Brazilian Journal of Probability and Statistics
2015, Vol. 29, No. 1, 1–33
DOI: 10.1214/13-BJPS226

Bayesian factor models for the detection of coherent patterns in gene expression data

Vinicius D. Mayrink[a] **and Joseph E. Lucas**[b]

[a]*Universidade Federal de Minas Gerais*
[b]*Duke University*

Abstract. A common problem in the analysis of gene expression microarray data is the identification of groups of features that are coherently expressed. For example, one often wishes to know whether a group of genes, clustered because of correlation in one data set, are still highly co-expressed in another data set. Alternatively, for some expression array platforms there are many, relatively short probes for each gene of interest. In this case, it is possible that a given probe is not measuring its targeted gene, but rather a different gene with a similar region (called cross-hybridization). Accurate detection of the collection of probe sets (groups of probes targeting the same gene) which demonstrate highly coherent expression patterns is the best approach to the identification of which genes are present in the sample. We develop a Bayesian Factor Model (BFM) to address the general problem of detection of coherent patterns in gene expression data sets. We compare our method to "state of the art" methods for the identification of expressed genes in both synthetic and real data sets, and the results indicate that the BFM outperforms the other procedures for detecting transcripts. We also demonstrate the use of factor analysis to identify the presence/absence status of gene modules (groups of coherently expressed genes). Variation in the number of copies of regions of the genome is a well known and important feature of most cancers. We examine a group of genes, representative of Copy Number Alteration (CNA) in breast cancer, then identify the presence/absence of CNA in this region of the genome for other cancers. Coherent patterns can also be evaluated in high-throughput sequencing data, a novel technology to measure gene expression. We analyze this type of data via factor model and examine the detection calls in terms of read mapping uncertainty.

1 Introduction

Multivariate statistical methods for analysis of high-dimensional data have been the topic of many publications in the last decade. A challenging issue, connected with high-dimensional data, is the fact that the number of variables is much larger than the number of samples. In particular, gene expression data from DNA microarrays are characterized by measurements of many different genes, often in

Key words and phrases. Coherent, copy number alteration, detection call, factor model, high-throughput data, microarray.

Received February 2012; accepted August 2013.

only a few samples. Although the number of genes is large, there may be only a few underlying gene components accounting for much of the variation in the data. The analysis of high dimensional data requires special techniques such as variable selection or dimension reduction. Factor models are a flexible and powerful tool to analyze multivariate dependence and to verify patterns and relationships in the data. Several types of factor models with different constraints and computational algorithms can be found in the literature. The Principal Component Analysis (PCA) is a well-known dimension reduction method used in various applications involving gene expression; see, for example, Yeung and Ruzzo (2001). However, the literature indicates that other techniques can be superior in terms of dimension reduction. As an example, consider the Partial Least Squares (PLS) method compared with PCA in Nguyen and Rocke (2002), and also evaluated in Boulesteix and Strimmer (2006). The regression response variable is taken into account in the PLS approach as opposed to PCA, and thus the PLS usually performs better than PCA in prediction problems. Another technique used to reduce the dimensionality of the data, defining a linear combination of a reduced set of factors, is called Non-negative Matrix Factorization (NMF). The method is applied to the analysis of gene expression data by Brunet et al. (2004) and Kim and Tidor (2003), and within this context the factors represent groups of genes (metagenes) strongly correlated in subsets of the data. The NMF might be useful to study biological subsystems, because it can identify local and global patterns of similarities between genes. In contrast, other techniques, such as PCA, focus on only global patterns.

Liu et al. (2005) use information from multiple chips to analyze gene expression data with the purpose of providing the uncertainty of the measured expression value for each gene. Their approach (multi-mgMOS) is designed to model the binding affinity of probe pairs and to capture the effect of specific binding to MM probes. The model includes a probe-specific parameter that is shared across chips to identify composition effects. They argue that the likelihood can be written in closed form and this allows fast gradient-based optimization to obtain the parameters.

Computational advances have been critical in enabling the application of complex models for analysis of large data sets. The development of iterative MCMC simulation methods has contributed to the increasing attention devoted to the Bayesian framework as a good alternative to work with factor models. In recent years, numerous studies have applied factor models combined with the Bayesian approach to the analysis of gene expression data, and their results often show an improvement in the identification and estimation of metagene groups and patterns related to the underlying biology. As an example, West (2003) introduced sparse latent factor models, as a natural extension of the sparse regression modeling. The study assumes typical Bayesian variable selection priors and demonstrates the ability of latent factor models to describe pattern/signature profiling in expression genomics. Lucas et al. (2006) also apply hierarchical sparsity priors and obtain substantial improvements in terms of identification of complex patterns of covariation among genes. The paper explores the Bayesian methodology for large-scale regression, ANOVA and latent factor models. Carvalho et al. (2008) is another reference working with sparsity priors to address dimension reduction on latent factor models applied to gene expression data. Stochastic simulation and evolutionary stochastic search methods are used in the paper to address questions of uncertainty about the number of latent factors. This same issue is also evaluated in Lopes and West (2004) via reversible jump MCMC methods.

In the present paper, we consider factor models to study the expression pattern across Affymetrix oligonucleotide microarrays. In the first application, our goal is to evaluate the presence/absence status of probe sets targeting genes. We compare the factor model with two other detection techniques proposed in the literature. In the second application, we study the Copy Number Alteration (CNA) in different regions of the genome. We examine a gene list, detected with CNA in breast cancer, to investigate whether the CNA is affecting other types of cancer at the same chromosomal location. We develop a third application (see Appendix A) to evaluate the coherent pattern across samples of RNA-Seq data. The introduction of high-throughput sequencing technologies has opened new doors into the field of gene expression. Our goal is to study the presence/absence status of genes and the corresponding read mapping uncertainty.

The outline of this paper is as follows. Section 2 introduces the first application with an overview of existing detection methods. The data and few strategies to adjust the measurements are described in Section 2.1. Next, the factor model is proposed in Section 2.2. Section 2.3 shows inference results to verify the performance of the BFM in terms of parameter estimation. Section 2.4 presents the comparison between BFM and other detection methods, using a simulated data set. Section 2.5 develops the comparison analysis based on a real data set from a spike-in study designed by Affymetrix. In Section 3, the factor analysis is used to test whether a group of genes, detected with CNA in breast cancer, is coherently expressed in other types of cancer. The sparse latent factor model is presented in Section 3.1. The performance of the model for inference results is evaluated in Section 3.2. In Section 3.3, we show a real data application involving 7 data sets. Finally, Section 4 presents the conclusions. The algorithms required to fit the factor models were implemented using the MATLAB programming language (http://www.mathworks.com).

2 Presence/absence calls for gene expression

In gene expression analysis, it is unlikely that all probe set sequences in a microarray will find its complementary sequence in the hybridization solution. An efficient method is then required to distinguish between the probe sets detecting transcripts of those genes present in the target sample and the probe sets expressing only nonspecific binding or background noise. Approaches exploring this problem have been proposed for different microarray platforms, for example, Affymetrix implemented the detection above background (DABG) method for detection calls in GeneChip® Exon Arrays (Affymetrix, 2005), Kapur et al. (2007) apply a method called "GeneBASE" to investigate presence/absence calls in the same type of array and Li and Wong (2001) study experiments involving replicate arrays using a software "dChip" which includes a presence/absence method similar to the MAS 5.0 detection procedure.

A popular detection method was developed for Affymetrix GeneChip® oligonucleotide arrays, and it is implemented as part of the preprocessing technique Microarray Suite version 5.0 or MAS 5.0 (Affymetrix, 2001). In this case, the detection calls can take the values absent, marginal or present. Both perfect match (PM) and mismatch (MM) probes are used to calculate the score $R = (\text{PM}-\text{MM})/(\text{PM}+\text{MM})$ for each probe pair within a probe set. First, the method tests whether all the probe pairs are saturated, if so the probe set is classified as present, otherwise a one-sided Wilcoxon's signed rank test is applied to obtain a p-value which is used to assign a detection call. The default parameters of this presence/absence (P/A) method are studied in Archer and Reese (2009) using spike-in data sets. McClintick and Edenberg (2006) and Liu et al. (2002) apply MAS 5.0 P/A and show that it works relatively well in practice; however, other frameworks improving this solution can be found in the literature. As an example, Wu and Irizarry (2005) define a procedure called Half-price method. They argue that using MM data may be problematic in gene expression analysis, and then define a method based on PM probes only. Their results indicate that the half-price technique outperforms MAS 5.0 P/A in terms of detection calls. Two recent studies using MAS 5.0 P/A in a multiple-array analysis are Ouandaogo et al. (2011) and Tiedermann et al. (2012).

Another interesting method was proposed by Warren et al. (2007), and it is named Presence Absence calls with Negative Probe sets (PANP). In brief, the authors identify Affymetrix probe sets which cannot hybridize to the intended target, because they have been designed in the reverse direction against their own transcripts. These probe sets are called Negative Strand Matching Probe sets (NSMPs) and can be obtained from chip annotation files available on Affymetrix webpage. The selected NSMPs are assumed as controls, and the empirical cumulative distribution of their intensities is used to derive a cutoff intensity; a particular probe set is classified as present if its expression value is higher than the threshold. Before applying this analysis, the data are preprocessed using any technique. In particular, the performance of PANP is evaluated in the paper with data obtained from RMA (Irizarry et al., 2003b), GCRMA (Wu et al., 2004) and MAS 5.0. Besides simplicity, the authors indicate that another advantage of PANP is the fact that it works with preprocessing techniques using (PM, MM) or PM-only probes. The paper shows that PANP combined with RMA, GCRMA and even MAS 5.0 outperforms the default detection method of MAS 5.0 in terms of several metrics of accuracy and precision. The choice of threshold in PANP is somewhat arbitrary and has a strong impact over the final result. If the threshold is slightly shifted, the detection calls of some probe sets will change.

In none of these techniques the coherent expression is used to identify the presence of a gene. Our Bayesian factor model is proposed to take advantage of the tremendous amount of information—available from studying the co-expression of probes across all samples—about the behavior of a probe set. The expression of one probe is the observed intensity value in one of the samples or microarray. A consistent expression pattern for all probes suggests presence, whereas probes randomly alternating intensities across arrays indicates absence; this characteristic will be considered to generate detection calls in the downstream analyses.

2.1 The data

Suppose n microarrays are available for analysis; each array contains K probe sets. The first step in a usual microarray analysis is to preprocess the data; the observed intensity values are transformed to remove the noise effect in an individual chip, and to adjust the information obtained from replicate microarrays. Details about three well-known preprocessing techniques can be found in Irizarry et al. (2003b), Wu et al. (2004) and Affymetrix (2001). Consider the following preprocessing steps which are used in the data analyses of Section 2.

1. In an individual chip, there are technical variables that affect the vast majority of spots on the microarray, such as total RNA in the sample and camera exposure time. When examining expression patterns across the samples, the overwhelming signal in the raw data reflects these effects. We address the indicated problem by dividing the probe intensities by the corresponding intensity mean computed for each array.
2. Even though samples may be extracted from the same type of tissue, the distribution of intensity values has a broad range and is highly skewed. Thus it is standard practice to log-transform expression data.
3. Consider a matrix X containing intensities of probe set $k \in \{1, \dots, K\}$ from each chip. Each row represents an individual probe within the probe set, and each column represents a microarray. The n intensities observed for any probe are assumed to follow a Normal distribution. Different probes may be associated with different mean and variance parameters. In order to evaluate whether the pattern of expression is consistent across samples for all probes belonging to probe set k, the rows of the indicated X matrix will be standardized.
4. Some microarrays are brighter than others due to technical effects introduced during their production (e.g., scanner setting and physical problems). In order to subtract away such effects, we compute the first principal component (pc_1) based on the entire data set. Note that pc_1 is computed only once, it is representative of the overall relative brightness of the chips and does not reflect the

biology of individual genes. For each row i (probe), we subtract the component of its expression in the direction of this principal component ($X_i - X_i pc_1 pc_1'$). See footnote 1 for further details about this step.

