iBATCGH: Integrative Bayesian Analysis of Transcriptomic and CGH Data

Alberto Cassese, Michele Guindani, and Marina Vannucci

Abstract We describe a method for the integration of high-throughput data from different sources. More specifically, `iBATCGH` is a package for the integrative analysis of transcriptomic and genomic data, based on a hierarchical Bayesian model. Through the specification of a measurement error model we relate the gene expression levels to latent copy number states which, in turn, are related to the observed surrogate CGH measurement via a hidden Markov model. Selection of relevant associations is performed employing variable selection priors that explicitly incorporate dependence information across adjacent copy number states. Posterior inference is carried out through Markov chain Monte Carlo techniques that efficiently explores the space of all possible associations. In this chapter we review the model and present the functions provided in `iBATCGH`, an R package based on a C implementation of the inferential algorithm. Lastly, we illustrate the method via a case study on ovarian cancer.

1 Introduction

In recent years, the field of genomics has seen the development of modern profiling high-throughput techniques that have resulted in the generation of large-scale data sets. The development of these modern techniques has made available several platforms to profile DNA, RNA and proteins, at different levels of accuracy. Integrating data from those different sources has emerged as a challenging problem in genomics, and a fundamental step in the understanding of many diseases. For example, it is now well known that cancer is the consequence of a dynamic interplay

A. Cassese
Maastricht University, Maastricht, The Netherlands
e-mail: alberto.cassese@maastrichtuniversity.nl

M. Guindani
UT MD Anderson Cancer Center, Houston, TX, USA
e-mail: mguindani@mdanderson.org

M. Vannucci (✉)
Department of Statistics, Rice University, Houston, TX, USA
e-mail: marina@rice.edu

A. Frigessi et al. (eds.), *Statistical Analysis for High-Dimensional Data*,
Abel Symposia 11, DOI 10.1007/978-3-319-27099-9_6

at different levels (DNA, mRNA and protein). Multilevel studies that try to integrate different types of data have therefore become of great interest.

Here we focus on the combined analysis of gene expression data and DNA copy number aberrations. Gene expression data are measurements of the abundance of a set of transcripted genes in a specific tissue. At the DNA level, many different kinds of aberration can occur and, for this reason, many different methods have been developed to detect them. Here we focus on Comparative Genomic Hybridization (CGH), a method able to detect copy number changes. This technique has a relatively high resolution and can span a large part of the genome in a single experiment. CGH data are well suited for cancer studies, since cancer is the result of a number of complex biological events and as such it cannot be attributed to a single mutation. Thus, discovering amplification of oncogenes or deletion of tumor suppressors is an important step for elucidating tumorigenesis.

Some methods that rely on regression models as a way to integrate gene expression data with copy number variants have been developed [21, 28]. These methods do not infer the underlying copy number information, but rather use their surrogate CGH measurements as regressors. Alternatively, methods have been proposed that first estimate copy number latent states using available methodology, and, as a second step, use these estimates as regressors [3, 33]. Here we describe a novel method we have proposed for the joint estimation of copy number aberrations and their association with copy number variants [7, 8]. More specifically, we have developed a model that regresses gene expression on copy number states, while simultaneously estimating the latent copy number states of the observed surrogate CGH data. This modeling strategy is able to take into account the uncertainty on the latent unobserved copy number states typical of CGH data, while simultaneously assessing their association with gene expression. The model employs selection priors to detect significant associations, incorporating information on the physical distance between neighboring DNA probes and their latent copy number and association status. The latent copy number states are estimated via a hidden Markov model, which is able to capture the peculiar stickiness of CGH data.

In this chapter, after reviewing the proposed methodology, we describe in details the implementation of the methods via the package `iBATGH`, released under the GNU General Public License within the R project, and freely available on the CRAN website. The package is mainly based on an algorithm using C-code and the interface with the R environment is handled using the packages `Rcpp` and `RcppArmadillo` [13]. This choice achieves a good performance in terms of computational speed, with the advantage of a user friendly interface as provided by the R environment.

The rest of this chapter is organized as follows. In Sect. 2 we describe the model, the priors and the posterior inferential algorithm. In Sect. 3 we describe the R package `iBATCGH`. In Sect. 4 we illustrate the method via a case study.

2 Model

Let us first introduce the notation that will be used throughout the following sections. Let $\boldsymbol{Y} = [Y_{ig}]_{n\times G}$ be the $n \times G$ matrix of gene expression measurements on G genes in n subjects. Let $\boldsymbol{X} = [X_{im}]_{n\times M}$ denote the matrix of CGH measurements on M DNA probes, on the same samples ($i = 1, \ldots, n$). We assume the CGH probes ordered according to their chromosomal location and refer to two consecutive probes as adjacent. Lastly, let $\boldsymbol{Z} = [\boldsymbol{Y}, \boldsymbol{X}]_{n\times(G+M)}$ denote the $(n \times (G + M))$ matrix containing all data measurements.

In our modeling approach we treat the observed CGH intensities $\boldsymbol{X}$ as surrogates for the unobserved copy number states. In particular, we introduce $\boldsymbol{\xi} = [\xi_{im}]_{n\times M}$, a latent matrix of copy number states, and consider a four copy number states classification [15]:

$\xi_{im} = 1$ for copy number loss (less than two copies of the fragment)

$\xi_{im} = 2$ for copy-neutral state (exactly two copies of the fragment)

$\xi_{im} = 3$ for a single copy gain (exactly three copies of the fragment)

$\xi_{im} = 4$ for multiple copy gains (more than three copies of the fragment).

We assume that, conditional on the latent state $\boldsymbol{\xi}$, the corresponding observed surrogate $\boldsymbol{X}$ does not contain additional information on the outcome $\boldsymbol{Y}$, that is, $f(\boldsymbol{Y}|\boldsymbol{\xi}, \boldsymbol{X}) = f(\boldsymbol{Y}|\boldsymbol{\xi})$. In the statistical literature this modeling framework is commonly referred to as a *non-differential measurement error model* [27], and allows the factorization of the joint distribution of $\boldsymbol{Z}$ as the product of two conditionally independent sub models: an outcome model, that in our modeling context relates the gene expressions with the latent copy number states, and a measurement model, that relates the latent states to the observed surrogate CGH measurements, $f(\boldsymbol{Z}|\boldsymbol{\xi}) = f(\boldsymbol{Y}|\boldsymbol{\xi})f(\boldsymbol{X}|\boldsymbol{\xi})$. As commonly done in the literature on integrative genomics, we assume conditional independence of the gene expression measurements, $\boldsymbol{Y}_i \perp \boldsymbol{Y}_j|\xi_1, \ldots, \boldsymbol{\xi}_M$. We also assume independence of the CGH measurements, conditional on their latent states, $\boldsymbol{X}_i \perp \boldsymbol{X}_j|\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M$. Given those assumptions, we can write our proposed model as

$$f(\mathbf{Z}|\boldsymbol{\xi}) = \prod_{i=1}^{n} \left\{ \prod_{g=1}^{G} f(Y_{ig}|\boldsymbol{\xi}_i) \prod_{m=1}^{M} f(X_{im}|\xi_{im}) \right\}. \tag{1}$$

Below we provide detailed specification of the outcome and the marginal models.

2.1 Outcome Model with Measurement Error

We follow current literature on models that integrate gene expression levels with genetic data and specify the outcome model as a linear regression of the gene expression measurements on the latent copy number states [21, 28]. For each gene $g = 1, \ldots, G$ the regression is defined as follows,

$$Y_{ig} = \mu_g + \boldsymbol{\xi}_i \boldsymbol{\beta}_g + \epsilon_{ig}, \qquad i = 1, \ldots, n, \tag{2}$$

where μ_g is a gene specific intercept and $\epsilon_{ig} \sim \mathcal{N}(0, \sigma_g^2)$ are independent normally distributed errors, with σ_g^2 a gene specific variance. We further assume conjugate prior distributions for the intercept, $\mu_g|\sigma_g^2 \sim \mathcal{N}(0, c_\mu^{-1}\sigma_g^2)$, and gene specific precisions, $\sigma_g^{-2} \sim \text{Ga}(\frac{\delta}{2}, \frac{d}{2})$, with c_μ, δ and d hyperparameters to be chosen.

