Title

Bivariate microarray analysis: statistical interpretation of two-channel functional genomics data

Journal

Systems and Synthetic Biology, 2(3-4)

ISSN

1872-5325

Authors

Hsiao, Albert
Subramaniam, Shankar

Publication Date

2008-12-01

DOI

10.1007/s11693-009-9033-8

Peer reviewed

Syst Synth Biol (2008) 2:95–104
DOI 10.1007/s11693-009-9033-8

RESEARCH ARTICLE

Bivariate microarray analysis: statistical interpretation of two-channel functional genomics data

Albert Hsiao · Shankar Subramaniam

Received: 5 October 2006 / Revised: 11 July 2009 / Accepted: 13 July 2009 / Published online: 13 August 2009

Abstract Conventional statistical methods for interpreting microarray data require large numbers of replicates in order to provide sufficient levels of sensitivity. We recently described a method for identifying differentially-expressed genes in one-channel microarray data 1. Based on the idea that the variance structure of microarray data can itself be a reliable measure of noise, this method allows statistically sound interpretation of as few as two replicates per treatment condition. Unlike the one-channel array, the two-channel platform simultaneously compares gene expression in two RNA samples. This leads to covariation of the measured signals. Hence, by accounting for covariation in the variance model, we can significantly increase the power of the statistical test. We believe that this approach has the potential to overcome limitations of existing methods. We present here a novel approach for the analysis of microarray data that involves modeling the variance structure of paired expression data in the context of a Bayesian framework. We also describe a novel statistical test that can be used to identify differentially-expressed genes. This method, bivariate microarray analysis (BMA), demonstrates dramatically improved sensitivity over existing approaches. We show that with only two array replicates, it is possible to detect gene expression changes that are at best detected with six array replicates by other methods. Further, we show that combining results from BMA with Gene Ontology annotation yields biologically significant results in a ligand-treated macrophage cell system.

Keywords Microarray · Statistical analysis · Bayesian · Bivariate microarray analysis

Abbreviations

BMA	Bivariate microarray analysis
VAMPIRE	Variance-modeled posterior inference with regional exponentials
FDR	False discovery rate
RMSD	Root mean square deviation
AfCS	Alliance for cellular signaling
GO	Gene ontology
KEGG	Kyoto encyclopedia of genes and genomes
MHC	Major histocompatibility complex
LPS	Lipopolysaccharide

Introduction

Microarrays are invaluable tools for measuring transcriptional responses of cells and tissues. The most common and perhaps the most fundamental question that is asked is

Electronic supplementary material The online version of this article (doi:10.1007/s11693-009-9033-8) contains supplementary material, which is available to authorized users.

Availability: The statistical approach presented here is implemented as a set of Java-based tools, accessible at our website, http://genome.ucsd.edu/microarray.

A. Hsiao · S. Subramaniam (✉)
Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA
e-mail: shankar@ucsd.edu

A. Hsiao
e-mail: alhsiao@ucsd.edu

S. Subramaniam
Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093, USA

A. Hsiao
Medical Scientist Training Program, University of California, San Diego, La Jolla, CA 92093, USA

whether observed differences in gene expression are statistically significant. Traditionally, a fold-change cutoff has been used as an indicator of significance. Although fold-change can be a consistent measure at high signal intensities, it is not very reliable when the signals are low. Low intensities are therefore filtered out prior to applying a fold change threshold, leaving these genes poorly interpreted. More rigorous statistical techniques are therefore desired to overcome these limitations.

The variance structure of microarray data has been widely known and modeled by a number of groups (Ideker et al. 2000; Li and Wong 2001; Rocke and Durbin 2001). It is believed that this variance structure can be decomposed into an expression-dependent and an expression-independent component. The approaches generally taken however, have been to "normalize" the microarray data to remove this undesired behavior, and process the resultant data with existing statistical tests (Sasik et al. 2002; Durbin and Rocke 2004; Irizarry et al. 2003). We have taken a fundamentally different approach to this problem. Recognizing that the sample variance is itself a poor estimator of noise at low replicate numbers ($n = 2$–3), we hypothesized that the variance structure itself could be used as a more precise estimator of variance (Hsiao et al. 2004). We subsequently devised a modeling procedure to identify the maximum likelihood estimates of expression-dependent and expression-independent variance. This model was then incorporated directly into a Bayesian statistical test.

In addition to normalization strategies, several statistical methods have been described to interpret gene expression data. Non-parametric methods such as Significance Analysis of Microarrays (SAM) (Tusher et al. 2001) are widely applicable because they do not rely on explicit assumptions about the error structure. These non-parametric methods, however, lack sensitivity in the absence of high levels of replication. On the other hand, parametric methods such as the *t*-test and Bayesian variants like Cyber-T (Baldi and Long 2001) can be powerful because of their ability to extrapolate the behavior of variability based on prior assumptions about the error structure. By parameterization, these methods reduce the minimum number of array replicates needed to identify significant changes in gene expression. Despite these advances however, existing methods still require many replicates to achieve sufficient sensitivity for biological interpretation.