We recognize that there are a number of different approaches to removing processing effects from the raw data. The factor model we utilize is general and can be applied with any data cleansing approach. Appendix D shows data sets preprocessed via RMA background adjustment and quantile normalization.

We are dealing with Affymetrix arrays that consist of groups of 22–40 probes (a probe set) that are all intended to target the same gene. These come in pairs (11 to 20 pairs per probe set) with 1 perfect match and 1 mismatch probe in each pair. The biggest probe set has 20 probe pairs, and thus 40 probes.

Figure 1 shows the image of the specified X matrix for different probe sets. The previous manipulations have been applied to the original data. The data correspond to 251 oligonucleotide microarrays containing 22,283 probe sets. The transcripts in the hybridization solution are extracted from breast cancer tumors. As a reference for this data set consider Miller et al. (2005) which investigates the importance of the p53 tumor suppressor functional status for predicting human breast cancer behavior. Two distinct patterns can be observed when comparing the images. Image (a) displays a consistent sequence of intensities for every probe across the arrays. The values decrease when moving from the left to the right, and this pattern is the same for all probes. In image (a), the gene is present and the associated probes indicate such presence by exhibiting similar intensities within each array. On the other hand, image (b) shows that the probes randomly alternate intensities across samples. No pattern can be observed and within each array the probes disagree from each other by expressing different levels of intensities. This situation indicates absence of the gene, and the displayed values are just noise effects.

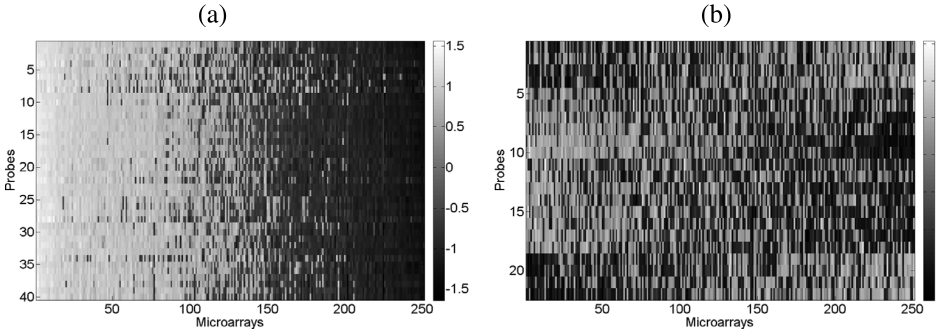

Figure 1 *Intensities of all probes within two different probe sets* (a) *and* (b). *Samples are along the* x*-axis and probes are on the* y*-axis. White corresponds to relatively high expression levels of the corresponding probe in the corresponding sample. The columns have been sorted so that the* 1*st principal component is monotone.* (a) *is an example displaying a strong and consistent pattern of intensities for every probe across samples.* (b) *is an example of a noise probe set with probes randomly alternating intensities across arrays.*

Any experimental design, controlling phenotypic information or any other variable possibly affecting individuals, has not been used to produce and select the microarrays included in our data set. In this context, it does not seem appropriate to extrapolate the interpretation of Figure 1 assuming associations with unknown clinical variables.

The existing detection methods (MAS 5.0 P/A and PANP) can provide presence/absence calls for every microarray and every probe set, whereas the factor model proposed in the next subsection uses the information from several microarrays to define a single detection call for each probe set. Our goal is to identify those genes that are potentially relevant to the experiment being conducted. In particular, genes may be present in some samples but not in others. In this case, our method would call the probe set as present because all of the probes to that probe set have low values on some samples (those where the gene is absent) and high values for other samples; see Figure 1(a). Alternatively, our method would count as absent any probe set whose probes are expressed at a non-constant level within each sample. In this case, the differential expression of the probes across the samples is explained by noise, and this would translate into a decoherent expression pattern.

2.2 Bayesian factor model for gene expression detection

Suppose X_{ij} is the log of light intensity of the probe i within a probe set on sample $j = 1, 2, \dots, n$. Assume further that the transformations indicated in Section 2.1 have been applied to the data. Let X be the matrix with element x_{ij} in the ith row and jth column. The following model[1] is assumed for the expression data of X:

$$X = \alpha\lambda + \varepsilon, \tag{1}$$

where α is an m-dimensional column vector reflecting the strength of hybridization between the target of the probe set and each of the m probes, component λ is a n-dimensional row vector describing the pattern of expression of the target across samples, and finally ε is the idiosyncratic noise term.

If the gene is present ($\alpha \neq 0$), all probes are expected to display expression values correlated with λ; therefore, we say "coherent pattern." Each element in λ is associated with the magnitude of the expression values in one of the samples.

We assume a mixture prior distribution on the factor loadings.

$$\alpha \sim (1 - q)N_m(0, \Phi_1) + qN_m(0, \Phi_2) \tag{2}$$

[1] We admit there is some variation that is systematic across all probe sets. Ideally, we could address this issue by assuming the model $X_k = \beta\delta + \alpha_k\lambda_k + \varepsilon$ for a probe set k, where $\beta\delta$ is the systematic variation. In this case, we would have to work with the full data set due to the dependence between probe sets, and this task is computationally expensive. Our normalization step 4 (Section 2.1) subtracts off an approximation of $\beta\delta$; hence, we can assume model (1) and work with each probe set independently.

with $\Phi_1 = \omega_1 I_m$ and $\Phi_2 = \omega_2 I_m$. We define q as the prior probability of detecting the probe set as present. A small and positive number is suggested for the scalar ω_1, while ω_2 is large and positive. Because we treat probe sets that have a high probability of having derived from the first component as noise, the relative sizes of ω_1 and ω_2 effectively define our estimated signal to noise ratio. A small ω_1 is chosen for defining a Normal component centered on zero with small variability, which indicates that the factor loadings are close to zero, and then the detection call "Absent" is appropriate. There are certainly pairs of probes that show correlation in the data, but we are inherently choosing between a model that explains everything as noise and a model that assumes that everything is correlated with everything else.

A point mass distribution at 0 has been considered as a replacement for component $N_m(0, \Phi_1)$ in the mixture prior (2). However, this is unsuitable for the factor loadings because the classification "Present" is obtained only if a completely random sequence of intensities is observed across samples. In practice, there is often structure in a probe set that has not been completely subtracted by the data cleaning techniques. Additionally, because the probes come in pairs with each pair different in only 1 of 25 locations, there is built in correlation between subsets of the probes even in the absence of the target gene of interest.

We complete the model specification with the conjugate priors $\lambda' \sim N_n(0, I_n)$ and $\varepsilon_{ij} \sim N(0, \sigma_i^2)$ with an Inverse Gamma distribution IG(a, b) for σ_i^2. Denote $\sigma^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)'$.

In order to sample from $p(\alpha, \lambda, \sigma^2|X)$, we implement the Gibbs Sampler algorithm; see Gamerman and Lopes (2006) for details about this Markov Chain Monte Carlo method. The likelihood function and the full conditional posterior distributions are shown in the Appendix B.[2] In particular, during the Gibbs sampling we draw α from the second mixture component with probability

$$q^* = \frac{N[0|0, \Phi_2]}{N[0|M_2^*, \Phi_2^*]} q \Big/ \left(\frac{N[0|0, \Phi_2]}{N[0|M_2^*, \Phi_2^*]} q + \frac{N[0|0, \Phi_1]}{N[0|M_1^*, \Phi_1^*]} (1 - q) \right), \qquad (3)$$

where M_1^*, M_2^*, Φ_1^* and Φ_2^* are specified in the Appendix B. It is the average of q^* across all draws from the MCMC that is used to assign presence/absence calls in the BFM.

2.3 Inference results of the Bayesian factor model

The main aim of this subsection is to verify the performance of the BFM in terms of inference results for its parameters. A simulated data set is considered in this application. First, the proposed BFM is fitted to the real data and the posterior

[2] In Section 2.4, we define a data generating procedure that requires the specification (1) including a mean expression parameter μ. The full model (with μ) is explored in Appendix B. Let $\mu = \mathbf{0}$ to identify results related to model (1).

estimates of the involved parameters are assumed as real values in the process of simulating data. Next, for each microarray a random value is generated from $N(0, \sigma_i^2)$ with $i = 1, 2, \dots, m$ representing a probe; this is the noise term in the BFM. The product $\alpha\lambda$ results in a $m \times n$ matrix which is added to the previous noise matrix. The pattern of a noise probe set can be simulated by letting α and/or λ be a null vector.

Assume as prior specifications: $\omega_1 = 0.01$ defining small variability for the first component within the mixture prior (2), $\omega_2 = 100$ defining large variability for the second component, $q = 0.5$ indicating equal prior probability for both components in the mixture. In addition, consider the Inverse Gamma prior distribution with $a = 2.1$ and $b = 1.1$, which has expected value 1, mode 0.3548 and variance 10. As initial values of the chain, consider $\lambda^{(0)}$ generated from its prior $N_n(0, I_n)$, and a null vector for $\alpha^{(0)}$; the value 1 is indicated for each $(\sigma_i^2)^{(0)}$.

We are aware of the identifiability problem related to the sign of α and λ. This issue does not affect q^* in (3); therefore, constraints are not imposed to address the problem. The Gibbs Sampler is run for 2000 iterations. The first 1000 elements of the chain are considered as burn-in period, and thus removed from the analysis. Fast convergence to the limiting distribution is observed for all chains we have examined.

Two scenarios are chosen to evaluate results: the first one is presented in Figure 1(a) showing a strong pattern of expression across samples, and the second one is indicated in Figure 1(b) suggesting a noise probe set. The corresponding simulated data sets are displayed in Figure 2. As can be seen, the simulated data are very similar to the real data in Figure 1.

Figure 3 presents box plots to study the distance between posterior estimates and real values of the parameters. The graphs for α and λ are concentrated around zero suggesting a small difference between the posterior estimates and the corresponding real values. The box plot for the ratio $\hat{\sigma}_i^2/\sigma_{i,\text{true}}^2$ is centered around 1 indicating

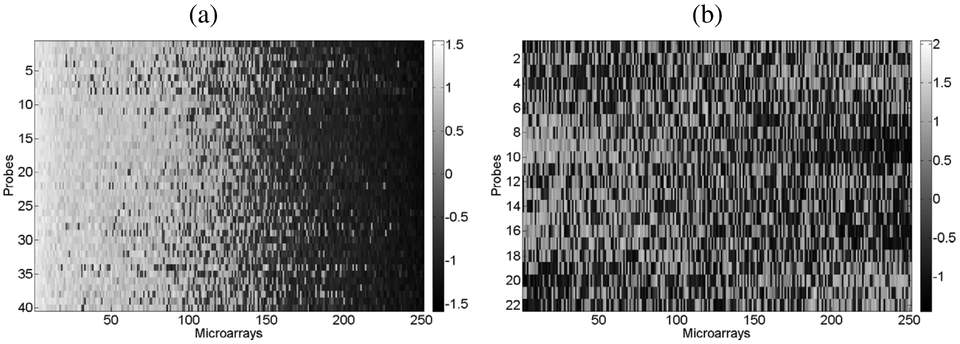

Figure 2 *Intensities of a probe set across microarrays (synthetic data).* (a) *and* (b) *reproduce the scenario of Figure* 1(a) *and* (b), *respectively.*

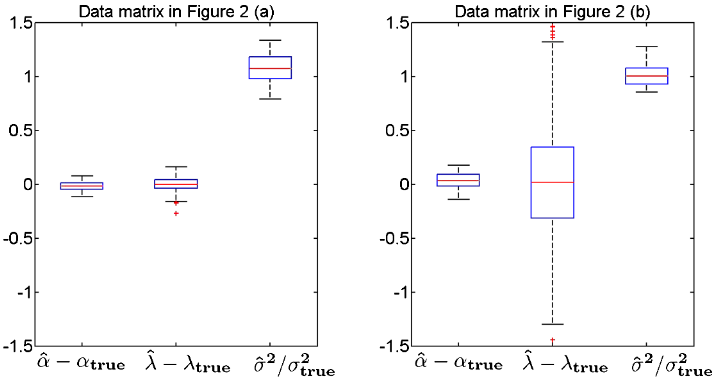

Figure 3 *Box plots summarizing the behavior of the bias (for α_i and λ_j) or the ratio (for σ_i^2) involving the posterior estimates and the true value of the parameters. Right: results associated with Figure* 2(a) *and Left: results associated with Figure* 2(b).

again a good approximation. In summary, these results show that the model is performing well in terms of inference. The comparison between the factor analyses developed in each panel (strong pattern probe set versus noise probe set) leads to the following additional conclusions: (i) the majority of the 95% intervals for parameters includes the real value, (ii) the noise probe set analysis produces wider 95% credible intervals (including zero) for λ and this higher posterior uncertainty reflects on the larger differences observed in the box plot for $\hat{\lambda} - \lambda_{\text{true}}$ (right panel), (iii) as expected, all 95% intervals for α do not include 0 in the strong pattern probe set analysis, while many of them includes 0 in the noise probe set analysis.