We look at the identification of copy number variations associated with gene expression levels as a variable selection problem. We therefore introduce a latent binary matrix $\boldsymbol{R} = [r_{gm}]_{G\times M}$, where the generic element r_{gm} is set to one if the corresponding regression coefficient β_{gm} is different from zero, and $r_{gm} = 0$ otherwise. As commonly done in the Bayesian variable selection literature, we adopt spike and slab priors on the conditional distribution of the regression coefficients,

$$\beta_{gm}|r_{gm}, \sigma_g^2 \sim r_{gm}\mathcal{N}(0, c_\beta^{-1}\sigma_g^2) + (1 - r_{gm})\delta_0(\beta_{gm}), \tag{3}$$

with $\delta_0(\cdot)$ a point mass at zero, σ_g^2 is a gene specific variance, and c_β an hyperparameter to be chosen [5, 14]. Equation (3) is completed with a prior on r_{gm}. The simplest choice is to assume an independent Bernoulli prior, $r_{gm} \sim \text{Bern}(p)$ with p a fixed hyperparameter. However, in Sect. 2.3 below, we describe two alternative priors that accounts for spatial information [7, 8], borrowing strength across genes.

2.2 Marginal Model

We now describe the marginal model that relates the latent copy number states with the surrogate CGH measurements. In the literature this problem has been tackled by using modeling strategies that employ either circular binary segmentation, a method that infers change points [35], or clustering based methods [6, 25]. In addition, Bayesian nonparametric methods have been applied to CGH modeling [1, 12, 38], as well as methods that rely on hidden Markov models [9, 15, 36]. Here we extend the latter approaches, in particular the one proposed by Guha et al. [15], to handle multiple samples in a single modeling framework.

A hidden Markov model is a state space model with discrete hidden states. It comprises of a Markov chain with stochastic measurements on the hidden states and, conditionally on the states, of an independent emission distribution. In the context of our specific application, conditional on the latent copy number states, the

observed CGH measurements are assumed independent and normally distributed, and the emission distribution is defined as

$$X_{im}|(\xi_{im} = j) \overset{iid}{\sim} \mathcal{N}(\eta_j, \sigma_j^2), \tag{4}$$

with η_j and σ_j^2 denoting respectively the expected $\log_2$ ratio and variance of all CGH probes in state j ($j = 1, \ldots, 4$). As for the latent copy number states, we choose a first order HMM, which assumes that the probability of being in a particular copy number state for a given probe m depends only on the state of the previous probe $(m-1)$,

$$P(\xi_{im}|\xi_{i1}, \ldots, \xi_{i(m-1)}) = P(\xi_{im}|\xi_{i(m-1)}) = a_{\xi_{i(m-1)}\xi_{im}}, \tag{5}$$

with $\boldsymbol{A} = [a_{hj}]_{4\times4}$ a matrix of transition probabilities with strictly positive elements. We also assume that the distribution of the first probe is given by the unique stationary distribution of $\boldsymbol{A}$, denoted by π_A. We further assume the samples to be independent, and that they share the same transition matrix. In summary, the proposed HMM can be factorized as

$$\begin{aligned} P(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M) = &\prod_{i=1}^{n} P(X_{i1}|\xi_{i1})P(\xi_{i1}) \\ &\prod_{m=2}^{M} P(X_{im}|\xi_{im})P(\xi_{im}|\xi_{i(m-1)}). \end{aligned} \tag{6}$$

To complete our prior specification of the HMM, we choose conjugate priors. In particular, we assume each row of the transition matrix $\boldsymbol{A}$ as a draw from independent Dirichlet distributions, $\boldsymbol{a}_h = [a_{h1}, a_{h2}, a_{h3}, a_{h4}] \sim \text{Dir}(\phi_1, \phi_2, \phi_3, \phi_4)$, for $h = 1, \ldots, 4$, with $\boldsymbol{\phi} = [\phi_1, \phi_2, \phi_3, \phi_4]$ a vector of hyperparameters to be chosen. As for the state specific mean η_j and variance σ_j^2, we assume $\eta_j \sim N(\delta_j, \tau_j^2)\boldsymbol{I}\{low_{\eta_j} < \eta_j < upp_{\eta_j}\}$ and $\sigma_j^{-2} \sim \text{Gamma}(b_j, l_j)\boldsymbol{I}\{\sigma_j^{-2} > upp_{\sigma_j}\}$, for $j = 1, \ldots, 4$. Lastly, we set $low_{\eta_1} = -\infty$, $upp_{\eta_4} = \infty$, while all other hyperparameters are defined by the user on the base of the platform [15].

2.3 Spatially Informed Variable Selection Priors

Our choice of the selection prior relies on the consideration that two contiguous regions of copy number variants might correspond to the same aberration. As a consequence, they are more likely to jointly affect gene expression. We therefore define a prior that accounts for the selection status of the adjacent probes. In other words, the prior probability $p(r_{gm})$ of an association between gene g and probe m depends on the values of $r_{g(m-1)}$ and $r_{g(m+1)}$. As a first step of the prior construction,

we define a probe specific quantity that incorporates, in a multiplicative fashion, information on the physical distance among probes and on the frequency of copy number change points across samples as

$$s_{(m-1)m} = \left\{ \frac{\exp\{1 - \frac{d_m}{D}\} - 1}{\exp\{1\} - 1} \right\} \left\{ \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{I}\{\xi_{im} = \xi_{i(m-1)}\} \right\}, \tag{7}$$

where d_m denotes the distance between adjacent probes $[m-1, m]$ and where D is the total length of the DNA fragment, for example the length of the chromosome under study. We then explore two different ways of employing these quantities in the definition of the selection prior [7, 8]. More specifically, we have investigated a mixture prior and a Probit prior.

(a) **Mixture prior**: We start defining the quantities,

$$\begin{aligned} \gamma_m &= \frac{\alpha}{\alpha + s_{(m-1)m} + s_{m(m+1)}}, \\ \omega_m^{(1)} &= \frac{s_{(m-1)m}}{\alpha + s_{(m-1)m} + s_{m(m+1)}}, \quad \omega_m^{(2)} = \frac{s_{m(m+1)}}{\alpha + s_{(m-1)m} + s_{m(m+1)}}, \end{aligned} \tag{8}$$

with α set to a positive real value, and then define a mixture prior with two components

$$\begin{aligned} \pi(r_{gm} | r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi}, \pi_1) = \gamma_m [\pi_1^{r_{gm}} (1-\pi_1)^{(1-r_{gm})}] \\ + \sum_{j=1}^{2} \omega_m^{(j)} \boldsymbol{I}\{r_{gm} = r_{g(m+(-1)^j)}\}. \end{aligned} \tag{9}$$