We present a novel statistical approach for the interpretation of microarray functional genomics data that can be performed on as few as one measurement per experimental condition. Our method, the BMA (Bivariate Microarray Analysis) is based on the Bayesian framework developed in Hsiao et al. (2004) for the interpretation of one-channel array data. In addition to modeling the relationship between signal intensity and variance, BMA also models the covariation between the two color channels. We demonstrate in simulated data the dramatically increased sensitivity of BMA over other standard approaches, particularly at low replicate numbers ($n = 2$). This remains true independent of whether we choose an array-wide false-positive rate (α_{Bonf}) or a false discovery rate (FDR) as a significance threshold. Lastly, we show that when BMA is applied to a time course study of lipopolysaccharide (LPS)-treated macrophages, the differential gene expression pattern found with a single dye-swapped pair is very similar to that detected with three dye-swap pairs.

Statistics and modeling

Bayesian framework

The mean of intensity measurements is typically used as a point estimate for gene expression and the variance from that mean is used as an indicator of variability. Our method relies on the notion that a model that fits the error structure of microarray data will be a better estimator of variability. By integrating this variance model into a Bayesian framework, which is discussed more thoroughly in Supporting Information Sect. I, we transform the mean intensity into a probability density for gene expression. The resulting probability density describes the likely values for "true expression level" (μ). We can then apply statistical tests on this density to determine which genes are differentially-expressed between two experimental conditions. The foundation of our approach, then, lies at accurately modeling the amount of noise in microarray data.

The error structure of microarray data

Microarray gene expression data displays greater fractional error at low signal intensities than at high intensities. When the same RNA samples are analyzed on multiple arrays, this relationship is still present, indicating that much of this variation comes from the microarray platform itself. We believe that platform-specific factors, such as the resolution of the microscope and dye system, fundamentally limit the reliability of low-intensity measurements. We refer to this type of noise as expression-independent variance (B). At greater signal intensities, this physical limit becomes less and less important. Here, expression-dependent variance (A) begins to dominate, but both kinds of error are small compared to the signal. By itself, this simple model can explain much of the variance behavior in a single channel.

In a two-channel microarray, measurements are obtained from two RNA samples labeled with different dyes. These samples are subsequently hybridized with the microarray probes (Brown and Botstein 1999). Since these

probes are spotted, rather than synthesized directly on the array surface, there can be considerable variability between the amounts of probe spotted between arrays. This variation in "probe intensity" alters the intensity measurements for both color channels, and leads to covariation of the paired signals. We therefore chose to model this covariation by introducing correlation coefficients for expression-dependent and expression-independent variance. The final variance model contains a total of six parameters. These parameters are related to the treatment variances (σ_1, σ_2) and correlation (ρ) through:

$$\sigma_1^2 = \mu_1^2 A_1 + B_1 \quad (1)$$

$$\sigma_2^2 = \mu_2^2 A_2 + B_2 \quad (2)$$

$$\rho\sigma_1\sigma_2 = \rho_A \mu_1 \mu_2 \sqrt{A_1 A_2} + \rho_B \sqrt{B_1 B_2} \quad (3)$$

where (A_1, A_2) are the expression-dependent variances, (B_1, B_2) are the expression-independent variances, ρ_A is the expression-dependent correlation, ρ_B is the expression-independent correlation, and μ_1, μ_2 are the "true" expression level of each treatment condition.

Parameter estimation

Parameters for the variance model can be computed by Markov Chain Monte Carlo (MCMC) simulation of the bivariate normal density. This simulation computes the maximum likelihood estimates for each of the six variance parameters. Since the "true" gene expression is not known prior to observing array measurements, the observed sample mean is used as a surrogate. This choice, although necessary, can be problematic at low intensities. With the sample mean as a surrogate for true expression, poor signal-to-noise ratios can cause underestimation of true variability. Therefore, we have devised an estimation technique that identifies parameters which are stable against an expression-level cutoff.

Model parameters should be the same regardless of how much data is used to estimate them. Thus, we may progressively discard increasing numbers of low-intensity features to avoid the downward bias caused by low signal-to-noise. In Fig. 1, we demonstrate the effect of these cutoffs on estimates of expression-independent error. This parameter has two regions that are relatively stable against successive quantile cutoffs. When smaller cutoffs are used, expression-independent error tends to be underestimated. When larger cutoffs are used, the parameters become unstable. The details of the modeling procedure are discussed in Supporting Information Sect. II. It is also important to note that the expression-level cutoff defined here is only used to improve the quality of parameter estimates. The resulting variance model is applied across the entire array. By combining this model with the Bayesian framework described earlier, we can then use a statistical test to determine the significance of observed differences in gene expression.

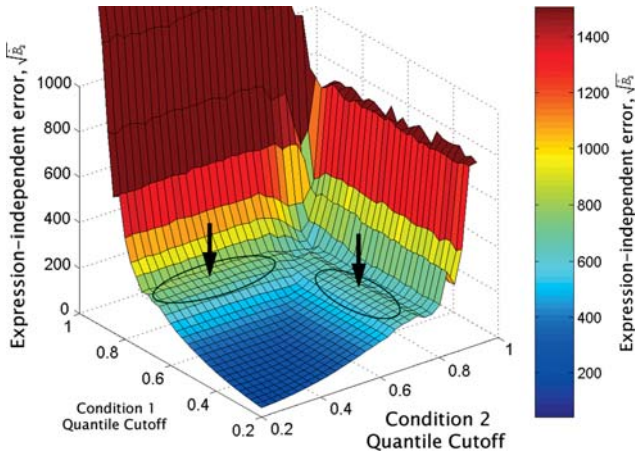

Fig. 1 Estimation of expression-independent error in a sample data set. The estimation procedure computes parameters for a series of expression-level quantile cutoffs. Array features with average measurements below the cutoffs are discarded, and the most likely parameters for the remaining data are computed. In this figure, each point on the colored surface describes the estimated value of expression-independent error ($\sqrt{\hat{B}_2}$) for a given pair of cutoffs. There are two rectangular regions where expression-independent error appears to be stable. These regions are identified by arrows