The Gibbs Sampler applied to the data shown in Figure 2(a) and (b) indicates that the posterior probability q^* converges to 1 and 0 respectively. This result confirms the visual interpretation of the images which suggests presence of the gene for panel (a) and absence for panel (b).

2.4 Comparison of detection methods, using a simulated data set

In a simulated scenario the true detection calls are known, and thus the performance of the methods can be evaluated. We will now estimate the characteristics of a real data set and use those estimates to generate a simulated data set for which we know whether each gene is present or absent. First, the real data are transformed as suggested in Section 2.1. We use 500 probe sets from this data set to generate 500 different simulated probe sets. The pairwise linear correlation coefficient between each pair of rows (probes) in the X matrix is computed, resulting in a $m \times m$ matrix, with m being the number of probes. An interesting aspect can be observed in the correlation matrix: probe sets with strong patterns across samples exhibit high correlations (close to 1), whereas a noise probe set is associated with low correlations (close to 0). This characteristic can be used to rank the probe sets by computing the average correlation in the matrix. A selection of 2000 probe sets are sorted in an increasing order of average correlations. Then, the first 200 cases are selected. In addition, starting with the 201st probe set and moving toward the case 2000, we select every 6th probe set for use in generating simulated data. This strategy ensures the selection of a wide range of cases including strong, intermediate and weak patterns across samples.

Real values of parameters are determined for each selected probe set. In the previous subsection, this task was accomplished by fitting the BFM to the transformed data; however, in this case other detection methods are considered in the analysis. MAS 5.0 P/A and PANP work with original data, that is, without the manipulations described in Section 2.1, which includes standardizing the data. Therefore, a slight modification is required in the BFM indicated in (1). Consider

$$X = \mu 1_n + \alpha\lambda + \varepsilon, \tag{4}$$

where μ is an m-dimensional column vector whose entries are mean intensities fixed for each probe, and component 1_n is a n-dimensional row vector of ones. The interpretation of other components remains the same. Assume the prior specification $\mu \sim N_m(0, 100I_m)$. The complete conditional posterior distributions for this model are specified in the Appendix B.

In summary, we consider the following steps to generate the data:

1. The factor model (4) is fitted to the 500 selected X matrices containing the original data (without manipulations). Again, convergence to the limiting distribution is fast, and the posterior mean is assumed in the next step as the real value of the involved parameters.
2. As described in the previous section, for each microarray a value is randomly generated from $N(0, \sigma_i^2)$ $(i = 1, 2, \dots, m)$ forming a noise matrix, which is added to the matrix $\mu 1_n + \alpha\lambda$. The group of 200 selected probe sets showing the weakest patterns across samples are chosen to represent noise probe sets. In other words, α and λ are set to be null vectors in these cases. The remaining 300 probe sets are generated using posterior estimates of all parameters.
3. The PANP method requires Negative Strand Matching Probe sets to be used as controls. In the generated data set, 100 probe sets simulated as noise cases are assumed as NSMPs. The detection call “Absent” is automatically assigned for these cases. In addition, the generated values are preprocessed using RMA when the method PANP is applied.
4. The MAS 5.0 P/A method is applied to the synthetic data obtained in step 2.
5. In the analysis of the BFM, we first apply the data cleaning procedure suggested in Section 2.1 to the generated data. Next, we fit the factor model (1) to the “clean” synthetic data.

Figure 4 displays Receiver Operating Characteristic (ROC) curves for the three detection methods. This graph plots true positive (TP) rates against false positive (FP) rates computed for different choices of a threshold parameter. In fact, two

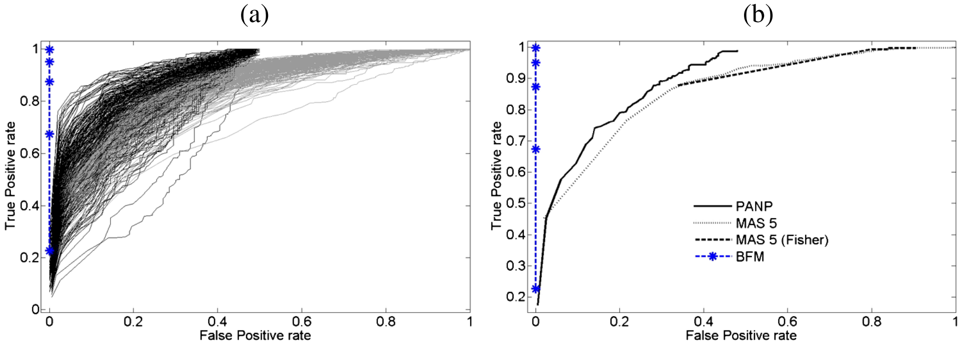

Figure 4 *ROC curves comparing true positive (TP) and false positive (FP) rates. In panel* (a) *we have: PANP (black solid lines) and MAS* 5.0 *P/A (gray solid lines). Panel* (b) *shows summarized results: we consider the most frequent call registered in* 251 *samples to derive the PANP (solid) and MAS* 5.0 *P/A (dotted) curves, and the dashed line represents the MAS* 5.0 *P/A result using Fisher's combined probability test (see Whitlock,* 2005*). The curve for the BFM (dashed line with asterisk) is built based on different sample sizes* (25, 50, 100, 150, 200 *and* 251 *arrays); the asterisk marks indicate the FP and TP rate for each size.*

thresholds are specified in PANP and MAS 5.0 P/A, below the first value the classification is "Present," and above the second value the classification is "Absent." In this analysis the detection "Marginal" will be suppressed by choosing the same number for both thresholds. In order to build the ROC curve, the detection methods (PANP and MAS 5.0 P/A) are applied to the data and their p-values are computed. Next, different choices of the single threshold are compared with the p-values, and then detection calls are defined. The true call is known because a simulated data set is used; therefore, TP and FP rates can be calculated for each case. The BFM defines a single list of detection calls taking into account all microarrays, whereas the other methods generate a list of calls for each chip, for this reason 251 ROC curves are shown for PANP and MAS 5.0 P/A in Figure 4(a). Panel (b) shows a single curve for PANP and MAS 5.0 summarizing the information from the multiple samples.

An interesting result is observed for the BFM: the posterior probability q^* in (3) converges to 1 for any probe set exhibiting an intermediate or strong pattern across samples. On the other hand, this probability converges to 0 for any probe set showing weak or no pattern. These extreme probabilities determine only presence or absence calls, and thus eliminate the need of defining thresholds for a "Marginal" detection call. The ROC curve for the BFM in Figure 4 shows the results for 6 different sample sizes ranging from 25 to 251; these samples are randomly selected from the group of 251 arrays. Given any threshold in the interval (0, 1), the (FP, TP) rates are summarized by the asterisk marks for each sample size.

The high level of certainty in presence/absence calls is in part a function of the size of the data sets. We have around 251 observations from a 22–40 dimensional (depending on the probe set) multivariate normal, which offers significant evidence for the presence or absence of a non-zero mean. We have evaluated the behavior of q^* in a MCMC run assuming a reduced number of probes and/or a reduced number of samples. The analysis involves a random selection of 5 rows and/or 5 columns of the two matrices presented in Figure 1. We have found that the probabilities q^* differing from 0 or 1 are common when small sample sizes and/or few probes are used in the study. However, the simulations we have presented are reflective of many publicly available data sets. In particular, the number of probes in a probe set is fixed and unchanging.

According to Figure 4, PANP outperforms MAS 5.0 P/A for detecting gene expression. The best result would be 0% false positives and 100% true positives which is represented by the point located in the top left corner of the graph. The closer the curve is to this point, the better is the performance of the method. As can be seen, most black solid lines (associated with PANP) are above the gray solid lines (representing MAS 5.0 P/A). This finding agrees with the result obtained by Warren et al. (2007). Only 5 asterisk marks are shown in Figure 4 because the results for 200 and 251 arrays are the same; we have FP = 0% for all cases and (TP = 22.67%) for 25 arrays, (TP = 67.33%) for 50 arrays, (TP = 87.33%) for 100 arrays, (TP = 95%) for 150 arrays, (TP = 99.67%) for 200 arrays and (TP = 99.67%) for 251 arrays. This clearly suggests that the BFM outperforms the other two methods when the sample size is large.

The number of microarrays available for analysis and the choice of ω_1 have an impact over the TP rate. Figure 5 explores these aspects. Panel (a) shows TP rates for different numbers of samples and different choices of ω_1. Note that, for a fixed ω_1, the rate increases as the number of arrays increases. The TP rate is 0% if only 10 arrays and $\omega_1 \geq 0.0075$ are considered. A strong pattern across 251 samples may not be displayed on 10 samples. The FP rate is 0% for all sample sizes. A random pattern across 251 samples is still random for smaller sample sizes. For a fixed number of arrays, the TP rate seems to decrease as ω_1 increases. This aspect is also explored in panel (b) which presents the TP rates for choices of

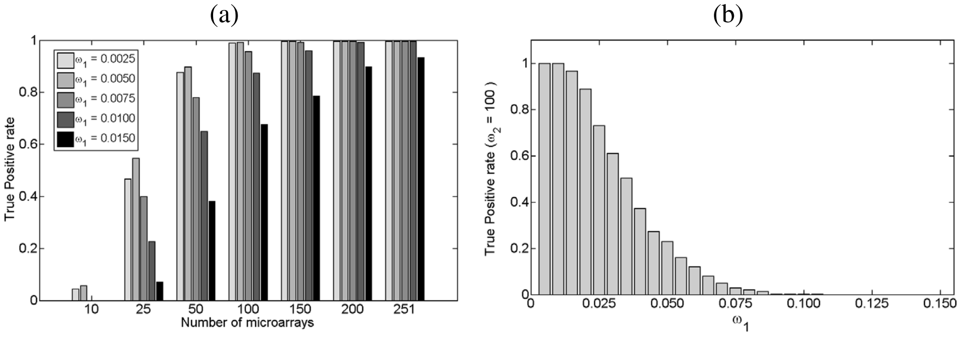

Figure 5 *BFM*: (a): *TP rates for different number of samples and different choices of* ω_1. (b): *TP rates for different choices of* ω_1 *and assuming* 251 *microarrays fixed.*

ω_1 ranging from 0.05 to 0.150; all 251 samples are used in the analysis. Recall that ω_1 controls the variability in the first component of the mixture prior (2) specified for the factor loadings. As can be seen, the largest TP rates are associated with small ω_1s, and again this rate decreases as ω_1 increases. If ω_1 is set to be 0.11, the TP rate is 0%, that is, the model cannot distinguish the two components of the mixture prior, and thus probe sets showing strong pattern will be incorrectly classified as "Absent." In summary, Figure 5 shows that the performance of the BFM depends on the sample size, and the choice of ω_1 can be used to calibrate the model by relaxing or strengthening the assumption of zero factor loadings for a noise probe set.