According to Eq. (9), with probability γ_m, the r_{gm}'s are distributed Bern(π_1), independently of the neighboring values. Otherwise, r_{gm} assumes the same value as in $r_{g(m-1)}$ or $r_{g(m+1)}$, with probability $\omega_m^{(1)}$ or $\omega_m^{(2)}$, respectively. We note that the weights in Eq. (8) sum up to one, i.e. $\gamma_m + \omega_m^{(1)} + \omega_m^{(2)} = 1$, and that the case $\gamma_m = 1$ reduces to an independent Bernoulli prior. We further note that $\alpha \to \infty$ implies $\gamma_m \to 1$, that is the independent prior, while when $\alpha = 0$, r_{gm} depends only on the values $r_{g(m-1)}$ and $r_{g(m+1)}$, with weights proportional to $s_{(m-1)m}$ and $s_{m(m+1)}$, respectively. As a consequence, lower values of α implies a stronger dependence on the selection status of the adjacent probes. In addition, larger values of $s_{(m-1)m}$ imply a stronger dependence of probe m on probe $(m-1)$, and viceversa. This reflects the assumption that two probes physically close and that share a similar copy number status are more likely to have the same association pattern. In [7] we suggest to chose α in the range $\alpha = [20, 50]$, for a good balance in terms of false positives and false negatives. We complete our prior specification by imposing a Beta hyperprior on $\pi_1 \sim \text{Beta}(e, f)$, and

integrating it out. This results in the following equation

$$\pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi}) = \gamma_m \frac{\Gamma(e+f)\Gamma(e+r_{gm})\Gamma(f+1-r_{gm})}{\Gamma(e+f+1)\Gamma(e)\Gamma(f)} + \sum_{j=1}^{2} \omega_m^{(j)} \boldsymbol{I}_{\{r_{gm}=r_{g(m+(-1)^j)}\}}. \tag{10}$$

(b) **Probit prior**: Let us first define the quantity Q_m as

$$Q_m = (-1)^{r_{g(m-1)}} s_{(m-1)m} + (-1)^{r_{g(m+1)}} s_{m(m+1)}. \tag{11}$$

Note that Q_m can either increase or decrease based on the selection status of the adjacent probes, and that the amount of increase or decrease depends on $s_{(m-1)m}$ and $s_{m(m+1)}$. We define the probability of inclusion for r_{gm} as

$$\pi(r_{gm} = 1|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi}) = 1 - \Phi(\alpha_0 + \alpha_1 Q_m), \tag{12}$$

where Φ indicates the c.d.f. of a standard normal distribution, and α_0 and $\alpha_1 > 0$ are hyperparameters to be set. In particular, α_0 represents a baseline intercept that can be set according to an a priori specified "level of significance", in absence of other covariates. Similarly, α_1 can be interpreted as a coefficient that captures the strength of the association between adjacent probes. We note that $\pi(r_{gm})$ is a monotonic function of Q_m, therefore it increases or decreases, based on the selection status of the adjacent probes, by an amount determined by $s_{(m-1)m}$ and $s_{m(m+1)}$. This reflects the assumption that two probes physically close and that share a similar copy number status are more likely to have the same association pattern. Even though the prior specifications (9) and (12) share the same assumptions and employ similar quantities, prior (12) is of more easy interpretation and has produced better results on simulated data [8].

2.4 Posterior Inference

In this section we describe the approach employed to perform posterior inference. In particular, the methodology aims at estimating the association matrix $\boldsymbol{R}$ and the matrix of copy number states $\boldsymbol{\xi}$. We rely on a Markov chain Monte Carlo algorithm that employs stochastic search variable selection techniques [5, 7, 8, 14, 28, 29, 31]. In order to simplify the algorithm and to improve the mixing of the chain, we integrate out the regression coefficients μ_g, β_g and σ_g^2 [5, 29, 32]. The marginal likelihood reduces to

$$f(\boldsymbol{Y}_g|\boldsymbol{\xi}, \boldsymbol{R}) = \frac{(2\pi)^{-\frac{n}{2}} \left(\frac{c_\mu}{c_\mu+n}\right)^{\frac{1}{2}} (c_\beta)^{\frac{k_g}{2}} \Gamma\left(\frac{n+\delta}{2}\right) \left(\frac{d}{2}\right)^{\frac{\delta}{2}}}{|\boldsymbol{U}_g|^{\frac{1}{2}} \Gamma\left(\frac{\delta}{2}\right) \left(\frac{d+q_g}{2}\right)^{\left(\frac{n+\delta}{2}\right)}}, \tag{13}$$

Algorithm 1 Selection of the subsets of rows of $\boldsymbol{R}$ and $\boldsymbol{\xi}$ to be updated at every MCMC iteration

```
set cumsum = 0
repeat
    Generate ζ from Geom(p)
    Sum ζ to cumsum
    Add cumsum to the set of selected features
until cumsum > F
```

with $q_g = \boldsymbol{Y}_g^T \boldsymbol{H}_s \boldsymbol{Y}_g - \boldsymbol{Y}_g^T \boldsymbol{H}_s \boldsymbol{\xi}_R \boldsymbol{U}_g^{-1} \boldsymbol{\xi}_R^T \boldsymbol{H}_s \boldsymbol{Y}_g$, $\boldsymbol{U}_g = c_\beta \boldsymbol{I}_{k_g} + \boldsymbol{\xi}_R^T \boldsymbol{H}_s \boldsymbol{\xi}_R$ and $\boldsymbol{H}_s = \boldsymbol{I}_n - \frac{\mathbf{1}_n \mathbf{1}_n^T}{n + c_\mu}$, where k_g indicates the number of selected regressors for the gth regression. Aiming at more efficient MCMC steps, we perform multiple updates of $\boldsymbol{R}$ and $\boldsymbol{\xi}$. This is accomplished relying on Algorithm 1 for the selection of a subset of rows to be updated at every MCMC iteration. The MCMC consists of four steps and is described below.

- Update $\boldsymbol{R}$ using a Metropolis step. First use Algorithm 1 with parameters $p = p_R$ and $F = G$ to select at random a set of genes. Then, for each gene in the set, choose between an Add/Delete or Swap moves, with probability ρ and $(1 - \rho)$ respectively. For the Add/Delete move, choose at random one element of the row and change its selection status. For the Swap move, select at random two elements of the row with different inclusion status and swap their values. In order to efficiently explore the space of all possible associations, we do not consider some CGH probes as possible regressors. In particular, we exclude those CGH probes that have been called in neutral state in a fraction of samples larger than p_{MC}, at the current MCMC iteration. Note that p_{MC} is a parameter set by the user, for example 10 % is the default argument in the package. We accept the proposed move with probability

$$\min \left[\frac{f(\boldsymbol{Y}|\boldsymbol{\xi}, \boldsymbol{R}^{new}) \pi(\boldsymbol{R}^{new}|\boldsymbol{\xi})}{f(\boldsymbol{Y}|\boldsymbol{\xi}, \boldsymbol{R}^{old}) \pi(\boldsymbol{R}^{old}|\boldsymbol{\xi})}, 1 \right].$$

- Update $\boldsymbol{\xi}$ using a Metropolis-Hastings step. First select a column of $\boldsymbol{\xi}$ and then use Algorithm 1 with parameters $p = p_\xi$ and $F = N$ to select at random a subset of samples. For each selected element sample a candidate state using the current transition matrix $\boldsymbol{A}$. In other words, we propose ξ_{im}^{new} conditional on ξ_{im}^{old}. Accept the proposed move with probability

$$\min \left[\frac{f(\boldsymbol{Y}|\boldsymbol{\xi}^{new}, \boldsymbol{R}) f(\boldsymbol{X}|\boldsymbol{\xi}^{new}) \pi(\boldsymbol{R}|\boldsymbol{\xi}^{new}) \pi(\boldsymbol{\xi}^{new}|\boldsymbol{\xi}^{old}, \boldsymbol{A}) q(\boldsymbol{\xi}^{old}|\boldsymbol{\xi}^{new})}{f(\boldsymbol{Y}|\boldsymbol{\xi}^{old}, \boldsymbol{R}) f(\boldsymbol{X}|\boldsymbol{\xi}^{old}) \pi(\boldsymbol{R}|\boldsymbol{\xi}^{old}) \pi(\boldsymbol{\xi}^{old}|\boldsymbol{\xi}^{old}, \boldsymbol{A}) q(\boldsymbol{\xi}^{new}|\boldsymbol{\xi}^{old})}, 1 \right].$$

- Update η_j, for $j = 1, \ldots, 4$ using a Gibbs step. Sample $\eta_j|\boldsymbol{X}, \boldsymbol{\xi}, \sigma_j \sim \mathcal{N}(\nu_j, \theta_j^{-2}) \mathbf{I}_{\{low_{\eta_j} < \eta_j < upp_{\eta_j}\}}$, with precision $\theta_j = \tau_j^{-2} + n_j \sigma_j^{-2}$ and weighted

mean $\nu_j = \theta_j^{-2}(\delta_j \tau_j^{-2} + \bar{X}_j n_j \sigma_j^{-2})$, where $n_j = \sum_{m=1}^{M} \sum_{i=1}^{n} \boldsymbol{I}\{\xi_{im} = j\}$ and $\bar{X}_j = \frac{1}{n_j} \sum_{m=1}^{M} \sum_{i=1}^{n} X_{im} \boldsymbol{I}\{\xi_{im} = j\}$.