Statistical test

In order to make BMA compatible with conventional concepts widely understood by the biological research community, we devised a statistical test that computes a *P*-value. This *P*-value may be used to control type I error, which is the rate at which the null hypothesis is incorrectly rejected. In BMA, we define the null hypothesis (H_0) to be that the "true" gene expression does not change across two experimental conditions. The alternative hypothesis (H_1) is accepted when the null hypothesis is rejected. In other words, we test

$$H_0 : |\mu_2 - \mu_1| = 0 \text{ against } H_1 : |\mu_2 - \mu_1| \neq 0 \quad (4)$$

where μ_1 and μ_2 represent the "true" expression of the two treatment conditions.

The *P*-value of BMA is defined in a way similar to that of a *t*-test. In a *t*-test, the *P*-value is defined as an integral of a *t*-distribution. When the "tail area" of the *t*-distribution is sufficiently small, we consider experimental results to be statistically significant. In BMA, we define the *P*-value as a two-dimensional integral of a bivariate normal density. In other words, we determine whether the "tail volume" of this distribution is sufficiently small. If so, we consider the gene to be differentially-expressed. A depiction of the bivariate integral is shown in Fig. 2. Explicitly,

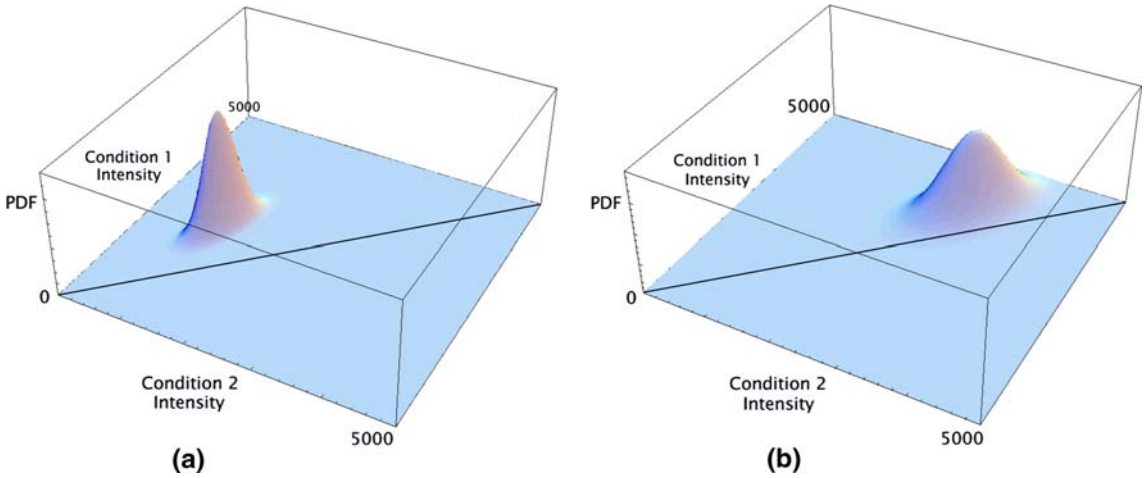

Fig. 2 Depiction of the significance integral. The statistical test performed by BMA involves integration of a bivariate normal density. The peak of the density appears on one side of the diagonal line. The *P*-value is defined as the "tail volume" on the side opposite of the peak. Statistical significance is achieved when the *P*-value is smaller than a specified threshold. The bivariate normal density depicted in (**a**) was sufficiently distant from the diagonal, and was considered a "significant" change ($\alpha_{\text{Bonf}} = 0.05$). The density depicted in (**b**) was too close to the diagonal, and therefore does not represent a significant difference in gene expression

$$p_i = \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\mu_2} \pi_i(\mu_1, \mu_2) d\mu_1 \cdot d\mu_2, & E[\mu_1] > E[\mu_2] \\ \int_{-\infty}^{\infty} \int_{\mu_2}^{\infty} \pi_i(\mu_1, \mu_2) d\mu_1 \cdot d\mu_2, & E[\mu_2] > E[\mu_1] \end{cases} \quad (5)$$

where π_i is the joint posterior density for the "true" expression levels of the *i*th feature.

A solution to the integration of the bivariate normal is given in Supporting Information Sect. III. When the *P*-value is less than the significance threshold, we reject the null hypothesis that $\mu_1 = \mu_2$, and consider the difference in gene expression to be statistically significant. The results of this test, when applied to a LPS-treated macrophage data set, are shown in Fig. 3.