The BFM is insensitive to the choice of ω_2, in the range of very small values for ω_1 that we are interested in. This result is expected since we work with a relatively large data set (251 microarrays) and this is sufficient to identify variance parameters. Additional graphs similar to Figure 5(b) are presented in Appendix B (Figure B.1), and they show how the factor model behaves, in terms of true positive rates, for other choices of ω_2.

For a different type of cell, the normalization procedure defined in Section 2.1 provides transformed data with variability of intensities similar to those observed in the study of breast cancer developed here. Therefore, the choice of ω_1 and ω_2 does not depend on the type of cell we are examining.

2.5 Comparison of detection methods, using a real data set

In order to compare the performance of detection methods, a data set containing information regarding the true presence/absence of a subset of genes is required. The spike-in study developed by Affymetrix for expression algorithm assessment is an interesting option involving the HG-U133A array (http://www.affymetrix.com/support/technical/sample_data/datasets.affx). According to Affymetrix (see the previous URL): "The data consist of 3 technical replicates of 14 separate hybridizations of 42 spiked transcripts in microarrays for human genome; therefore, the number of arrays available for analysis is 42. Different concentrations, ranging from 0 pM to 512 pM, are used for the spiked transcripts. Four spikes are bacterial controls, eight spikes are artificial sequences believed to be unique in the human genome, and thirty spikes correspond to cDNA clones isolated from total RNAs of a lymphoblast cell line." In other words, the chip contains sequences of 42 genes known to be absent in a non-spiked array, and the hybridization solution contains transcripts from these special genes at different concentrations. An efficient detection method is supposed to identify those 42 spiked genes as "Present," this is the point being evaluated in this subsection. Further details about this data set can be found in the Affymetrix website previously indicated. A similar data set designed for the same purpose has been used by Irizarry et al. (2003a) to evaluate the effectiveness of expression measures produced by MAS 5.0 and RMA.

In the real data analysis, two packages of functions written in the open source statistical language **R** are used. The detection method MAS 5.0 P/A is implemented in "affy" Gautier et al. (2004), and PANP can be applied via "panp." Both packages are integrated into the Bioconductor project (http://www.bioconductor.org), a collaborative effort providing softwares for computational biology and bioinformatics (Gentleman et al., 2004).

Recall that MAS 5.0 P/A and PANP require the specification of two thresholds defining regions for detection calls (present, marginal, absent). The choice of such thresholds is a crucial aspect to be considered in a study comparing three different detection methods. The number of presence/absence calls varies depending on the chosen values. As an example, assume the default thresholds (0.04 and 0.06 in MAS 5.0 P/A, 0.01 and 0.02 in PANP) to analyze the set of 42 microarrays in the spike-in study. These techniques generate detection calls for each array, and this information is summarized in a single list of calls by selecting the most frequent classification for each probe set. MAS 5.0 P/A and PANP detect 46.95% and 32.03% of the probe sets as present, respectively. The BFM applied to the same data set identifies 0.77% of the probe sets as present. Different percentages of presence calls suggest different false positive rates for each method.

In this real application, assume again the prior specifications and MCMC configuration indicated in Section 2.3. The BFM detects 172 probe sets out of 22,300 as "Present." Taking the previous discussion into account, we define thresholds for PANP and MAS 5.0 P/A such that the number of presence calls is close to the result from the BFM. It is not possible to select thresholds providing exactly 172 presence calls for PANP and MAS 5.0 P/A, because some p-values are the same (precision of 4 decimal places) which implies the same classification for a group of probes sets. Therefore, the threshold providing the smallest number of presence calls larger than 172 is chosen for the analysis (0.00180 for MAS 5.0 P/A, and 0.00015 for PANP). The first threshold is the most important, because in the study of the spike-in data interest lies in the presence calls. The second threshold is not considered here, and any probe set with p-value larger than the first threshold will be classified as "not present" combining marginal and absence calls.

Figure 6 compares detection calls for the 42 spiked probe sets. As can be seen, the BFM correctly identifies all 42 cases as "Present,"[3] whereas PANP and MAS 5.0 P/A indicate 16 and 14 presence calls, respectively. This result based on real data reinforces the conclusion of the simulated study in the previous subsection, where the BFM outperforms the other two methods. Figure 7 displays images of two probe sets across samples in the spike-in data. Panel (a) shows a consistent and strong pattern for each probe across samples. The corresponding probe set belongs to the group of 42 spiked-in cases, and it was detected as "Present" by the BFM

[3] The factor model takes into account the co-expression of probes across all arrays to generate the P/A call. The detection call for a certain probe set is not generated based only on a single array hybridized with concentration 0 of the target sequence.

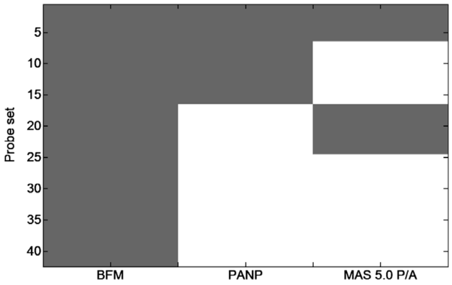

Figure 6 *Detection calls of BFM, PANP and MAS* 5.0 *P/A for* 42 *spiked probe sets. Presence calls (gray) and non-presence calls (white). All* 42 *spiked probe sets are detected as present using BFM, which is a significant improvement over PANP or MAS* 5.0 *P/A.*

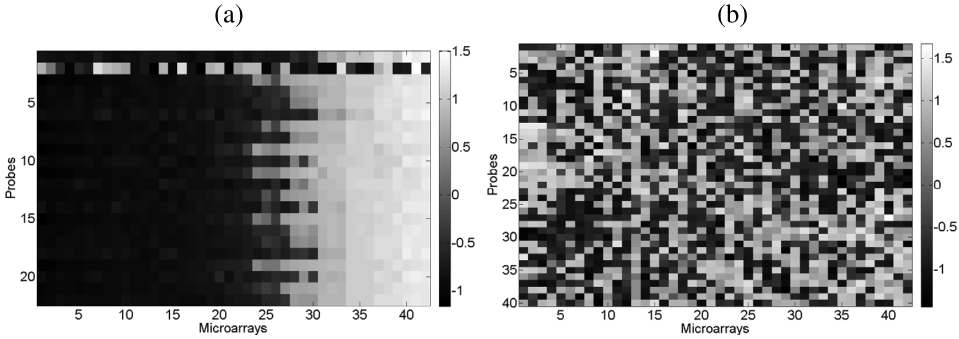

Figure 7 *Spike-in study*: (a) *is an example of a probe set showing a consistent pattern across samples (detection call: BFM = "Present," Other methods = "Absent")*. (b) *is an example of a probe set detected as "Absent" by the BFM.*

and as "Absent" by the other techniques. Panel (b) presents a probe set whose probes randomly alternate intensities across samples. This is a typical "Absent" case correctly identified by the BFM.

3 Factor model to study DNA copy number alteration

The number of copies of a gene in a chromosome can be modified as a consequence of problems during cell division. These alterations play an important role in human cancer. Several methods have been developed for identification of such chromosomal abnormalities; see for example Lai et al. (2005), Diskin et al. (2006), Wieringen et al. (2006) and Rueda and Uriarte (2007). However, none of these methods compare data sets representing different tumors to investigate whether a gene list, detected with CNA in one cancer type, is also representative of CNA in other types of tumors.

We focus on regions of the genome that are known to exhibit CNA, and identify those genes that are showing evidence of it. This is critically important because these are the genes in the region that may be relevant in cancer progression. The genes that do not show coherent expression in the region are evidently unexpressed or otherwise regulated and are therefore less likely to be driver mutations.

Consider seven data sets evaluated in Chin et al. (2006), Miller et al. (2005), Sotiriou et al. (2006), Wang et al. (2005), Bild et al. (2006), Marks et al. (1991) and Freije et al. (2004).[4] The first four studies involve microarrays for breast cancer, and the remaining references are associated with lung, ovarian and brain cancer, respectively. In addition, assume a collection of genes known to be coherently expressed. These co-expressed genes are located altogether in the chromosome, and their locations are known. An annotation file identifying the chromosome location of each probe set can be obtained from the Affymetrix webpage, and it can be used to determine the gene lists investigated in this section. We generate the lists by collecting all genes within a fixed range, defined around some central point representing the location of the group.

The selected genes have been shown to be over-expressed in certain breast cancers due to duplications of their DNA segment (see Pollack et al., 2002; Lucas et al., 2010).[5] Extra copies of the DNA leads directly to a higher concentration of mRNA for those genes through a dosing effect, and thus the measurements in the microarrays are affected. The question driving the present study is: does the same chromosomal duplication occur in other types of cancer cells? In other words, given a chromosomal region that is known to exhibit CNA in breast cancer, can we measure that abnormality in breast cancer gene expression and can we assess whether that same region exhibits CNA in other tumor tissue types? Our study focuses on the effects of CNA over the gene expression pattern across samples. The factor analysis can be used to statistically assess the CNA effect.

3.1 The factor model

Denote X as the $(m \times n)$ matrix containing the RMA output of probe set i in sample j. The RMA is used to preprocess the data analyzed in Section 3 and we

[4] We consider Affymetrix HG-U133A oligonucleotide arrays explored in Chin, Miller, Sotiriou, Wang, Marks and Freije. The experiments to generate signatures use HG U133 2.0 plus arrays (Affymetrix) in Bild. Only the RMA output was available for the dat sets Chin and Bild.

[5] In Lucas et al. (2010) the expression scores of 56 latent factors were assessed on both breast cancer data set as well as breast tumor cell lines. These scores were then compared with CGH clones in the corresponding tumor and cell line samples using Pearson correlation. Approximately, $1/3$ of the factors show a significant degree of association with the CGH clones in small chromosomal regions in both tumor and cell line. We consider an interval around the central point of the chromosomal region where the indicated correlations are significant to select probe sets for our analysis. Note that the breast cancer data is linked to CNAs in specific chromosomal regions through latent factors identified in a factor model.

are no longer looking at the expression of probes. In the scenario without CNA, we expect the expression of probe sets across samples to randomly alternate values due to the activities of pathways that are relevant to each gene individually. In the scenario with chromosomal duplications, the pathway activity is swamped by the impact of CNA, and coherent patterns can be detected. We choose to standardize the rows of X; as a result, we can consider a more parsimonious model without a mean expression parameter μ. The factor analyses with and without μ provide very similar results which makes the parsimonious version more attractive. Consider the one-factor formulation $X = \alpha\lambda + \varepsilon$ shown in (1). The model defined in Section 2 and the current one differ in terms of the mixture prior distribution specified for the factor loadings.

$$\begin{aligned} &\alpha_i \sim (1-h_i)\delta_0(\alpha_i) + h_i N(0,\omega), \\ &h_i \sim \text{Bernoulli}(q) \quad \text{and} \quad q \sim \text{Beta}(\gamma_1,\gamma_2), \end{aligned} \tag{5}$$

where $\delta_0(\alpha_i)$ means $\alpha_i = 0$ with probability 1. Because each row of X represents a different probe set and some genes may exhibit distinct patterns, the mixture prior is specified for each loading i. Note that h_i is a binary latent variable indicating whether $\alpha_i = 0$ or not. The probability q measures the overall level of sparsity in the factor loadings, and we express our uncertainty about this parameter through the Beta distribution. The posterior estimate of q is an interesting measure of coherent patterns in X. Conjugate priors are specified for the remaining parameters: $\lambda_j \sim N(0,1)$, $\varepsilon_{ij} \sim N(0,\sigma_i^2)$ and $\sigma_i^2 \sim \text{IG}(a,b)$. The Gibbs Sampler algorithm is implemented to generate values from $p(\alpha,\lambda,\sigma^2,h,q|X)$. The likelihood function and the full conditional posterior distributions are presented in Appendix C. The conditional probability $p(h_i = 1|\omega,\lambda,\sigma^2,q,X) = q/\{q + (1-q)N[0|M_\alpha,V_\alpha]/N[0|0,\omega]\}$ can be used to measure how strong is the expression pattern of probe set i; see Appendix C to identify M_α and V_α.