- Update σ_j, for $j = 1, \ldots, 4$ using a Gibbs step. Sample $\sigma_j | \boldsymbol{X}, \boldsymbol{\xi}, \eta_j \sim \text{IG}(b_j + \frac{n_j}{2}, l_j + \frac{V_j}{2}) \boldsymbol{I}_{\{\sigma_j^{-2} > upp_{\sigma_j}\}}$, where $n_j = \sum_{m=1}^{M} \sum_{i=1}^{n} \boldsymbol{I}_{\{\xi_{im}=j\}}$ and $V_j = (X_{im} - \eta_j)^2 \boldsymbol{I}_{\{\xi_{im}=j\}}$.
- Update $\boldsymbol{A}$ using a Metropolis step. Generate for each row of $\boldsymbol{A}$ a new vector as $\boldsymbol{A}_{\cdot j}^{new} | \boldsymbol{\xi} \sim Dir(\phi_1 + o_{h1}, \phi_2 + o_{h2}, \phi_3 + o_{h3}, \phi_4 + o_{h4})$, where $o_{hj} = \sum_{i=1}^{n} \sum_{m=1}^{M-1} \boldsymbol{I}_{\{\xi_{im}=h, \xi_{i(m+1)}=j\}}$, and accept it with probability

$$\min \left[1, \prod_{i=1}^{n} \frac{\pi_{A^{new}}(\xi_{i1})}{\pi_{A^{old}}(\xi_{i1})} \right].$$

We summarize the output of the MCMC as follows. Inference on $\boldsymbol{R}$ is performed computing the marginal probability of inclusion (PPI) for each of its elements. More specifically, PPIs are obtained by counting for each element of the matrix the number of iterations it was set to one after burn-in, and dividing by the total number of iteration after burn-in. A selection of the most relevant associations can be made by thresholding the PPIs based on some decision theoretic criterion, see for example [2, 23]. As for the inference on $\boldsymbol{\xi}$ we select for every position of the matrix the modal state after burn-in, i.e. the state that shows the highest count. Inference on the HMM parameters $\boldsymbol{A}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ is performed by averaging their values across the MCMC iterations, after burn-in.

3 The iBATCGH Package

The package `iBATCGH` is released under the GNU General Public License within the R project, and is freely available on the CRAN website. It uses the libraries `Rcpp` and `RcppArmadillo`, and it is based on a backbone C implementation [13]. The package comprises of nine functions and two data sets. The first data set, `NCI_60`, consists of the processed and filtered NCI-60 cancer cell lines data, as used in [7]. The second one, `TCGA_lung`, is the processed and filtered TCGA lung squamous cell carcinoma data set, as used in [8]. In the rest of the section, we describe the main R functions and provide an example on how to run them. A summary of the functions with their arguments, output and a brief description can be found in Table 1. Throughout this Section, the variables in the code will be written using `teletypefont`, while the corresponding notation used in Sect. 2 will be reported in parenthesis.

The functions `Scenario1` and `Scenario2` generate two types of simulated data sets, with `Scenario2` explicitly assuming dependence among the regression coefficients [7]. Their only argument is the error variance of the regression model `sigmak` (σ_ϵ), set to 0.1 by default. Those functions return a list composed by the

Table 1 Available functions of the package `iBATCG`, their arguments, output and brief description

Function	Arguments	Output	Description
`Center`	$\boldsymbol{Y}$	$\bar{\boldsymbol{Y}}$	*Preprocessing*, center each column of the gene expression matrix
`iBAT`	$\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{d}$, D, *intercept*, $\boldsymbol{\xi}$, $\boldsymbol{R}$, $\boldsymbol{A}$, $\boldsymbol{\eta}$, $\boldsymbol{\sigma}$, c_μ, c_β, δ, d, e, f, α, $\boldsymbol{\delta}$, $\boldsymbol{tau}$, $\boldsymbol{upp}_\eta$, $\boldsymbol{low}_\eta$, $\boldsymbol{b}$, $\boldsymbol{l}$, $\boldsymbol{upp}_\sigma$, $\boldsymbol{\phi}$, *niter*, *burnin*, *Cout*, ϕ, p_R, p_{MC}, p_ξ	**MCMC output**	*Main*, function for the model that employs the mixture prior
`iBATProbit`	$\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{d}$, D, *intercept*, $\boldsymbol{\xi}$, $\boldsymbol{R}$, $\boldsymbol{A}$, $\boldsymbol{\eta}$, $\boldsymbol{\sigma}$, c_μ, c_β, δ, d, α_0, α_1, $\boldsymbol{\delta}$, $\boldsymbol{tau}$, $\boldsymbol{upp}_\eta$, $\boldsymbol{low}_\eta$, $\boldsymbol{b}$, $\boldsymbol{l}$, $\boldsymbol{upp}_\sigma$, $\boldsymbol{\phi}$, *niter*, *burnin*, *Cout*, ϕ, p_R, p_{MC}, p_ξ, *indep*	**MCMC output**	*Main*, function for the model that employs the Probit prior
`Inference`	**MCMC output**, G, M, *niter*, *burnin*, *threshold*	$\hat{\boldsymbol{R}}$, $\hat{\boldsymbol{\xi}}$, $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\sigma}}$	*Postprocessing*, perform posterior inference on the output of the main function
`InitMu`	$\boldsymbol{\delta}$, $\boldsymbol{\tau}$, $\boldsymbol{low}_\eta$, $\boldsymbol{upp}_\eta$	$\boldsymbol{\mu}$	*Initialization*, initialize $\boldsymbol{\mu}$, sampling from the prior distribution
`InitXi`	$\boldsymbol{X}$	$\boldsymbol{\xi}$	*Initialization*, initialize $\boldsymbol{\xi}$, using a crude estimator on $\boldsymbol{X}$
`Scenario1`	σ_ϵ	$\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\xi}$, $\boldsymbol{A}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{B}$, $\boldsymbol{d}$, D	*Simulated data*, simulate the data as described in [7, 8]
`Scenario2`	σ_ϵ	$\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\xi}$, $\boldsymbol{A}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{B}$, $\boldsymbol{d}$, D	*Simulated data*, simulate the data as described in [7, 8]
`Tran`	$\boldsymbol{\xi}$	$\boldsymbol{A}$	*Preprocessing*, compute the transition matrix corresponding to a specific $\boldsymbol{\xi}$

A **bold** variable in the arguments/output represents a vector, matrix or list

two data matrices `Y` ($\boldsymbol{Y}$) and `X` ($\boldsymbol{X}$), and the empirical parameters of the HMM, `Xi` ($\boldsymbol{\xi}$), `A` ($\boldsymbol{A}$), `mu` ($\boldsymbol{\mu}$) and `Sd` ($\boldsymbol{\sigma}$). Also, they return a matrix of regression coefficients `coeff` ($\boldsymbol{B}$), a vector of distances between probes `distance` ($\boldsymbol{d}$) and the total length of the DNA fragment `disfix` (D).