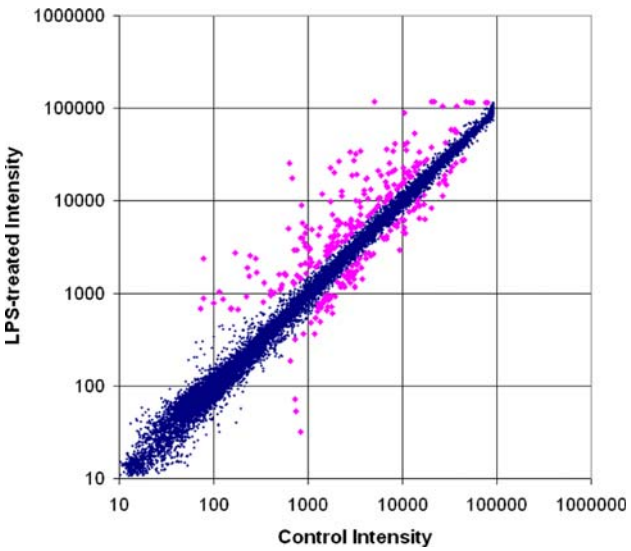

Fig. 3 A scatter plot demonstrating the results of BMA on macrophages treated with LPS for 1 h ($\alpha_{\text{Bonf}} = 0.05$). Each point represents the average signal measured from an array feature ($n = 2$). Highlighted dots appear farthest from the diagonal, indicating statistical significance

Specifying a significance threshold

There are two commonly used methods for determining a useful significance threshold (α) when analyzing microarray data. If we are interested in maintaining an array-wide false positive rate, we may set a Bonferroni-corrected threshold (α_{Bonf}). This is equivalent to the number of features we expect to appear as significant by random chance alone. If we set $\alpha_{\text{Bonf}} = 0.05$, then we expect only one feature to be significant when analyzing 20 batches of microarray experiments. Needless to say, this is a tremendously strict significance threshold. When the number of features is large, the *P*-value threshold for a two-sided test is

$$\alpha = \alpha_{\text{Bonf}}/2n, \quad (6)$$

where n is the number of array features.

In some cases, the false discovery rate (*FDR*) may be preferred. It is typically less stringent, but can still provide meaningful results. Assuming that the *P*-value defined here gives a reasonable estimate for the type I error, the false discovery rate is related to the array-wide false positive rate by

$$\text{FDR} = \frac{2n \cdot \alpha}{i} = \frac{\alpha_{\text{Bonf}}}{i}, \quad (7)$$

where i is the number of significant features.

Results

Simulated data

We tested the accuracy of several statistical methods on simulated data sets derived from the estimated variance structure of a two-color data set recently published by Rome et al. (2003). The authors analyzed the RNA content of human skeletal muscle biopsies collected before and after application of a hyperinsulinemic euglycemic clamp. Sample measurements for our simulated set were obtained from lognormal distributions centered on the pre-clamp mean intensities. The second treatment condition was obtained by "spiking" 10% of the features. The sample means of these randomly selected features were multiplied or divided by a random number between 1 and 20. Since the signal-to-noise ratio (SNR) is much greater at higher signal intensities, it is likely that subtle changes are more readily picked up at high levels of gene expression. This model problem therefore contains both subtle and dramatic gene expression changes at a variety of intensity levels.

Prior to identifying the "spiked" features, we tested the accuracy of two different modeling procedures. An equivalent model for the variance structure of two-color microarray data was previously proposed by Ideker et al. (2000). We applied their modeling approach, VERA (Variability and Error Assessment) along with our own, to compare parameter accuracy (Fig. 4). Even with two replicates, BMA showed striking accuracy in its estimate of expression-independent variance. VERA underestimated this parameter even with large replicate numbers ($n = 16$). We attribute the accuracy of BMA to the cutoff procedure, which reduces the effect of poor-quality measurements at low intensities. Since expression-independent variance dominates at low intensities, this precision is essential for interpreting faintly-expressed genes. In addition, both approaches had a tendency to underestimate expression-dependent variance when only two replicates were used, but improved with more replicates.

With these variance parameters, we then compared the accuracy of four different methods in identifying "differentially-expressed" genes. This comparison was performed using three different thresholds of statistical significance. The quality of the predictions is displayed in Table 1. Each of the statistical tests demonstrates improved sensitivity as significance thresholds were loosened and as larger numbers of experimental replicates were used. As we would expect, the FDR significance threshold was often predictive of the true false-positive rate. There is only one clear case where the false-positive rate substantially exceeded the desired FDR. At $n = 2$ and FDR = 0.05, BMA shows a loss in specificity. This results because expression-dependent variance was underestimated at $n = 2$. When the variance model derived from $n = 6$ data was used, the loss of specificity was reversed. In all other cases, our method appears to be much more sensitive than existing methods without sacrificing specificity.

These results show that BMA can accurately predict genes that are differentially-expressed. The variance-modeled approach is particularly beneficial at low replicate numbers where other statistical methods fail. In these conditions, the variance model estimates variability more accurately than traditional statistics. Applying this model in a Bayesian framework translates into improved sensitivity for subtle changes in gene expression.