3.2 Simulated study

Assume $\omega = 10$ and $\gamma_1 = \gamma_2 = 1$ in (5), $\sigma_i^2 \sim \text{IG}(2.1, 1.1)$. The MCMC algorithm is set to perform 2000 iterations (burn-in period = 1000). The initial values of the chains are $\alpha_i^{(0)} = 0$, $(\sigma_i^2)^{(0)} = 1$, $q^{(0)} = 0.5$, $h_i^{(0)} \sim \text{Bernoulli}(0.5)$ and $\lambda_j^{(0)} \sim N(0,1)$. Fast convergence to the target distribution is observed. The procedure to generate the data is simple: we select a real breast cancer data set (Miller et al., 2005) representing the expressions of 23 probe sets across 251 samples. The model in Section 3.1 is fitted to the real data and the posterior estimates of α, λ and σ^2 are assumed as the real values. In short, our procedure to simulate data is a reconstruction of X with its most important features. We generate ε_{ij} from $N(0,\sigma_i^2)$ and compute $\alpha\lambda + \varepsilon$ to obtain the synthetic X.

The conditional posterior distribution of α_i is a mixture of $\delta_0(\alpha_i)$ and a Gaussian component. Figure 8(a) suggests a good performance of the proposed model; in

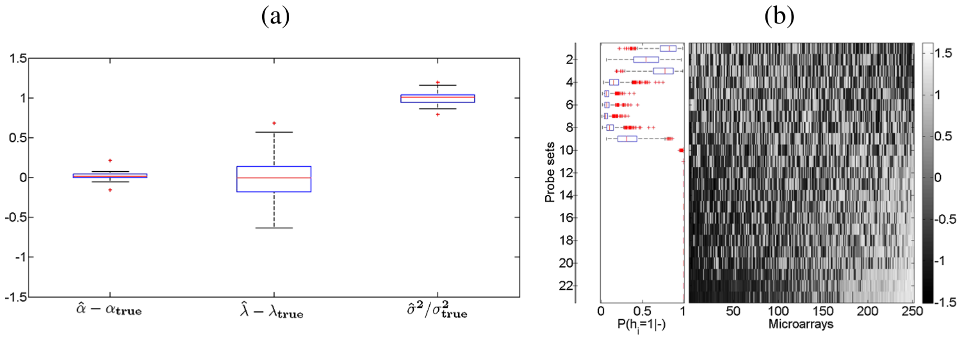

Figure 8 *Panel* (a): *box plots summarizing the behavior of the bias* (*for* α_i *and* λ_j) *or the ratio* (*for* σ_i^2) *involving the posterior estimates and the true value of the parameters. Panel* (b): *box plots indicating the distribution of* $p(h_i = 1|\cdots)$, *and the image of the simulated* X *with columns sorted so that the* 1*st principal component is monotone.*

most cases the true value is close to the posterior mean and located inside the 95% credible interval. Panel (b) shows the heat map image of the simulated data and box plots representing the posterior distribution of $p(h_i = 1|\cdots)$. Note that the box plots concentrate the probability mass around 1 for some probe sets, which means that these cases have higher posterior probability that $\alpha_i \neq 0$, and it suggests a coherent pattern across samples. Some probe sets located in the top have box plots more dispersed in the unit interval. In particular, the box plots in rows 4–8 indicate probability mass below 0.5. The expression patterns observed in the image are in accordance with the interpretation obtained from the box plots.

3.3 Real data analysis

Here, we investigate the CNA effect over the gene expression across samples of 4 different types of cancer. Again, CNA is known to occur in some chromosome regions for breast tumors (see Pollack et al., 2002; Lucas et al., 2010); we identify the group of genes located in those regions, and then examine the expressions of these genes for other types of cancer. The same priors and MCMC configurations described in Section 3.2 are applied in this study; fast convergence to the target distribution is observed.

Figures 9 and 10 show images of matrices X with rows representing probe sets. We apply the factor model described in Section 3.1 to each data set, and generate box plots representing the distribution of $p(h_i|\cdots)$. In addition, we perform the factor analysis for probe-level data using the model (1). The presence/absence calls for each probe set are presented in the color bar displayed to the left of some panels; the probe-level data are not available for some data sets.

Note that few probe sets are detect as “absent” in the breast cancer data sets. This result is expected since the genes were selected in a region of the genome where CNA occurs for breast cancer. In the ovarian case, most probe sets are detected

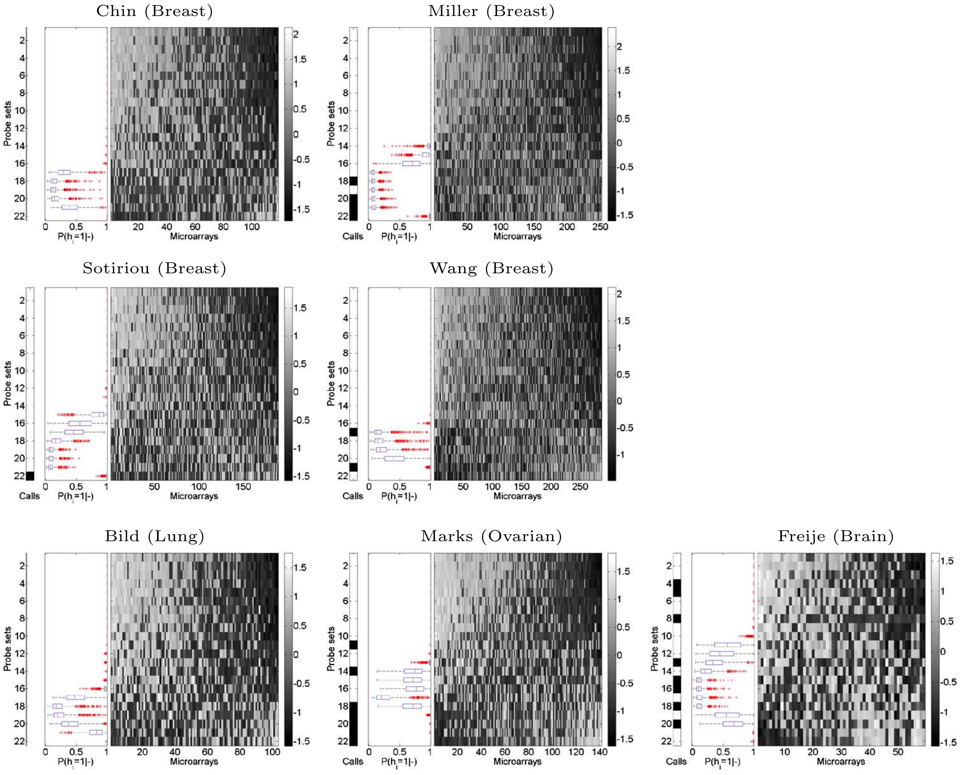

Figure 9 *Probe-level analysis with model* (1): *Presence* (*white*) *and Absence* (*black*) *calls are shown on the left of some images*. *Gene-level analysis with model* (5): *Box plots representing the distribution of* $p(h_i = 1|\cdots)$ *are shown in the middle of each panel*. *All panels investigate the same gene list located between positions* 108,000,000 *and* 112,000,000 *in Chromosome* 6. *Columns are sorted so that the* 1*st principal component is monotone*.

as “Present”; however, the number of absence calls increases a bit as compared to the breast cancer. The brain cancer indicates distinct results, that is, a majority of presence calls in Figure 9 and a majority of absence calls in Figure 10. In most cases, presence calls seem associated with a row displaying increasing or decreasing patterns in the image, and absence calls seem associated with random patterns.

The box plots in Figures 9 and 10 show high probability mass above 0.5 for most probe sets in almost all panels. Note that coherent patterns in the image graph are related to box plots located above 0.5, and random patterns correspond to dispersed box plots centered below 0.5. The brain cancer panel in Figure 10 is the case showing the largest number of box plots not concentrated around 1.

The box plots in Figure 11 represent the posterior distribution of the probability q specified in (5). As can be seen, the graph for the brain cancer (Figure 10) is the only one suggesting high probability mass below 0.5. This result indicates that the sparsity level in the factor loadings vector is high, and thus the CNA can be

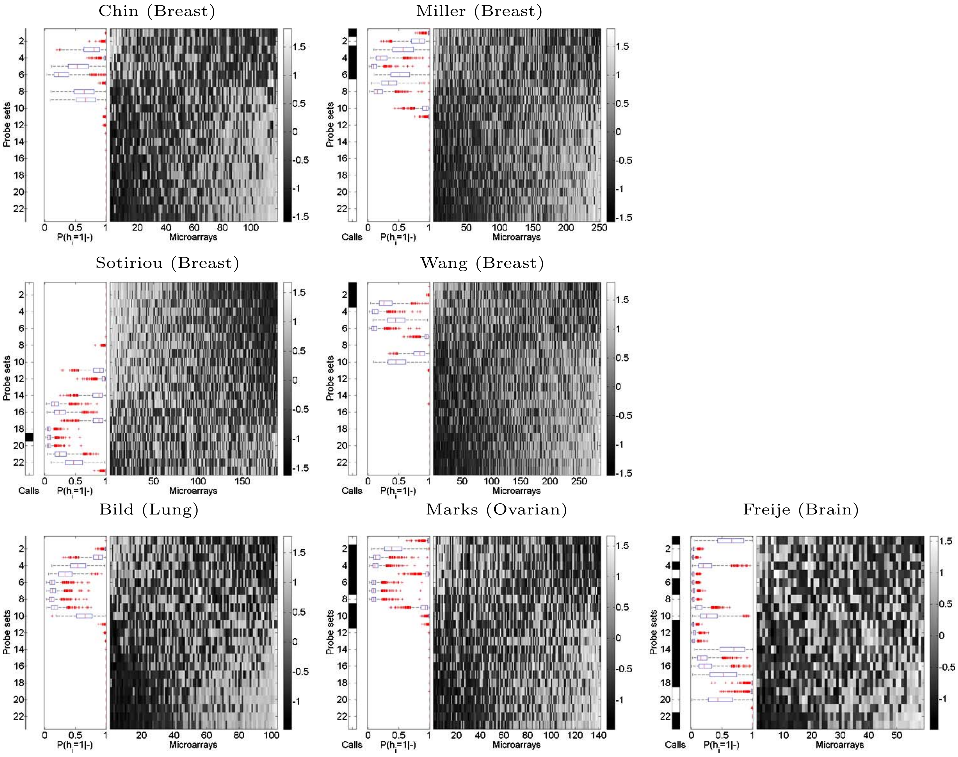

Figure 10 *Probe-level analysis with model* (1): *Presence* (*white*) *and Absence* (*black*) *calls are shown on the left of some images. Gene-level analysis with model* (5): *Box plots representing the distribution of* $p(h_i \mid \cdots)$ *are shown in the middle of each panel. All panels investigate the same gene list located between positions* 208,000,000 *and* 212,000,000 *in Chromosome* 1. *Columns are sorted so that the* 1*st principal component is monotone.*

considered absent. All other data sets have box plots suggesting: low sparsity level in α, strong coherent pattern for most probe sets, and presence of CNA.

In Figure 12, we compare the behavior of the factor models (1) and (5). As can be seen, the posterior estimates for the breast cancer data are similar for both models; several intervals do not contain the value 0, and the posterior means are located in similar positions. The posterior estimates for α reflect the absence of CNA effect identified for brain cancer in Figure 11. Note that the model in Column 1 provides intervals centered around zero, and the model in Column 2 estimates several α_i as zero.

We have applied the model (1) to the data sets X shown in Figures 9 and 10. The results are in accordance with the interpretation of the box plots in Figure 11, that is, the only case associated with a small probability q^* in (3) is the brain cancer with genes from Chromosome 1. A small q^* suggests no CNA effect in that case.