The function `Center` takes as argument a matrix of gene expression measurements `Y` ($\boldsymbol{Y}$) and returns the matrix obtained after centering each column with respect to its mean. As for the initialization of the parameters, `InitMu` initializes the state specific mean vector, by sampling each element independently from its prior, i.e.

Algorithm 2 Crude estimator employed to initialize $\boldsymbol{\xi}$

```
for i=1 to i=N do
  for m=1 to m=M do
    if X_im < bounds_1 then
      set ξ_im = 1
    else if X_im < bounds_2 then
      set ξ_im = 2
    else if X_im < bounds_3 then
      set ξ_im = 3
    else
      set ξ_im = 4
    end if
  end for
end for
```

truncated normal distribution. This function has default arguments and can be run simply as `> mu = InitMu()`. However the user can change the arguments of the vector of state specific means `deltak` ($\boldsymbol{\delta}$), standard deviations `tauk` ($\boldsymbol{\tau}$), lower bound `low_bounds` ($\boldsymbol{low}_{\eta}$) and upper bound `upp_bounds` ($\boldsymbol{upp}_{\eta}$). The function `InitXi` takes a matrix of CGH data as the only argument and returns a crude estimate of the corresponding latent copy number states. More specifically, given a vector of threshold `bounds`, set by default to $[-0.5, 0.29, 0.79]$, the function simply applies the thresholding to the data and groups them into four subsets. Each subset is associated to a specific latent state as described in Algorithm 2. Given a matrix of latent states $\boldsymbol{\xi}$, an empirical transition matrix can be computed. In order to initialize $\boldsymbol{A}$, we use this empirical transition matrix computed on the initialized value of $\boldsymbol{\xi}$. The function implemented for this purpose is `Tran` and takes `xi` as only argument.

We now focus on the two main functions implemented for the integrative Bayesian analysis of gene expression and CGH data. More specifically, `iBAT` employs the mixture prior and `iBATProbit` employs the Probit prior. Tables 2 and 3 show a description of their arguments. In particular, the first column reports the arguments used in the code, the second column recalls the corresponding math notation as introduced in Sect. 2, and the third column adds a brief description. Note that the case of spatially independent variable selection priors can be obtained by setting the option `indep` to one or `alpha1` to zero, respectively in `iBAT` and `iBATProbit`. The output consists of an R list composed by $4*$`niter`$+3$ objects, where `niter` is the number of MCMC iterations. The first `niter` objects of the list are vectors, each containing the positions of the association matrix set to one, at the corresponding MCMC iteration. Each of the following `niter` objects of the list are the transition matrices at the corresponding MCMC iteration, while the third and the fourth set of `niter` objects are the vectors of state specific mean and state specific variance, respectively. The last three objects of the list consist of three matrices counting the number of times the corresponding latent state has been set to 1, 3 and 4, respectively.

Table 2 Arguments of `iBAT`: name in the code, corresponding math notation, and brief description

Argument	Math	Description
`Y`	$\boldsymbol{Y}$	Matrix of gene expression data
`X`	$\boldsymbol{X}$	Matrix of CGH data
`distance`	$\boldsymbol{d}$	Vector of distance between CGH probes
`disfix`	D	Length of the chromosome under investigation
`intercept`	NA	If set to one an intercept is included in the regression model
`xi`	$\boldsymbol{\xi}$	Initialized matrix of latent states
`R`	$\boldsymbol{R}$	Initialized association matrix in a vector form. Default set to -1, that automatically creates a vector with all the positions set to zero
`tran`	$\boldsymbol{A}$	Initialized transition matrix
`mu`	$\boldsymbol{\mu}$	Initialized state specific mean vector
`sigma`	$\boldsymbol{\sigma}$	Initialized state specific standard deviation vector
`cmu`	c_μ	Parameter that controls the variance of the prior on the intercept
`c`	c_β	Parameter that determines the shrinkage in the model
`delta`	δ	Parameter of the Inverse-Gamma prior on the error variance
`d`	d	Parameter of the Inverse-Gamma prior on the error variance
`e`	e	Parameter of the Beta prior on the inclusion probability
`f`	f	Parameter of the Beta prior on the inclusion probability
`alpha`	α	Parameter that regulates the strength of the independent part of the mixture
`deltak`	$\boldsymbol{\delta}$	Vector of mean of the prior on the state specific mean
`tauk`	$\boldsymbol{\tau}$	Vector of standard deviation parameters of the prior on the state specific mean
`upp_bounds`	$\boldsymbol{upp}_\eta$	Vector of upper bounds of the prior on the state specific mean
`low_bounds`	$\boldsymbol{low}_\eta$	Vector of lower bounds of the prior on the state specific mean
`alpha_IG`	$\boldsymbol{b}$	Vector of parameters of the prior on the state specific standard deviation
`beta_IG`	$\boldsymbol{l}$	Vector of parameters of the prior on the state specific standard deviation
`low_IG`	$\boldsymbol{upp}_\sigma$	Truncation of the prior on the state specific standard deviation
`a`	$\boldsymbol{\phi}$	Vector of parameters of the prior on the transition matrix
`niter`	NA	Number of Markov Chain Monte Carlo iterations
`burnin`	NA	Burn-in
`Cout`	NA	Print the number of iterations ran every `Cout` iterations
`phi`	ϕ	Probability of an A/D step
`pR`	p_R	Parameter of the distribution used to select the rows to be updated at every MCMC iteration
`selectioncgh`	p_{MC}	Number of samples not in neutral state in order to consider a CGH as a potential candidate for association with gene expression. Default set to -1 that automatically set it to 10 % of the samples
`pXI`	p_ξ	Parameter of the distribution used to select the rows to be updated at every MCMC iteration
`indep`	NA	If set to an integer different from zero, run the analysis with an independent prior, i.e. setting $\alpha \to \infty$

Table 3 Arguments of `iBATProbit`: name in the code, corresponding math notation, and brief description

Argument	Math	Description
`Y`	$\boldsymbol{Y}$	Matrix of gene expression data
`X`	$\boldsymbol{X}$	Matrix of CGH data
`distance`	$\boldsymbol{d}$	Vector of distance between CGH probes
`disfix`	D	Length of the chromosome under investigation
`intercept`	NA	If set to one an intercept is included in the regression model
`xi`	$\boldsymbol{\xi}$	Initialized matrix of latent states
`R`	$\boldsymbol{R}$	Initialized association matrix in a vector form. Default set to -1, that automatically creates a vector with all the positions set to zero
`tran`	$\boldsymbol{A}$	Initialized transition matrix
`mu`	$\boldsymbol{\mu}$	Initialized state specific mean vector
`sigma`	$\boldsymbol{\sigma}$	Initialized state specific standard deviation vector
`cmu`	c_μ	Parameter that controls the variance of the prior on the intercept
`c`	c_β	Parameter that determines the shrinkage in the model
`delta`	δ	Parameter of the Inverse-Gamma prior on the error variance
`d`	d	Parameter of the Inverse-Gamma prior on the error variance
`alpha0`	α_0	Baseline intercept of the selection prior
`alpha1`	α	Parameter that regulates the strength of the spatially informed dependence
`deltak`	$\boldsymbol{\delta}$	Vector of mean of the prior on the state specific mean
`tauk`	$\boldsymbol{\tau}$	Vector of standard deviation parameters of the prior on the state specific mean
`upp_bounds`	$\boldsymbol{upp}_\eta$	Vector of upper bounds of the prior on the state specific mean
`low_bounds`	$\boldsymbol{low}_\eta$	Vector of lower bounds of the prior on the state specific mean
`alpha_IG`	$\boldsymbol{b}$	Vector of parameters of the prior on the state specific standard deviation
`beta_IG`	$\boldsymbol{l}$	Vector of parameters of the prior on the state specific standard deviation
`low_IG`	$\boldsymbol{upp}_\sigma$	Truncation of the prior on the state specific standard deviation
`a`	$\boldsymbol{\phi}$	Vector of parameters of the prior on the transition matrix
`niter`	NA	Number of Markov Chain Monte Carlo iterations
`burnin`	NA	Burn-in
`Cout`	NA	Print the number of iterations ran every `Cout` iterations
`phi`	ϕ	Probability of an A/D step
`pR`	p_R	Parameter of the distribution used to select the rows to be updated at every MCMC iteration
`selectioncgh`	p_{MC}	Number of samples not in neutral state in order to consider a CGH as a potential candidate for association with gene expression Default set to -1, that automatically set it to 10 % of the samples
`pXI`	p_ξ	Parameter of the distribution used to select the rows to be updated at every MCMC iteration