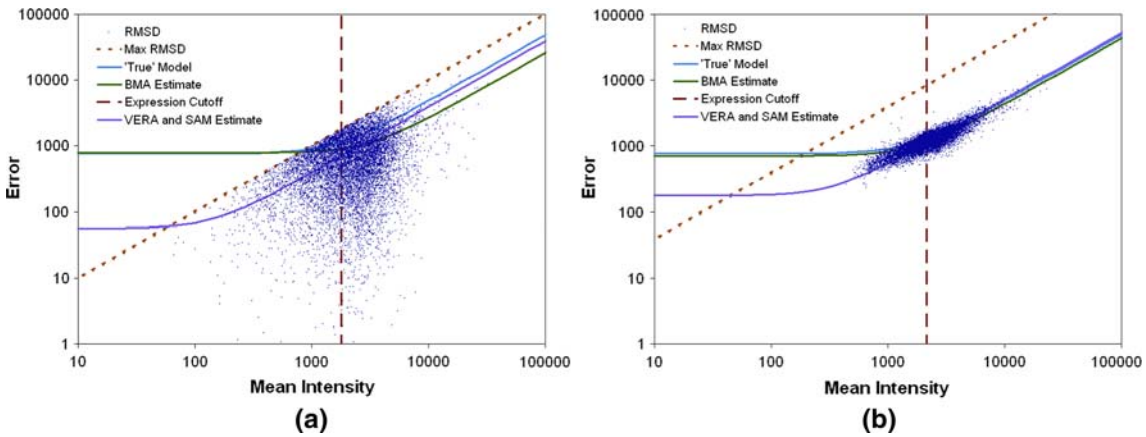

Fig. 4 The variance structure estimated by BMA tightly fits the RMSD (root mean square deviation) as well as the underlying variance model. The simulated data set shown here was obtained by sampling from a 'true' model derived from the Rome et al. (2003) two-color data set. The quality of fit to RMSD is shown for **a** $n = 2$ and **b** $n = 16$ in non-transformed coordinates. The vertical dashed line indicates the location of the quantile cutoff used to estimate the variance parameters. The diagonal line labeled "Max RMSD" shows the maximum possible value for RMSD at each expression level. The variance model estimated by BMA tightly matches the 'true' model, particularly at low intensities. This accuracy was not matched by VERA and SAM, even when large numbers of replicates were used

Table 1 Simulated data sets of 20,000 features, 2,000 spiked genes were analyzed with four different methods to determine the accuracy of predictions of differential-expression

Threshold	Method	TP	FP	FPR	FNR	Sensitivity	Specificity
n = 2							
$\alpha_{Bonf} = 0.05$	BMA	404	0	0.000	0.081	0.202	1.000
	Cyber-T paired	78	0	0.000	0.096	0.039	1.000
$\lambda_c = 23.8$	VERA and SAM	274	9	0.032	0.088	0.137	1.000
FDR = 0.001	BMA	657	14	0.021	0.069	0.329	0.999
	Cyber-T paired	139	0	0.000	0.094	0.070	1.000
	SAM paired	0	0	N/A	0.100	0.000	1.000
FDR = 0.05	BMA	1,171	442	0.274	0.045	0.586	0.975
	BMA (*n* = 6 model)	472	5	0.010	0.078	0.236	1.000
	Cyber-T paired	542	3	0.006	0.075	0.271	1.000
	SAM paired	130	1	0.008	0.094	0.065	1.000
n = 6							
$\alpha_{Bonf} = 0.05$	BMA	937	0	0.000	0.056	0.469	1.000
	Cyber-T paired	269	0	0.000	0.088	0.135	1.000
$\lambda_c = 23.8$	VERA and SAM	1,155	4	0.003	0.045	0.578	1.000
FDR = 0.001	BMA	1,211	3	0.002	0.042	0.606	1.000
	Cyber-T paired	387	0	0.000	0.082	0.194	1.000
	SAM paired	0	0	N/A	0.100	0.000	1.000
FDR = 0.05	BMA	1,514	138	0.084	0.026	0.757	0.992
	Cyber-T paired	925	2	0.002	0.056	0.463	1.000
	SAM paired	266	8	0.029	0.088	0.133	1.000
n = 16							
$\alpha_{Bonf} = 0.05$	BMA	1,479	3	0.002	0.028	0.740	1.000
	Cyber-T paired	829	0	0.000	0.061	0.415	1.000
$\lambda_c = 23.8$	VERA and SAM	1,554	34	0.021	0.024	0.777	0.998
FDR = 0.001	BMA	1,589	16	0.010	0.022	0.795	0.999
	Cyber-T paired	1,096	0	0.000	0.048	0.548	1.000
	SAM paired	930	0	0.000	0.056	0.465	1.000
FDR = 0.05	BMA	1,718	160	0.085	0.016	0.859	0.991
	Cyber-T paired	1,441	6	0.004	0.030	0.721	1.000
	SAM paired	1,601	61	0.037	0.022	0.801	0.997

The number of true-positives (*TP*), number of false-positives (*FP*), false-positive rate (*FPR*), false-negative rate (*FNR*), sensitivity, and specificity are reported for each method when either the Bonferroni-corrected threshold (α_{Bonf}) or false discovery rate (*FDR*) are controlled. BMA demonstrates substantial improvements in sensitivity across all significance thresholds investigated. A considerable loss of specificity was only observed at low replicate numbers ($n = 2$) when the significance threshold is loose (FDR = 0.05)

Time-course study of LPS-treated macrophages

We also examined the effectiveness of our method along with several others on two-channel microarray data available from the Alliance for Cellular Signaling (AfCS). [Reference: http://www.signaling-gateway.org]. In this data set, three pairs of dye-swapped measurements were obtained for each of six time points during lipopolysaccharide (LPS) treatment of RAW 264.7 macrophages, for a total of 36 arrays. Data sets such as this are valuable for understanding the dynamics of gene expression in response to ligand. In addition, much is already known about the behavior of macrophages in the presence of LPS. This is an ideal system then, to determine whether it is possible to interpret gene expression responses while using fewer microarray replicates.