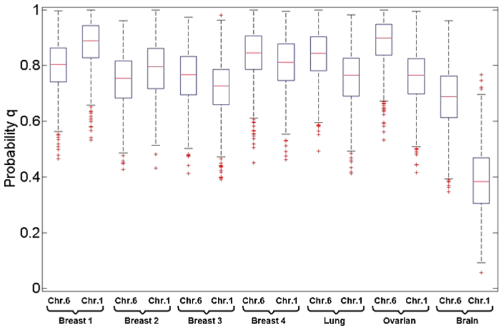

Figure 11 *Box plots comparing the posterior distribution of q for all data sets.*

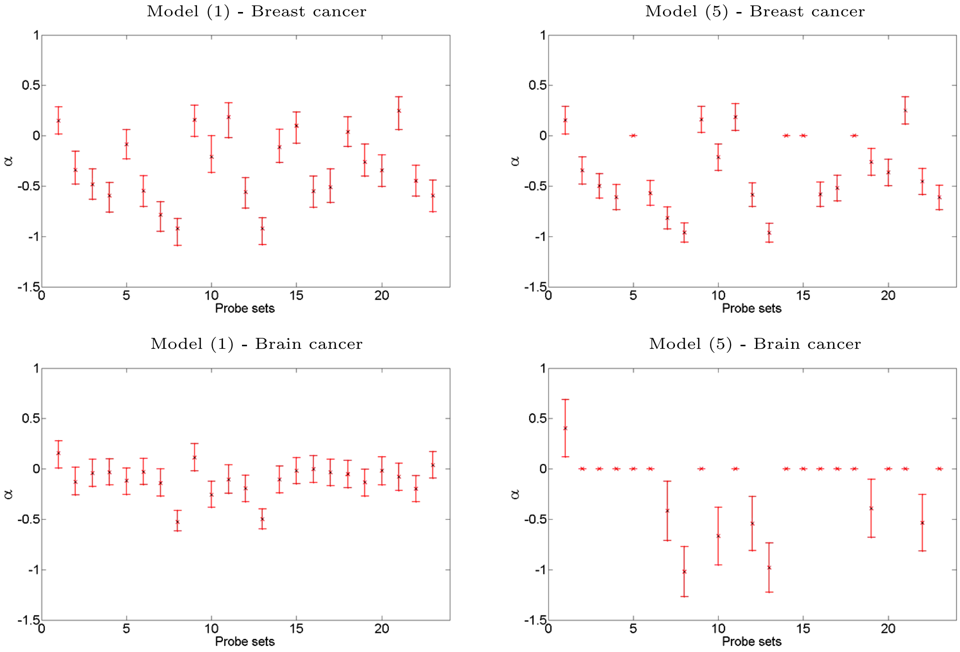

Figure 12 *Posterior mean* (*x mark*) *and* 95% *credible interval* (*bar*) *for* α. *Gene list identified in Chromosome* 1.

4 Conclusions

In this paper, factor models are proposed to evaluate the gene expression of Affymetrix oligonucleotide microarrays. Our first study is focused on detection calls to examine the presence/absence status of probe sets targeting genes. The factor model (BFM) evaluates whether the probes within a probe set exhibit a consistent pattern of intensities across arrays. In a simulated study to investigate the inference performance, the factor model provides good approximation between posterior estimates and real values of the parameters. In a second simulated study, we compare the BFM with two other detection methods suggested in the literature. The main conclusion was obtained from ROC curves comparing the methods in terms of true positive and false positive rates. The BFM indicates the best combination of high true positive and low false positive rates. However, the performance of the proposed factor model depends on the number of microarrays available for analysis. The smaller the sample size, the lower the TP rate. The study suggests that the BFM is preferred to PANP and MAS 5.0 P/A, particularly if a large number of samples is available. We have also considered a real data set related to a spike-in study to compare the detection methods. In the designed experiment, transcripts from 42 spiked genes were known to be present in the arrays. The BFM correctly detected the presence of all spikes, whereas PANP and MAS 5.0 P/A indicated absence call for some cases.

The second application presented in this paper involves a factor model for probe set expression data instead of probe-level data. Chromosomal duplications affecting the number of transcripts of genes located in a specific region of the genome are known to occur in breast cancer tumors. The main idea of this study was to examine copy number changes for the same group of genes in other types of cancer. Images of probe sets across arrays can be inspected for visual identification of patterns suggesting CNA; however, the factor analysis provides a more accurate answer than the visual inspection. The rows of X represent genes, and different genes may exhibit distinct expression patterns; therefore, we have proposed in Section 3 a factor model with a mixture prior for each loading α_i. We induce sparsity assuming $\delta_0(\alpha_i)$ as one of the components of the mixture, and a Beta distribution is used to express our uncertainty about the probability of $\alpha_i \neq 0$. In a simulated study to verify performance, we have concluded that the model can estimate well the true values of the parameters. Next, we have analyzed real data for different types of cancer. The main conclusions are obtained from box plots representing the posterior distributions of $p(h_i = 1|\cdots)$ and q. There is association between the expression pattern, displayed in the heat map image, and the position of the box plot for $p(h_i = 1|\cdots)$. The probability mass is concentrated above 0.5 for coherent patterns, and below 0.5 for random patterns. The posterior estimate of q indicates the overall level of sparsity in the factor loadings. We use this result to measure our posterior uncertainty about the presence/absence of CNA. The CNA effect was detected for almost all data sets, the only exception was the brain cancer with genes from Chromosome 1.

Appendix A: Analysis of high-throughput sequencing data

RNA-Seq is a promising new technology to measure gene expression. Its main steps are (i) RNA's are isolated from a sample and converted to cDNA fragments,

(ii) a high-throughput sequencer is used to generate millions of reads (short nucleotide sequences) from the cDNA fragments, (iii) an alignment tool is used to map the reads to a reference genome, and (iv) counts of reads mapped to each gene are used to estimate expression levels. Because the outputs of RNA-Seq are counts, they are referred to as "digital" gene expression, as opposed to the "analog" fluorescence intensities from microarrays. Although the technology is still young, its output data have been analyzed in several scientific publications (e.g., Marioni et al., 2008, Mortazavi et al., 2008 and Wang et al., 2009). RNA-Seq data have some advantages over microarrays, such as low background noise, an ability to detect novel transcripts, and the requirement of less RNA samples. On the other hand, some experimental challenges must be addressed, such as read mapping uncertainty. The reads are much shorter than the transcripts from which they are derived, and there is the possibility that a single read may map to multiple genes, complicating the expression analysis. Different approaches have been proposed in the literature to deal with this problem; one might discard reads that map to multiple locations (Marioni et al., 2008), allocate the reads to genes heuristically (Faulkner et al., 2008), assume a statistical model with latent variables representing the true mappings (Li et al., 2010).

In this section, we apply the BFM in (4) to analyze the coherent pattern across samples of RNA-Seq data, and then determine the presence/absence status of the corresponding gene. This strategy can be an interesting approach for the read mapping uncertainty issue, where a random pattern is potentially observed for a gene with a large number of incorrectly mapped sequences. On the other hand, a large number of nucleic acid sequences targeting the correct gene will contribute to a strong pattern across samples.

The expression data are obtained via the high-throughput sequencing system Illumina Genome Analyzer 2. This sequencer generates images of size approximately 1 Tb. The data are then processed and passed through a quality control where bad reads with a chastity score lower than 0.2 are discarded. Next, the remaining reads with a quality score larger than 15 are consolidated into a frequency. Finally, the reads are mapped to the human genome wherever possible. The data set is composed of 32 samples related to Ovarian tumors (source: Harvard Medical School; access: The Cancer Genome Atlas data portal, http://tcga-data.nci.nih.gov/tcga/).

For each sample, the data set contains a list of sequences (17 nucleotides in length), the total number of reads per sequence and the associated gene symbol. Each nucleic acid sequence has a single entry in the list, and more than one sequence may target the same gene. The number of nucleotide sequences and the number of identified genes may vary between samples. The total number of genes in the union of all samples is 44,320; however, only 16,082 cases can be found in the intersection. Our study is focused on those genes in the intersection. For each gene, any row of matrix X represents a DNA sequence targeting that gene in at least one sample. In general it is possible that, as a consequence of read mapping uncertainty, the sequence belongs to another gene.

Of the 16,082 genes that we consider, 852 are represented by a single nucleotide sequence (i.e., X is a n-dimensional row vector). In addition, 1634 genes are associated with more than 100 sequences, most of which are identified in only 1 sample. Neither of these situations provide significant information regarding coherence of expression. With this in mind, we filter the 16,082 genes by selecting those cases with corresponding matrix X having more than 20 rows with very few rows full of zeros (criterion: at least 70% of the DNA sequences detected in at least 5 samples). It is important to highlight that the focus of this application is to show how the analysis of coherent patterns across samples can be useful to handle the presence/absence detection problem for a subset of genes identified in RNA-Seq data.

The normalization procedure described in Section 2.1 to remove background noise in microarray data is not required in the analysis of RNA-Seq data. The counts of reads are not significantly affected by background noise; therefore, no transformation is applied to the values in X. Figure A.1 shows heat maps represent-

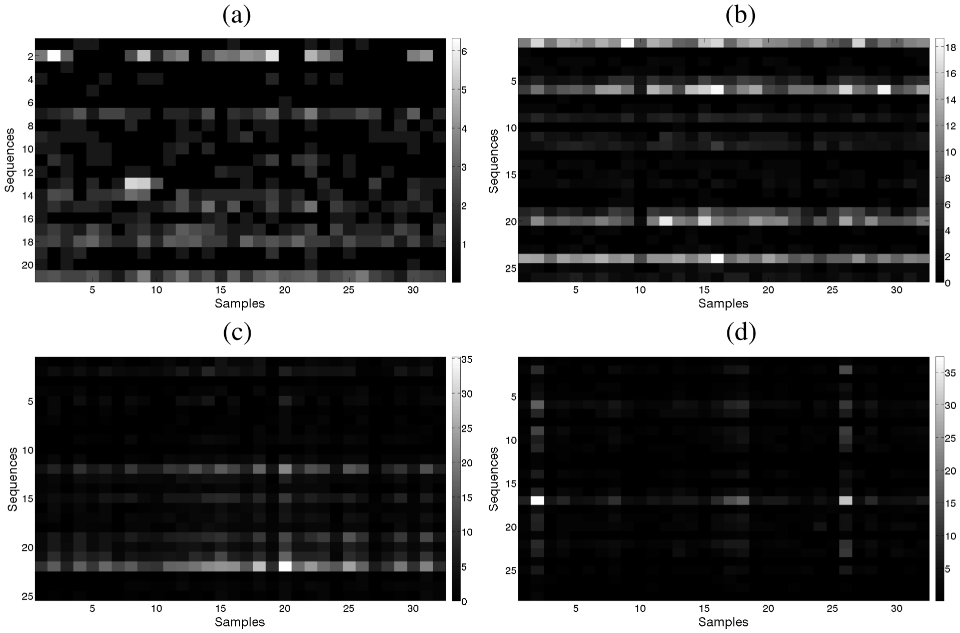

Figure A.1 *Count of reads for each sequence across samples. The sequences target gene "C6orf162" (chr. 6) in* (a), *"TMED4" (chr. 7) in* (b), *"ARL4C" (chr. 2) in* (c), *and "PHLDA1" (chr. 12) in* (d). *The square root of the original count data are displayed (columns are not sorted and rows are not standardized).*

ing matrices X for 4 different genes. As can be seen, the magnitude of the count values can be different between rows. Note that the image displayed in panel (b) indicates rows 1, 6, 20 and 24 containing the largest values in that matrix. Given this difference between rows, it seems appropriate to introduce the mean expression parameter μ in the model. We use the model specification (4), and assume the prior distribution $\mu \sim N_m(\hat{\mu}, 100I_m)$ with $\hat{\mu}$ being a m-dimensional column vector containing the minimum value of each row of X. In addition, we consider the same prior specifications defined in Section 2.3 for α, λ and σ_i^2. The MCMC algorithm is set to perform 2000 iterations (burn in period = 1000). The initial values of parameters are the same as those defined in Section 2.3; we also set $\mu^{(0)} = \mathbf{1}$.