The last function that we discuss summarizes the output of the MCMC chains and, therefore, allows to perform posterior inference. More specifically, the arguments of the function `Inference` are the output of the MCMC `listComplete`, the number of gene expression probes `G`, the number of CGH probes `M`, the number of MCMC iteration `niter`, the number of iterations to be discarded as burn-in `burnin`, and the threshold on the PPIs `threshold`, set by default to 0.5. The output of the function is a list made by five elements: `R`, the binary matrix of estimated association, `Xi`, the matrix of estimated copy number states, `A`, the estimated transition matrix, `Mu`, the estimated vector of state specific means, and `Sd`, the estimated vector of state specific standard deviations.

We now give an example of the use of the code. First get the data,

```
> data(NCI_60)
> Y = NCI_60$Affy
> X = NCI_60$aCGH
> distance = NCI_60$distance
```

then initialize ξ, $\boldsymbol{A}$ and $\boldsymbol{\eta}$ and center the measurements of each gene in the gene expression matrix,

```
> xi = InitXi(X)
> tran = Tran(xi)
> mu = InitMu()
> Y = Center(Y)
```

finally run the main function and summarize the results.

```
> res = iBAT(Y=Y,X=X,distance=distance,
disfix=146274826,xi=xi,tran=tran,
mu=mu,d=0.2587288)
> summRes = Inference(res,G=dim(Y)[[2]],
M=dim(X)[[2]],niter=niter,bi=bi,threshold=0.5)
```

The example above takes approximately 5 min to run 1000 iterations, using a double core ® Intel ® Xeon processor with 16 GB of memory, 2.2 GHz.

4 Case Study

In order to illustrate the performance of the model and the type of inference it allows, we further used the function `iBATProbit` to analyze data from a study of ovarian cancer obtained from The Cancer Genome Atlas (TCGA) data portal, currently not included in the package. We used the level 3 Affymetrix HG-U133A array data as gene expression levels, and the normalized 415K array as CGH data. We selected a total of 350 samples, as those for which both data types were available. We further focused our attention on a subset of genes with highest variation across the samples (coefficient of variation). As for the CGH probes, we focused on the probes belonging to chromosome 17, which is highly involved in ovarian cancer [24]. We further reduced the complexity of the CGH data by smoothing the original

signal via replacing each value with the median in a window of three elements, using the R function `runmed`. We then selected one probe every 5. After reduction, the data set consisted of $G = 119$ genes and $M = 2265$ probes.

We ran our analysis using the default parameters of the function `iBATProbit`, except for $\boldsymbol{upp}_{\eta} = [-0.3, 0.3, 0.73, \infty]$, $\boldsymbol{low}_{\eta} = [-Inf, -0.3, 0.3, 0.73]$ and $p_{MC} = 5$. The default choice for $\alpha_1 = 1$ and $\alpha_0 = 2.32$ were used. We ran our MCMC for 500,000 iterations and we discarded the first 300,000 as burn-in. We report here the results by selecting associations with $PPI > 0.5$, that is the modal model selected by our method [2]. The selected set of associations contained 16 unique target genes and 22 unique CGH probes. As an example of the output of our analysis, Fig. 1 shows the marginal posterior probability of inclusion for each element of the matrix $\boldsymbol{R}$. The horizontal red line corresponds to the threshold used for our analysis. As for the inference on $\boldsymbol{\xi}$, Fig. 2 shows the frequency of copy number gains (states 3 and 4) and loss (state 1) across the samples for every CGH probe considered in our analysis. Note that there are a large number of CGH probes with high frequency of copy number loss, in line with what reported in the literature [34].

We assessed the biological relevance of our findings by using the database for annotation visualization and integrated discovery (DAVID) tool [11]. In particular, we performed a gene ontology (GO) analysis on the list of selected gene expressions and CGHs separately. We first focused on the enrichment analysis for the gene expressions, see Fig. 3 for a schematic representation of our findings. We found enrichment for six different GO terms. More specifically, we found enrichment

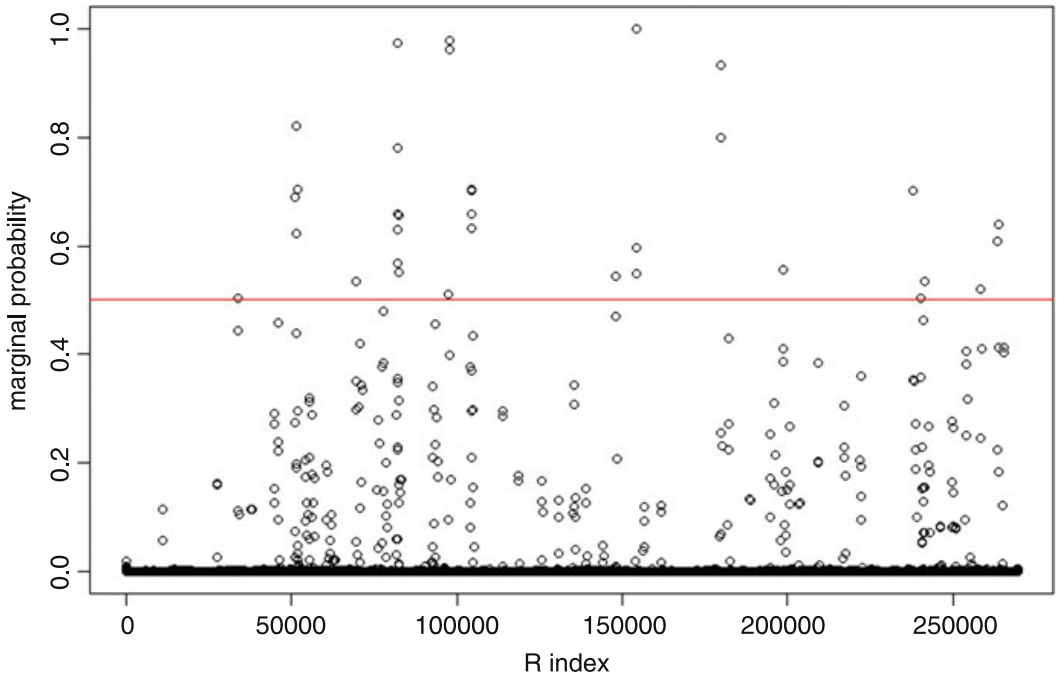

Fig. 1 Marginal posterior probabilities of inclusion of the single elements of the matrix $\boldsymbol{R}$, with the *horizontal red line* corresponding to the threshold used in our analysis