When all three dye-swap replicates are available, similar sets of differentially-expressed genes can be obtained by any of the previously described methods. At the 8 h time point, for example, 1,492 genes were detected as differentially-expressed by BMA, SAM, and Cyber-T (FDR = 0.001). In contrast, only 300 were detected solely by BMA, 279 solely

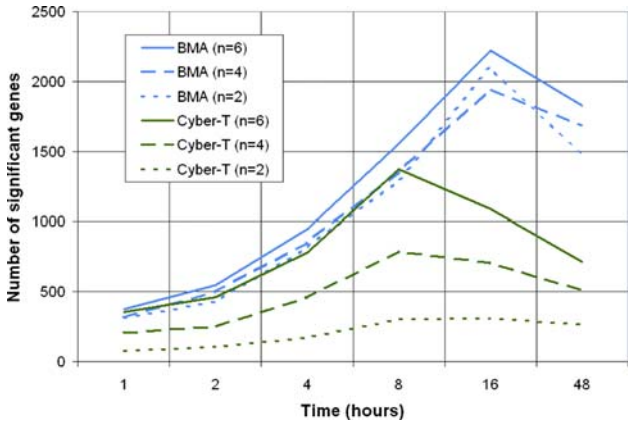

Fig. 5 BMA can use as little as a single pair of dye-swapped measurements to identify an equivalent number of statistically significant gene expression changes in the LPS-treated RAW 264.7 data set at all six time points. The quantity of significant gene changes ($\alpha_{Bonf} = 0.05$) stays relatively constant whether or not more replicates are used. Cyber-T's results approach BMAs as the number of replicates increase

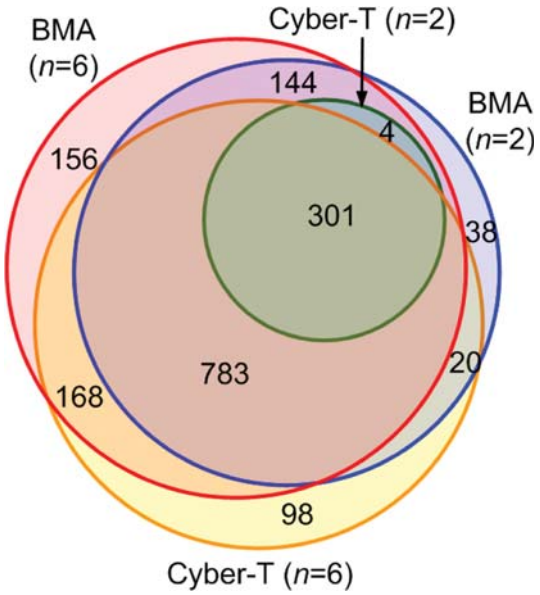

Fig. 6 The identity of genes identified by BMA and Cyber-T are very similar at the 8 h when all replicate data is used ($n = 6$). BMA also obtains similar results when fewer replicates are used ($n = 2$). A significance threshold of $\alpha_{Bonf} = 0.05$ was used for all methods

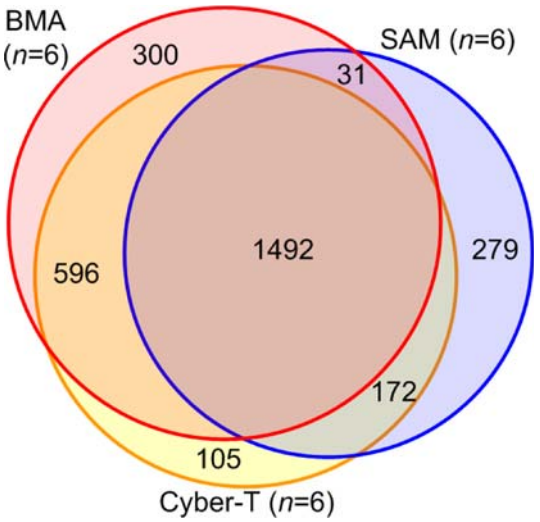

Fig. 7 The identity of genes identified by BMA, Cyber-T, and SAM are similar at 8 h when all replicate data is used ($n = 6$). A significance threshold of FDR = 0.001 was used for all methods

by SAM, and 105 solely by Cyber-T. The methods differ tremendously with fewer array replicates, however. When SAM is given fewer than three dye-swap pairs, no significance threshold can be found that satisfies the desired false-discovery rate. With Cyber-T, similar to what is observed in simulated data, far fewer numbers of significant gene expression changes are detected when either two dye-swap pairs ($n = 4$) a single dye-swap pair ($n = 2$) are used (Fig. 5). In contrast, our method detects similar numbers of gene changes regardless of how many replicates are used in analysis. This is further confirmed by comparing the sets of differentially expressed features (Figs. 6, 7). Despite using only the information contained in a single dye-swap pair, BMA identifies equivalent sets of significant changes. Furthermore, we found that similar Gene Ontology (GO) terms were statistically enriched whether we used a single dye pair or all of the available data (Table 2). In particular, genes involved in cell death and proliferation were strongly regulated by LPS. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were also similarly enriched (Supporting Information). Therefore, although increased sensitivity can be gained by increasing replicate number, BMA allows us to make meaningful biological interpretations of two-channel functional genomics data with only a single dye-swap pair.

