **18**(12), 1593–1599.

Lonnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistical Sinica* **12**, 31–46.

Naef, F. and Magnasco, M. O. (2003). Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Physical Review E* **68**, 011906.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F., and Tsui, K. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

Pan, W., Lin, J., and Le, C. (2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Functional & Integrative Genomics* **3**(3), 117–124.

Peyser, B. D., Irizarry, R. A., Tiffany, C., Chen, O., Yuan, D. S., Boeke, J. D., and Spencer, F. A. (2005). Improved statistical analysis of budding yeast tag microarrays revealed by defined spike-in pools. *Submitted* .

Schena, M., Shalon, D., Davis, R. W., and O, B. P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235), 467–70.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P., and Davis, R. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences U S A* **93**, 10614–10619.

Singh-Gasson, S., Green, R. D., Yue, Y., Nelson, C., Blattner, F., Sussman, M. R., and Cerrina, F. (1999). Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nature Biotechnology* **17**, 974–978.

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

Tusher, V., Tibshirani, R., and Chu, C. (2001). Significance analysis of microarrays applied to ionizing radiation response. *Proceedings of the National Academy of Sciences U S A* **98**, 5116–5121.

Wolfinger, R., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**(6), 625–637.

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

Wu, Z. and Irizarry, R. (2004). Stochastic models inspired by hybridization theory for short oligonucleotide arrays. In: *Proceedings of RECOMB 2004*.

Wu, Z., Irizarry, R., Gentlemen, R., Martinez-Murillo, F., and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association* **99**(468), 909–917.

Yang, I. V., Chen, E., Hasseman, J. P., Liang, W., Frank, B. C., Wang, S., Sharov, V., Saeed, A. I., White, J., Li, J., Lee, N. H., Yeatman, T. J., and Quackenbush, J. (2002a). Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biology* **3**(11), research0062.1–0062.12.

Yang, Y. H., Buckley, M. J., Dudoit, S., and Speed, T. P. (2002b). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational & Graphical Statistics* **11**(1), 108–136.

Yuan, D., Pan, X., Ooi, S., Peyser, B., Spencer, F., Irizarry, R., and Boeke, J. (????) .