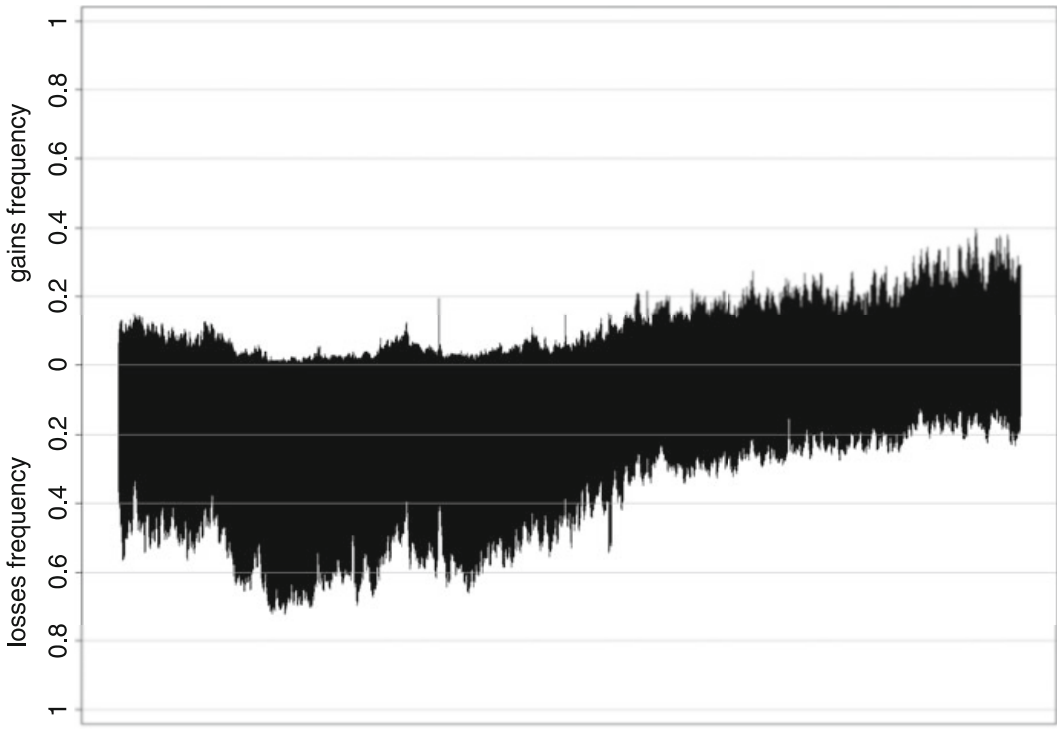

Fig. 2 Frequency of copy number gains (states 3 and 4) and loss (state 1) across the samples for every CGH probe considered in our analysis

for genes that code for skeletal system development, acute phase response, proteolysis, response to wounding, acute inflammatory response and regulation of lipid metabolic process. In particular, genes IGF2 and SERPINA3 are members of the acute phase response, response to wounding, acute inflammatory response and regulation of lipid metabolic process terms, and IGF2 is also a member of the skeletal system development. IGF2 is an insulin-like growth factor and has been previously found inhibited after treatment of responsive RMS-13 cells with 5-fluorouracil and betulinic acid in a study on the effect of the combination of these two acids on ovarian carcinoma cells [37]. SERPINA3, a serpin peptidase inhibitor, has been found over expressed in recurrent ovarian tumors, when compared to the expression in the primary tumor [17]. Focusing on the single genes identified by our analysis a large number of them have been found as associated with tumors in general (HOXB6, FAP, MMP1, GSTA, TAX3, FABP6, SERPINB2, HGMA2) and with ovarian cancer in particular (POSTN). In addition HGMA2 is a promising target for ovarian cancer silencing therapy [20]. Also, although it has not been confirmed yet, MAGEA4 may play a role in embrional development and tumor transformation or aspects of tumor progression [4, 16, 26]. Lastly, we also found enrichment for PPAR signaling KEGG pathway. This pathway inhibits proliferation in PC3 prostate carcinoma cells [30].

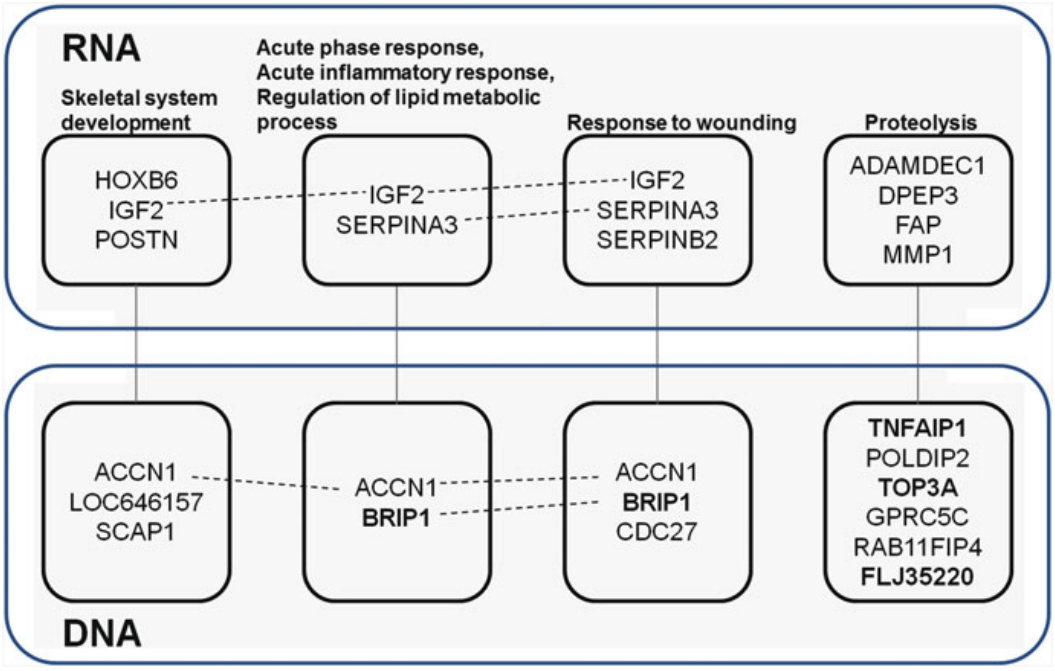

Fig. 3 Schematic representation of a GO analysis of the gene expressions identified by our model, via thresholding the posterior probabilities of inclusion. The *upper box* (labeled RNA) shows the enriched molecular functions together with the corresponding lists of target genes. The *lower box* (labeled DNA) reports the lists of CGH probes that our model found to be associated with the gene expressions. Bold CGH probes highlight those elements found enriched for molecular functions in the GO analysis on the CGH probes. The *solid connecting lines* indicate estimated associations between target genes and CNVs; *dashed lines* indicate genes that appear in multiple lists

As for the enrichment analysis of the CGH probes identified by our method, we found enrichment for six GO terms. More specifically, we found enrichment for the DNA metabolic process, cellular response to stress, DNA repair, cell projection organization, cellular response to DNA damage stimulus and cell projection assembly. A very interesting result is that we found enrichment for three terms that are linked to cellular response to DNA damage (DNA repair and DNA damage stimulus) and external stimulus (cellular response to stress). The DNA repair mechanism is often altered in ovarian cancer. Indeed mutations on the genes BRCA1 and BRCA2 are genes involved in DNA repair, and are very well know for their association with increased breast and ovarian cancer risks [19]. We also found enrichment for two terms related to cell movement and migration, cell projection organization and cell projection assembly. Cellular reorganization is often involved in acquisition of motility ability by cancer cells. This is a key process for the cell invasion and metastasis, two important stages of tumor progression. Further confirmation of our results comes from inspecting the single genes. As an example we identified BRIP1, the official symbol of BRCA1, and MAP2K4, the mitogen-activated protein kinase kinase 4, both well known tumor suppressor genes in ovarian cancer [10, 19]. Also, we identified TNFAIP1, a tumor necrosis factor, SFRS1, a proto oncogene upregulated in various tumors and SUMO2, a gene that play a role in DNA

replication and repair, among other cellular processes, and that has been already associated with many tumors and with ovarian cancer in particular [18, 22].