We further examined the 985 features that were uniquely found by BMA with only two replicates at the 8 h time point. Among these were 17 features associated with components of the toll-like receptor signaling pathway as defined by KEGG. This list includes prominent figures such as toll-like receptor 2 (Tlr2) and toll-like receptor 4 (Tlr4). Tlr4 is believed to be required for LPS-induced signaling (Ulevitch and Tobias 1999), while expression of Tlr2 and Tlr4 have previously been confirmed to be induced in murine alveolar macrophages by LPS (Oshikawa and Sugiyama 2003). Furthermore, it has long been known that LPS stimulation induces macrophage expansion in vivo (Yokochi et al. 1985). This effect on gene expression is striking at the 8 h time point, as BMA found 115 additional features involved in cell proliferation. As

Table 2 Statistically enriched GO terms among differentially-expressed features after 1 h of LPS treatment (BMA, $\alpha_{Bonf} = 0.05$)

GO ID	GO term name	Count	*P*-value
n = 2			
GO:0006952	Defense response	46	0.0000000000
GO:0006954	Inflammatory response	20	0.0000000000
GO:0006955	Immune response	41	0.0000000000
GO:0009607	Response to biotic stimulus	48	0.0000000000
GO:0045087	Innate immune response	20	0.0000000000
GO:0009605	Response to external stimulus	51	0.0000000001
GO:0009611	Response to wounding	22	0.0000000001
GO:0009613	Response to pest/pathogen/parasite	24	0.0000000041
GO:0050896	Response to stimulus	53	0.0000000047
GO:0005125	Cytokine activity	22	0.0000000060
n = 6			
GO:0006952	Defense response	48	0.0000000000
GO:0006955	Immune response	44	0.0000000000
GO:0009607	Response to biotic stimulus	51	0.0000000000
GO:0009611	Response to wounding	25	0.0000000000
GO:0006954	Inflammatory response	21	0.0000000001
GO:0045087	Innate immune response	21	0.0000000001
GO:0006915	Apoptosis	30	0.0000000007
GO:0012501	Programmed cell death	30	0.0000000010
GO:0009613	Response to pest/pathogen/parasite	27	0.0000000021
GO:0009605	Response to external stimulus	54	0.0000000060

The number of features annotated with the GO term and the *P*-value are also displayed. Only the 10 GO terms with the lowest *P*-values are shown. All displayed GO terms are significantly enriched ($\alpha_{Bonf} = 0.05$). The complete lists are provided in the Supporting Information

macrophages play a major role in antigen presentation of exogenous peptides, it is also interesting to note that seven additional features involved in MHC class I and one additional feature involved in MHC class II antigen presentation were also upregulated. This included several HLA loci and β_2-microglobulin. Class II MHCs are commonly thought to present exogenous antigens, but recent evidence has shown that immunity against *Mycobacterium tuberculosis* requires presentation on class I MHCs in a 'detour' pathway (Schaible et al. 2003). The consistency of these results with known macrophage physiology is strongly suggestive of the quality of genes detected by BMA. Furthermore, these 985 features would have been entirely missed by other methods if only two replicates were used.

Discussion

In our analysis of the performance of BMA against present statistical methods, BMA provides improved sensitivity at all significance thresholds and replicate numbers tested. A loss of specificity was only observed in a simulated data set when both (a) the variance model was not sufficiently accurate and (b) a loose false discovery rate was desired. In the LPS-treated macrophage data set, the significant features detected with three dye-swap replicates were not substantially different from the features detected with a single dye-swap replicate. Thus, a loss of specificity does not appear be an issue when a stringent significance threshold is applied. In addition, since no additional experiments need to be performed to compute this threshold, our method is immediately applicable to existing data sets.

The scope of microarray expression profile studies has been limited by the costs of producing sufficient numbers of arrays to accommodate present statistical methods. BMA approaches this fundamental limitation by modeling the relationships between variability and gene expression, and applying this array-wide model as a more accurate indicator of error. Although some previous attempts at modeling the sources of noise have been described (Ideker et al. 2000; Li and Wong 2001; Rocke and Durbin 2001), much of the literature is devoted to "normalizing" array data to reduce variability. While these forms of normalization are not necessarily incompatible with BMA, we believe they may actually introduce additional artifacts, particularly when their assumptions are too strong. For example, quantile and lowess normalization fail to account for the poor signal-to-noise ratios at low intensities. When we modeled variance parameters for unnormalized data, we found a dramatic decrease in expression-independent

variance (Supporting Information Sect. II). Expression-dependent variance was slightly increased. More importantly though, we found that the expression-independent correlation coefficient (ρ_B) converged on a more realistic value, and was vastly more stable against quantile cutoffs. This suggests to us that the underlying variance structure at low intensities is clouded by these forms of normalization, and that caution should be exercised in using these procedures. With BMA, meaningful results can be obtained from raw data with minimal manipulation.

Furthermore, we believe that the current approaches for normalizing variability may ultimately be limited by the use of conventional statistical tests, which rely on large levels of replication to demonstrate significance. The levels of replication demanded by these methods do not frequently exist because of financial and labor constraints. They are even less likely to exist in multiple-ligand and time-course studies. BMA takes advantage of the parallel nature of microarrays to reduce replicate number. Two channel arrays have tremendous untapped potential for large-scale studies of functional genomics. We anticipate that because of BMAs improved sensitivity at low replicate numbers, it will provide (1) more consistent interpretations of array data and (2) an avenue for implementing more sophisticated experimental designs.