Consider Figure A.2 where the rows of the selected matrices are standardized and the columns are ordered. Note that, panels (a) and (b) show images suggesting a more random pattern across samples; whereas panels (c) and (d) present cases where the count of reads indicates a strong coherent pattern. We fit the BFM to the four matrices displayed in Figure A.1 and the result confirms the visual interpretation of Figure A.2, that is, the "Absent" call is obtained for the genes represented in panels (a) and (b), and the "Present" call is determined for the genes in panels (c) and (d).

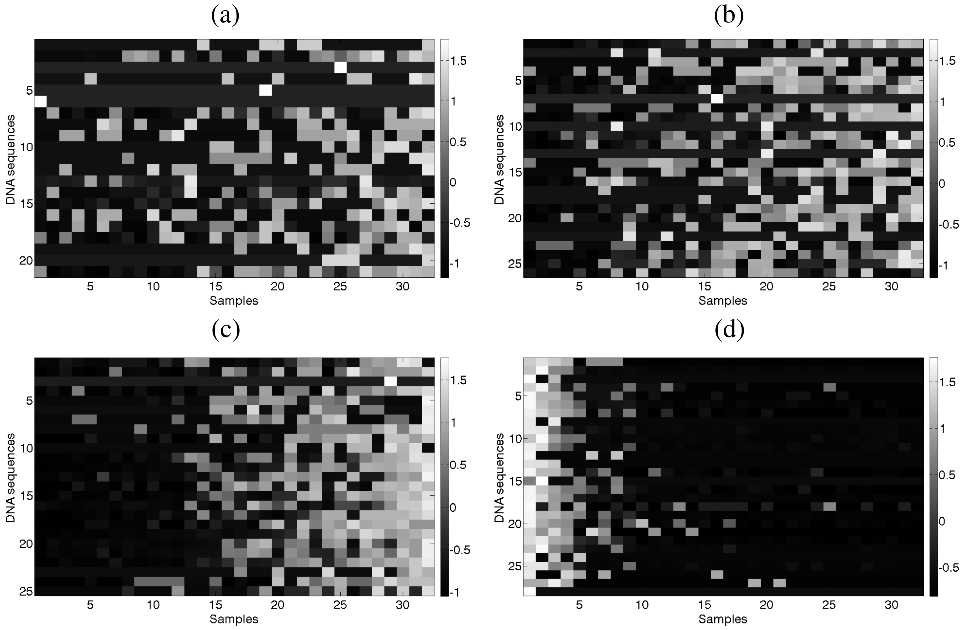

Figure A.2 *Count of reads for each sequence across samples. The sequences target the gene "C6orf162" in* (a), *"TMED4" in* (b), *"ARL4C" in* (c), *and "PHLDA1" in* (d). *In order to improve the expression pattern visualization, the rows are standardized and the columns are sorted in the direction of the first principal component.*

The Basic Local Alignment Search Tool (BLAST) is a program that can be used to identify nucleotide sequences; see Altschul et al. (1990) and http://blast.ncbi.nlm.nih.gov/Blast.cgi. The algorithm verifies whether an input sequence has similarities with other sequences in a public database. In fact, BLAST is a group of programs available for different types of query sequences, and in this application, we are interested in the BLAST search for short DNA sequences (17 nucleotides) associated with each row of matrix X. Therefore, we consider the "blastn" program to search a nucleotide database using a nucleotide query. The goal of this analysis is to verify whether the RNA-Seq data and BLAST identify the same gene for the sequences used for the genes in Figure A.2. The BLAST output[6] for a particular sequence is a list of the genes which contain nucleotide sequences that resemble the input sequence above certain threshold.

In summary, the main aspects of the current study are: (i) A fast alignment tool (not BLAST) is used to map sequence reads to a gene, (ii) BLAST—a more reliable tool—is applied to identify the target gene for each sequence read (both methods may agree or not), (iii) The BFM is applied to evaluate the relation between detection calls and the number of agreements between BLAST and the fast alignment tool.

We performed a BLAST search for every sequence (row) of the matrices presented in Figure A.2; the results are reported in Table A.1. Note that the genes classified as "Present" through the BFM have a large number of sequences showing agreement between BLAST and the RNA-Seq gene identification. In particular, this agreement result is observed for all sequences related to gene "ARL4C" in panel (c) and 25 of 28 for panel (d). On the other hand, the genes with detection call "Absent" indicate a large number of disagreement cases, that is, the gene suggested in the data set was not found in the BLAST search. The disagreement between BLAST and the data set gene identification can be considered an indication of incorrect sequence mapping. The analysis of the coherent pattern across samples via the BFM seems a reasonable strategy for detection of genes containing a relatively large number of sequences with incorrect mapping. Only four genes were analyzed in this section, but we have obtained similar conclusions from other cases not described here.

We assume the Gaussian distribution in this analysis; however, a more natural choice for count data would be the Poisson or Negative Binomial distributions. The Negative Binomial has two parameters, and the second parameter can be used to adjust the variance independently of the mean, which makes this distribution useful for cases of overdispersed data. Another alternative is the Generalized Poisson (GPois): $p(Y = y|\psi, \tau) = [(\psi/y!)/(\psi + y\tau)^{y-1}]\exp\{-\psi - y\tau\}$, where $\psi > 0$, $0 \leq \tau \leq 1$, and $y = 0, 1, 2, \ldots$; $(\tau = 0) \Rightarrow \text{Poisson}(\psi)$. It can be shown that $E(Y|\psi, \tau) = \psi(1 - \tau)^{-1}$ and $\text{Var}(Y|\psi, \tau) = \psi(1 - \tau)^{-3}$. Consider

[6] BLAST is a computationally expensive, but very high fidelity technique for mapping short nucleotide sequences back to the transcriptome. Its results are more reliable than the fast alignment tools used to map millions of sequence reads in RNA-Seq data.

Table A.1 *Number of DNA sequences where the gene identification from the RNA-Seq data is confirmed via BLAST*

Gene symbol	BFM P/A call	# of agreements	# of sequences
C6orf162	Absent	13	21
TMED4	Absent	8	26
ARL4C	Present	25	25
PHLDA1	Present	25	28

the result $\mathrm{Var}(Y|\psi,\tau)/E(Y|\psi,\tau) = (1-\tau)^{-2} \geq 1$, i.e., the GPois is suitable for count data with sample variance considerably larger than the sample mean. In a Poisson outlier context, one could assume the following mixture model: $X_{ij} \sim \mathrm{GPois}(\psi_{ij}, z_{ij}\tau)$ with $z_{ij} \sim \mathrm{Bernoulli}(\xi_i)$ and $0 < \tau \leq 1$. Factor loadings and scores can be associated with $\log(\psi_{ij})$ in a hierarchical structure. Let $\xi_i = 0$ to define a model for standard Poisson observations. We can develop a fully Bayesian analysis via Gibbs Sampler for this model. The full conditional distribution of τ is log-concave with respect to its arguments; therefore, the Adaptive Rejection Sampling algorithm can be applied.

Appendix B

Here, we present the full conditional posterior distributions associated with the model $X = \mu 1_n + \alpha\lambda + \varepsilon$ in (4). All prior specifications are defined in Section 2.2, except $\mu \sim N_m(\mu_0, \Sigma)$. Because we standardize the rows of X, the model (1) does not contain μ. The formulations can be easily adapted to that case by letting $\mu = \mathbf{0}$. Define $X_{\cdot j}$ as the m-dimensional column vector representing the jth column of X. Note that $(X_{\cdot j}|\mu,\alpha,\lambda,\sigma^2) \sim N_m[\mu + \alpha\lambda_j, D]$ with $D = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$. We assume conditional independence between samples; therefore, $p(X|\mu,\alpha,\lambda,\sigma^2) = \prod_{j=1}^n p(X_{\cdot j}|\mu,\alpha,\lambda,\sigma^2)$. The Bayes theorem provides:

- $(\mu|\alpha,\lambda,\sigma^2,X) \sim N_m(M_\mu, V_\mu)$ with $V_\mu = [nD^{-1} + \Sigma^{-1}]^{-1}$ and $M_\mu = V_\mu[\Sigma^{-1}\mu_0 + D^{-1}\sum_{j=1}^n (X_j - \alpha\lambda_j)]$.
- $(\alpha|\mu,\lambda,\sigma^2,X) \sim (1-q^*)N_m(M_1^*, \Phi_1^*) + q^* N_m(M_2^*, \Phi_2^*)$ where $\Phi_l^* = [\Phi_l^{-1} + D^{-1}\lambda\lambda']^{-1}$ and $M_l^* = \Phi_l^*[D^{-1}(X - \mu 1_n)\lambda']$ for $l \in \{1,2\}$.
- $q^* = (\frac{N[0|0,\Phi_2]}{N[0|M_2^*,\Phi_2^*]}q)/(\frac{N[0|0,\Phi_2]}{N[0|M_2^*,\Phi_2^*]}q + \frac{N[0|0,\Phi_1]}{N[0|M_1^*,\Phi_1^*]}(1-q))$.
- $(\lambda|\mu,\alpha,\sigma^2,X) \sim N(M_\lambda, V_\lambda)$ with $V_\lambda = (\alpha' D^{-1}\alpha + 1)^{-1} I_n$ and $M_\lambda = V_\lambda[(X - \mu 1_n)' D^{-1}\alpha]$.
- Denote $\sigma_{-i}^2 = (\sigma_1^2, \ldots, \sigma_{i-1}^2, \sigma_{i+1}^2, \ldots, \sigma_m^2)$. $(\sigma_i^2|\mu,\alpha,\lambda,\sigma_{-i}^2,X) \sim \mathrm{IG}[a + (n/2), b + B]$ with $B = \frac{1}{2}[\sum_{j=1}^n X_{ij}^2 - 2\mu_i \sum_{j=1}^n X_{ij} - 2\alpha_i \sum_{j=1}^n \lambda_j X_{ij} + \alpha_i^2 \sum_{j=1}^n \lambda_j^2 + 2\mu_i\alpha_i \sum_{j=1}^n \lambda_j + n\mu_i^2]$.

Figure B.1 shows additional graphs complementing the analysis of Figure 5(b).

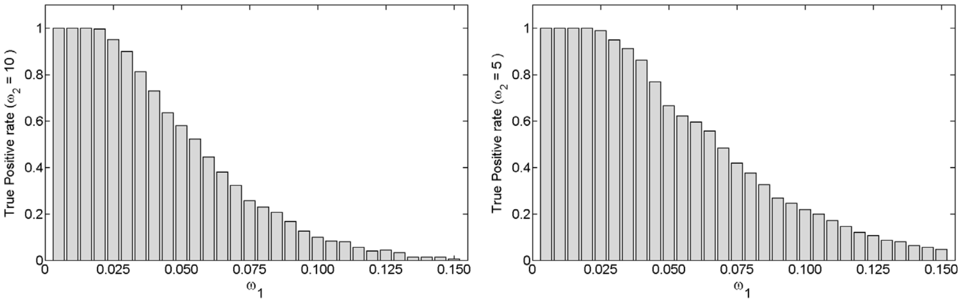

Figure B.1 *True positive rates obtained via BFM for different choices of* ω_1 (*assume* 251 *microarrays*). *Panel* 1 ($\omega_2 = 10$) *and panel* 2 ($\omega_2 = 5$).

Appendix C

Consider $X = \alpha\lambda + \varepsilon$, where α is $(m \times 1)$ and λ is $(1 \times n)$. We can write the likelihood function in two equivalent forms:

- Likelihood 1: Let $X_{\cdot j}$ represent the column j of X, and λ_j is the jth element of λ. Note that $(X_{\cdot j}|\alpha, \lambda_j, \sigma^2) \sim N_m[\alpha\lambda_j, D]$. Assume conditional independence between samples: $p(X|\alpha, \lambda, \sigma^2) = \prod_{j=1}^n p(X_{\cdot j}|\alpha, \lambda, \sigma^2)$.
- Likelihood 2: Let $X_{i\cdot}$ represent the row i of X, and α_i is the ith element of α. Note that $(X'_{i\cdot}|\alpha_i, \lambda, \sigma_i^2) \sim N_n[\lambda'\alpha_i, \sigma_i^2 I_n]$. Assume conditional independence between rows of X: $p(X|\alpha, \lambda, \sigma^2) = \prod_{i=1}^m p(X_{i\cdot}|\alpha, \lambda, \sigma^2)$.