References

1. Airoldi, E.M., Costa, T., Bassetti, F., Leisen, F., Guindani, M.: Generalized species sampling priors with latent Beta reinforcements. J. Am. Stat. Assoc. **109**(508), 1466–1480 (2014)
2. Barbieri, M.M., Berger, J.O.: Optimal predictive model selection. Ann. Stat. **32**(3), 870–897 (2004)
3. Barnes, C., Plagnol, V., Fitzgerald, T., et al.: A robust statistical method for case-control association testing with copy number variation. Nat. Genet. **40**, 1245–1252 (2008)
4. Brasseur, F., Rimoldi, D., Liénard, D., et al.: Expression of MAGE genes in primary and metastatic cutaneous melanoma. Int. J. Cancer **63**(3), 375–380 (1995)
5. Brown, P., Vannucci, M., Fearn, T.: Multivariate Bayesian variable selection and prediction. J. R. Stat. Soc. Ser. B **60**, 627–641 (1998)
6. Cardin, N., Holmes, C., Donnelly, P., Marchini, J.: Bayesian hierarchical mixture modeling to assign copy number from a targeted CNV array. Genet. Epidemiol. **35**, 536–548 (2011)
7. Cassese, A., Guindani, M., Tadesse, M., Falciani, F., Vannucci, M.: A hierarchical Bayesian model for inference of copy number variants and their association to gene expression. Ann. Appl. Stat. **8**(1), 148–175 (2014)
8. Cassese, A., Guindani, M., Vannucci, M.: A bayesian integrative model for genetical genomics with spatially informed variable selection. Cancer Informat. **13**(S2), 29–37 (2014)
9. Colella, S., Yau, C., Taylor, J., et al.: QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. **35**(6), 2013–2025 (2007)
10. Davis, S.J., Choong, D.Y., Ramakrishna, M., Ryland, G.L., Campbell, I.G., Gorringe, K.L.: Analysis of the mitogen-activated protein kinase kinase 4 (MAP2K4) tumor suppressor gene in ovarian cancer. BMC Cancer **1**(11), 173 (2011)
11. Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., Lempicki, R.A.: DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. **4**(5), P3 (2003)
12. Du, L., Chen, M., Lucas, J., Carlin, L.: Sticky hidden Markov modeling of comparative genomic hybridization. IEEE Trans. Signal Process **58**(10), 5353–5368 (2010)
13. Eddelbuettel, D., Francois, R.: Rcpp: seamless R and C++ integration. J. Stat. Softw. **40**(8), 1–18 (2011)
14. George, E., McCulloch, R.: Approaches for Bayesian variable selection. Stat. Sin. **7**, 339–373 (1997)
15. Guha, S., Li, Y., Neuberg, D.: Bayesian hidden Markov modelling of array cgh data. J. Am. Stat. Assoc. **103**(482), 485–497 (2008)
16. Imaia, Y., Shichijo, S., Yamada, A., Katayama, T., Yano, H., Itoh, K.: Sequence analysis of the MAGE gene family encoding human tumor-rejection antigens. Gene **160**(2), 287–290 (1995)
17. Jinawath, N., Vasoontara, C., Jinawath, A., et al.: Oncoproteomic analysis reveals co-upregulation of RELA and STAT5 in carboplatin resistant ovarian carcinoma. PLoS One **5**(6), e11198 (2010)
18. Karni, R., de Stanchina, E., Lowe, S.W., Sinha, R., Mu, D., Krainer, A.R.: The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat. Struct. Mol. Biol. **14**(3), 185–193 (2007)
19. King, M.C., Marks, J.H., Mandell., J.B.: Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science **302**(5645), 643–646 (2003)
20. Malek, A., Bakhidze, E., Noske, A., et al.: HMGA2 gene is a promising target for ovarian cancer silencing therapy. Int. J. Cancer **132**(2), 348–356 (2008)

21. Monni, S., Tadesse, M.: A stochastic partitioning method to associate high-dimensional responses and covariates. Bayesian Anal. **4**(3), 413–436 (2009)
22. Morris, J.R., Boutell, C., Keppler, M., et al.: The SUMO modification pathway is involved in the BRCA1 response to genotoxic stress. Nature **462**(7275), 886–890 (2009)
23. Newton, M.A., Noueiry, A., Sarkar, D., Ahlquist, P.: Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics **5**(2), 155–176 (2004)
24. Pharoah, P.D., Tsai, Y.Y., Ramus, S.J., et al.: GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. Nat. Genet. **45**(4), 362–370 (2013)
25. Picard, F., Robin, S., Lebarbier, E., Daudin, J.: A segmentation-clustering model for the analysis of array CGH data. Biometrics **63**(3), 758–766 (2007)
26. Resnick, M.B., Sabo, E., Kondratev, S., Kerner, H., Spagnoli, G.C., Yakirevich, E.: Cancer-testis antigen expression in uterine malignancies with an emphasis on carcinosarcomas and papillary serous carcinomas. Int. J. Cancer **101**(2), 190–195 (2002)
27. Rihardson, S., Gilks, W.R.: Conditional independence models for epidemiological studies with covariate measurement error model. Stat. Med. **12**, 1703–1722 (1993)
28. Richardson, S., Bottolo, L., Rosenthal, J.: Bayesian models for sparse regression analysis of high dimensional data. Bayesian Stat. **9**, 539–569 (2010)
29. Sha, N., Vannucci, M., Tadesse, M., et al.: Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. Biometrics **60**(3), 812–819 (2004)
30. Shappell, S.B., Gupta, R.A., Manning, S., et al.: 15S-Hydroxyeicosatetraenoic acid activates peroxisome proliferator-activated receptor gamma and inhibits proliferation in PC3 prostate carcinoma cells. Cancer Res. **61**(2), 497–503 (2001)
31. Stingo, F., Chen, Y., Vannucci, M., Barrier, M., Mirkes, P.A.: Bayesian graphical modelling approach to microRNA regulatory network inference. Ann. Appl. Stat. **4**(4), 2024–2048 (2010)
32. Stingo, F., Chen, Y., Tadesse, M., Vannucci, M.: Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes. Ann. Appl. Stat. **5**(3), 1978–2002 (2011)
33. Subirana, I., Diaz-Uriarte, R., Lucas, G., Gonzalez, J.: CNVassoc: association analysis of CNV data using R. BMC Med. Genomics **4**, 47 (2011)
34. Tavassoli, M., Ruhrberg, C., Beaumont, V., Reynolds, K., Kirkham, N., Collins, W.P., Farzaneh F.: Whole chromosome 17 loss in ovarian cancer. Genes Chromosom. Cancer **8**(3), 195–198 (1993)
35. Venkatraman, E., Olshen, A.: A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics **23**(6), 657–663 (2007)
36. Wang, K., Li, M., Hadley, D., et al.: PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. **17**(11), 1665–1674 (2007)
37. Wang, Y.J., Liu, J.B., Dou, Y.C.: Sequential treatment with betulinic acid followed by 5-fluorouracil shows synergistic cytotoxic activity in ovarian cancer cells. Int. J. Clin. Exp. Pathol. **8**(1), 252–259 (2015)
38. Yau, C., Papaspiliopoulos, O., Roberts, G.O., Holmes, C.: Bayesian nonparametric hidden Markov models with application to the analysis of copy-number-variation in mammalian genomes. J. R. Stat. Soc. Ser. B **73**(1), 37–57 (2011)