Conclusions

We have demonstrated here the effectiveness of our variance-modeled Bayesian approach on paired microarray data. BMA provides improved sensitivity at all significance thresholds and replicate numbers tested. The present method provides a reliable approach to identifying a set of differentially-expressed features with an a priori-specified significance threshold. Since existing statistical tests are unable to produce comparable results at low replicate numbers, this represents a substantial advance in microarray statistics.

Methods

In order to assess the effectiveness of several current statistical approaches, we created a simulated data set based on the error structure of the previously published data set of Rome et al. (2003). We computed variance parameters from this data set and used this model to generate random paired-samplings from a lognormal density centered on the sample means. The simulated data set contains 20,000 features. 10% of these features were randomly chosen to be spiked by multiplying the sample mean by e^s, where s is a sampling from a uniform random variate with support on $[-\log_e(20), \log_e(20)]$.

In our analysis with VERA and SAM (Ideker et al. 2000), we used a likelihood ratio cutoff of $\lambda_c = 23.8$ for statistical significance. For Cyber-T, we performed paired analysis with $\nu = 10$, and detected significance in two ways: using $\alpha_{Bonf} = 0.05$ and manually adjusting α_{Bonf} to achieve the desired false-discovery rates of 0.001 and 0.05. False-discovery rates in SAM were adjusted by adjusting Δ until the desired FDR were achieved. For both versions of BMA, we implemented an automated procedure that identifies the appropriate α_{Bonf} for the desired FDR, based on golden section root finding. A lower bound for α_{Bonf} was set at the minimum achievable FDR. The range of α_{Bonf} was subsequently adjusted using Eq. 7 until α_{Bonf} could be narrowed to within 10^{-5}. To determine the minimum achievable FDR, we similarly used a golden section minimization strategy.

To demonstrate the effectiveness of each method on real data, the AfCS LPS-treated RAW 264.7 Agilent inkjet-deposited oligo data set was obtained from the AfCS web site at http://www.signaling-gateway.org/. The processed signal intensities were normalized with intensity-dependent lowess normalization to standardize signals between arrays, although similar results were obtained by BMA without normalization. In order to identify a stable set of variance parameters, a single variance model was obtained from a pooled data set containing all six time points. This model was then applied to all six time points to determine differential expression.

Statistical enrichment of Gene Ontology terms and KEGG pathways was computed by comparing the number of "significant" features annotated with a particular GO term to the background. For the background, we used all of the features on the Agilent array. Exact likelihoods (P-values and q-values) were computed directly from the hypergeometric distribution.

The modeling procedure and the bivariate microarray analysis were implemented in a set of command-line JAVA tools. These tools have been added to the VAMPIRE (Variance-Modeled Posterior Inference with Regional Exponentials) statistical package, and are available from our website at http://genome.ucsd.edu/microarray. The MCMC framework used in VAMPIRE and BMA is a freely available JAVA package known as Hydra. The CustomMetropolisHastingsSampler is used to carry out the simulation, with a NormalMetropolisComponentProposal generating potential states. In addition, we used the open source library COLT for random number generation.

Acknowledgments We thank Yi-Xiong Zhou and Trey Ideker for our valuable discussions, and Wendy Ching, Dorothy Sears and Jason Papin for their comments on the manuscript. This work was supported

by grants from the National Institutes of Health, NIGMS K54-GM62114 and NHLBI R33-HL087375, NIDDK P01-DK074868, a Life Sciences Informatics grant from the State of California and Pfizer La Jolla, and a grant from the Hilblom Foundation. AH is a graduate student in the UCSD Medical Scientist Training Program and is supported by a Fellowship from the Whitaker Foundation and the UCSD MSTP Training Grant T35-GM07198.

References

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

Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21:33–37

Durbin BP, Rocke DM (2004) Variance-stabilizing transformations for two-color microarrays. Bioinformatics 20:660–667

Hsiao A, Worrall DS, Olefsky JM, Subramaniam S (2004) Variance-modeled posterior inference of microarray data: detecting gene-expression changes in 3T3-L1 adipocytes. Bioinformatics 20: 3108–3127

Ideker T, Thorsson V, Siegel AF, Hood LE (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. J Comput Biol 7:805–817

Irizarry RA et al (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264

Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA 98:31–36

Oshikawa K, Sugiyama Y (2003) Gene expression of toll-like receptors and associated molecules induced by inflammatory stimuli in the primary alveolar macrophage. Biochem Biophys Res Commun 305:649–655

Rocke DM, Durbin B (2001) A model for measurement error for gene expression arrays. J Comput Biol 8:557–569

Rome S et al (2003) Microarray profiling of human skeletal muscle reveals that insulin regulates approximately 800 genes during a hyperinsulinemic clamp. J Biol Chem 278:18063–18068

Sasik R, Calvo E, Corbeil J (2002) Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. Bioinformatics 18:1633–1640

Schaible UE et al (2003) Apoptosis facilitates antigen presentation to T lymphocytes through MHC-I and CD1 in tuberculosis. Nat Med 9:1039–1046

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98:5116–5121

Ulevitch RJ, Tobias PS (1999) Recognition of Gram-negative bacteria and endotoxin by the innate immune system. Curr Opin Immunol 11:19–22

Yokochi T et al (1985) In vivo effects of bacterial lipopolysaccharide on proliferation of macrophage colony-forming cells in bone marrow and peripheral lymphoid tissues. Infect Immun 47: 496–501