The posterior computation for each parameter can be simplified by choosing the appropriate version of the likelihood. Denote $h = (h_1, \ldots, h_m)'$, $\alpha_{-i} = (\alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_m)'$ and $\lambda_{-j} = (\lambda_1, \ldots, \lambda_{j-1}, \lambda_{j+1}, \ldots, \lambda_n)$. The Bayes theorem provides:

- If $h_i = 1$, $(\alpha_i|\alpha_{-i}, \lambda, \sigma^2, h, X) \sim N(M_\alpha, V_\alpha)$ with $V_\alpha = [\frac{1}{\omega} + \frac{1}{\sigma_i^2}\sum_{j=1}^n \lambda_j^2]^{-1}$ and $M_\alpha = V_\alpha[\frac{1}{\sigma_i^2}\sum_{j=1}^n \lambda_j X_{ij}]$.
- If $h_i = 0$, the full conditional posterior of α_i is $\delta_0(\alpha_i)$.
- $(q|h) \sim \text{Beta}(\gamma_1 + \sum_{i=1}^m h_i, \gamma_2 + m - \sum_{i=1}^m h_i)$.
- $p(h_i = 1|\alpha, \lambda, \sigma^2, q, X) = q/\{q + (1-q)N[0|M_\alpha, V_\alpha]/N[0|0, \omega]\}$.
- $(\lambda_j|\alpha, \lambda_{-j}, \sigma^2, X) \sim N(M_\lambda, V_\lambda)$ with $V_\lambda = (\alpha' D^{-1}\alpha + 1)^{-1}$ and $M_\lambda = V_\lambda[\alpha' D^{-1} X_{\cdot j}]$.
- $(\sigma_i^2|\alpha, \lambda, \sigma_{-i}^2, X) \sim \text{IG}[a + (n/2), b + B]$ with $B = \frac{1}{2}[X_{i\cdot}X'_{i\cdot} - 2\alpha_i\lambda X'_{i\cdot} + \alpha_i\lambda\lambda'\alpha'_i]$.

Appendix D

Figure D.1 shows three matrices displaying the preprocessed data obtained from the RMA background correction and quantile normalization. The RMA back-

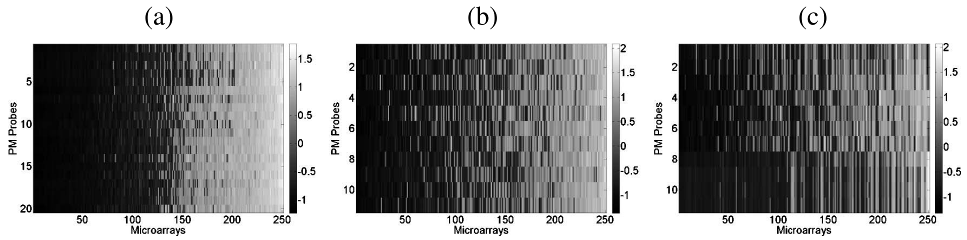

Figure D.1 *Intensities of PM probes within three different probe sets. Samples are along the x-axis and PM probes are on the y-axis. The data are preprocessed through the RMA steps of background adjustment and quantile normalization. Panels* (a) *and* (b) *correspond to the cases in Figure* 1(a) *and* (b), *respectively. The columns are sorted so that the* 1*st principal component is monotone.*

ground adjustment is designed for PM probes only; therefore, the MM probes are excluded from the analysis, and thus the number of rows of matrix X is reduced.

In Section 2, our study includes the MM probes. We could have deleted the MM probes from the study, however, we found that, even though the MM probes are often lower in expression level, that expression change is consistent and they tend to show similar expression patterns across samples as the PM probes. We therefore believe that the MM probes offer increased information for model fitting and inference.

As opposed to RMA, the preprocessing procedure defined in Section 2.1 can be used with PM and MM probes. We have observed coherent patterns for most probe sets, including the three cases in Figure D.1. It is common sense that some probe sets represent just noise effects; therefore, the large number of cases indicating coherent patterns might suggest that the RMA does not subtract very well the overall brightness from the chips. We apply the factor model (1) to the data sets in Figure D.1 assuming the same configuration of priors and MCMC indicated in Section 2.3. The detection call “Presence” is obtained for the three cases. In order to distinguish between weak and strong coherent patterns, one could try to increase a little bit the choice of ω_1 in (2).

Acknowledgments

This study was developed in the Ph.D. program of the Department of Statistical Science at Duke University where the first author received his Ph.D. degree. This study was supported by Award Number U54CA112952 from the National Cancer Institute, by funding from the Measurement to Understand Re-Classification of Disease of Cabarrus and Kannapolis (MURDOCK) Study and by a grant from the NIH CTSA (Clinical and Translational Science Award) 1UL1RR024128-01 to Duke University. The authors would like to thank the anonymous referee for constructive comments.

References

Affymetrix Technical Report (2001). Statistical algorithms reference guide. Available at http://www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf.

Affymetrix Technical Report (2005). Exon array background correction. Available at http://www.affymetrix.com/support/technical/whitepapers/exon_background_correction_whitepaper.pdf.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.

Archer, K. J. and Reese, S. E. (2009). Detection call algorithms for high-throughput gene expression microarray data. *Briefings in Bioinformatics* **2**, 244–252.

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., Joshi, M. B., Harpole, D., Lancaster, J. M., Berchuck, A., Olson, J. A. Jr, Marks, J. R., Dressman, H. K., West, M. and Nevins, J. R. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357.

Boulesteix, A. L. and Strimmer, K. (2006). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics* **8**, 32–44.

Brunet, J. P., Tamayo, P., Golub, T. R. and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4164–4169.

Carvalho, C., Chang, J., Lucas, J., Nevins, J. R., Wang, Q. and West, M. (2008). High-dimensional sparse factor modelling: Applications in gene expression genomics. *Journal of the American Statistical Association* **103**, 1438–1456. MR2655722

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W. L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B. M., Esserman, L., Albertson, D. G., Waldman, F. M. and Gray, J. W. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529–541.

Diskin, S. J., Eck, T., Greshock, J., Mosse, Y. P., Naylor, T., Stoeckert, C. J., Weber, B. L., Maris, J. M. and Grant, G. R. (2006). STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Research* **16**, 1149–1158.

Faulkner, G. J., Forrest, A. R., Chalk, A. M., Schroder, K., Hayashizaki, Y., Carninci, P., Hume, D. A. and Grimmond, S. M. (2008). A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* **91**, 281–288.

Freije, W. A., Castro-Vargas, F. E., Fang, Z., Horvath, S., Cloughesy, T., Liau, L. M., Mischel, P. S. and Nelson, S. F. (2004). Gene expression profiling of gliomas strongly predicts survival. *Cancer Research* **64**, 6503–6510.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd ed. *Texts in Statistical Science* **68**. Boca Raton, FL: Chapman & Hall/CRC. MR2260716

Gautier, L., Cope, L., Bolstad, B. M. and Irizarry, R. A. (2004). Affy—Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. and Speed, T. P. (2003a). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**, e15.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

Kapur, K., Xing, Y., Ouyang, Z. and Wong, W. (2007). Exon arrays provide accurate assessments of gene expression. *Genome Biology* **8**, R82.

Kim, P. M. and Tidor, B. (2003). Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Research* **13**, 1706–1718.

Lai, W. R., Johnson, M. D., Kucherlapati, R. and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763–3770.

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. and Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biology* **2**, R32.

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

Liu, X., Milo, M., Lawrence, N. D. and Rattray, M. (2005). A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. *Bioinformatics* **21**, 3637–3644.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41–67. MR2036762

Lucas, J. E., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R. and West, M. (2006). Sparse statistical modelling in gene expression genomics. In *Bayesian Inference for Gene Expression and Proteomics* (P. Muller, K. Do and M. Vannucci, eds.) 155–176. Cambridge: Cambridge University Press. MR2269095

Lucas, J. E., Kung, H. N. and Chi, J. T. (2010). Cross-study projections of genomic biomarkers: An evaluation in cancer genomics. *PLoS Computational Biology* **6**, e1000920. DOI:10.1371/journal.pcbi.1000920.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18**, 1509–1517.

Marks, J. R., Davidoff, A. M., Kerns, B. J., Humphrey, P. A., Pence, J. C., Dodge, R. K., Clarke-Pearson, D. L., Iglehart, J. D., Bast, R. C. and Berchuck, A. (1991). Overexpression and mutation of p53 in epithelial ovarian cancer. *Cancer Research* **51**, 2979–2984.

McClintick, J. N. and Edenberg, H. J. (2006). Effects of filtering by present call on analysis of microarray experiments. *BMC Bioinformatics* **7**, 49.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T. and Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 13550–13555.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628.

Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**, 39–50.

Ouandaogo, Z. G., Haouzi, D., Assou, S., Dechaud, H., Kadoch, I. J., Vos, J. D. and Hamamah, S. (2011). Human cumulus cells molecular signature in relation to oocyte nuclear maturity stage. *PLoS ONE* **6**, e27179.

Pollack, J. R., Sorlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Dale, A. L. B. and Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12963–12968.

Rueda, O. M. and Uriarte, R. D. (2007). Flexible and accurate detection of genomic copy-number changes from aCGH. *PLoS Computational Biology* **3**, 1115–1122. MR2367258

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Kains, B. H., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Vijver, M. J. V. D., Bergh, J., Piccart, M. and Delorenzi, M. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute* **98**, 262–272.

Tiedermann, R. E., Zhu, Y. X., Schimidt, J., Shi, C. X., Sereduk, C., Yin, H., Mousses, S. and Stewart, A. K. (2012). Identification of molecular vulnerabilities in human multiple myeloma cells by RNA interference lethality screening of the druggable genome. *Cancer Research* **72**, 757–768.

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63.

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Gelder, M. E. M. V., Yu, J., Jatkoe, T., Berns, E. M. J. J., Atkins, D. and Foekens, J. A. (2005). Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679.

Warren, P., Taylor, D., Martini, P. G. V., Jackson, J. and Bienkowska, J. (2007). PANP—A new method of gene detection on oligonucleotide expression arrays. In *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering* 108–115. Boston, MA: IEEE. DOI:10.1109/BIBE.2007.4375552.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics 7* (J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West, eds.) 723–732. New York: Oxford University Press. MR2003537

Wieringen, W. N. V., Belien, J. A. M., Vosse, S. J., Achame, E. M. and Ylstra, B. (2006). ACE-it: A tool for genome-wide integration of gene dosage and RNA expression data. *Bioinformatics* **22**, 1919–1920.

Whitlock, M. C. (2005). Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology* **18**, 1368–1373.

Wu, Z. and Irizarry, R. A. (2005). A statistical framework for the analysis of microarray probe-level data. Working Paper 73, Johns Hopkins Univ., Dept. Biostatistics. Available at http://www.bepress.com/jhubiostat/paper73.

Wu, Z., Irizarry, R. A., Gentleman, R., Murillo, F. M. and Spencer, F. (2004). A model based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association* **99**, 909–917. MR2113309

Yeung, K. Y. and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics* **17**, 763–774.

Departamento de Estatistica, ICEx
Universidade Federal de Minas Gerais
Av. Antonio Carlos, 6627, Pampulha
Belo Horizonte, MG 31270-901
Brazil
E-mail: vdm@est.ufmg.br

Institute for Genome Sciences & Policy
Duke University
Box 3382
Durham, North Carolina 27708
USA
E-mail: joseph.lucas@duke.edu