Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

A Bayesian Integrative Model for Genetical Genomics with Spatially Informed Variable Selection

Alberto Cassese[1,2], Michele Guindani[2] and Marina Vannucci[1]

[1]Department of Statistics, Rice University, Houston, TX, USA. [2]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

ABSTRACT: We consider a Bayesian hierarchical model for the integration of gene expression levels with comparative genomic hybridization (CGH) array measurements collected on the same subjects. The approach defines a measurement error model that relates the gene expression levels to latent copy number states. In turn, the latent states are related to the observed surrogate CGH measurements via a hidden Markov model. The model further incorporates variable selection with a spatial prior based on a probit link that exploits dependencies across adjacent DNA segments. Posterior inference is carried out via Markov chain Monte Carlo stochastic search techniques. We study the performance of the model in simulations and show better results than those achieved with recently proposed alternative priors. We also show an application to data from a genomic study on lung squamous cell carcinoma, where we identify potential candidates of associations between copy number variants and the transcriptional activity of target genes. Gene ontology (GO) analyses of our findings reveal enrichments in genes that code for proteins involved in cancer. Our model also identifies a number of potential candidate biomarkers for further experimental validation.

KEYWORDS: Bayesian hierarchical models, copy number variants, gene expression, measurement error, variable selection

SUPPLEMENT: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

CITATION: Cassese et al. A Bayesian Integrative Model for Genetical Genomics with Spatially Informed Variable Selection. *Cancer Informatics* 2014:13(S2) 29–37 doi: 10.4137/CIN.S13784.

RECEIVED: February 28, 2014. **RESUBMITTED:** April 10, 2014. **ACCEPTED FOR PUBLICATION:** April 16, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Original Research

FUNDING: Authors disclose no funding sources.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: mguindani@mdanderson.org

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants. Provenance: the authors were invited to submit this paper.

Introduction

Copy number variants (CNVs) are chromosomal aberrations that result in an abnormal number of copies of specific DNA segments in comparison with a reference genome. Studies have reported that as much as 12% of the human genome varies in copy number.[1] It is believed that some CNVs have no obvious phenotypic consequence or are merely related to normal phenotypic variations, while others may be related to genomic disorders and susceptibility to disease. For example, the amplification of a DNA segment in a gene that promotes cell replication may cause the cell to begin dividing excessively, as usually happens in cancer cells.

The challenge of detecting CNVs has received a lot of attention, and several methods have been developed to infer CNVs from high-throughput array-based technologies, such as comparative genomic hybridization (CGH) and single nucleotide polymorphism arrays. These methods mostly rely on hidden Markov models (HMMs)[2,3] and circular binary segmentation.[4] Another question of interest is the identification of CNVs associated with biological functions and complex human diseases. Procedures commonly used include univariate tests or simple linear regression models, with multiple testing correction, to relate the normalized intensity measurements to the outcomes of interest.[3,5] A stochastic partitioning method for a multivariate model has been recently developed.[6]

The model identifies sets of correlated gene expression levels and sets of chromosomal aberrations that jointly affect mRNA transcript abundances. A disadvantage of all such methods is that they do not infer copy number states. Indeed, the high noise level in the raw signal intensities may lead to the identification of a large number of false positives (FPs).[7] An approach widely used to address this problem is to perform the analysis in two steps, first by estimating the copy number states and then using those as the true states in a subsequent association analysis. However, using the estimated copy numbers as if they were the true states ignores the uncertainty in the estimation process and can introduce bias. Some methods that incorporate the uncertainty in copy number estimation into the association analysis have been proposed.[8,9]

Here, we consider a Bayesian hierarchical model that handles CNV detection and association analysis in a unified manner, by integrating array CGH and gene expression data collected on the same set of subjects. The framework takes advantage of a recently proposed measurement error model[10] that relates the gene expression levels to latent copy number states. In turn, the latent states are related to the observed surrogate CGH measurements via an HMM. The model incorporates a variable selection procedure with a prior distribution on the latent selection indicator that exploits dependencies across adjacent DNA segments. In this study, we investigate an alternative formulation of the spatially dependent variable selection prior, that is the basis of the measurement error model in Ref. 10, and show that it allows for increased flexibility, remarkably easy interpretation of the key parameters and major performance improvements. More specifically, the selection prior that we propose herein is based on a latent probit link; therefore, it can easily accommodate additional available covariate information to improve detection of significant associations. Model fitting and posterior inference are accomplished via Markov chain Monte Carlo (MCMC) stochastic search techniques. We explore the performance of this model in simulations and demonstrate an overall better performance of the model with the newly proposed prior. We also show an application to data from a genomic study on lung squamous cell carcinoma, where we identify potential candidates of associations between CNVs and the transcriptional activity of target genes. GO analyses of our findings reveal enrichments in genes that code for proteins involved in cancer.

The rest of the article is organized as follows: in Section 2, we introduce the integrative Bayesian model and its major components. In Section 3, we report the results from a simulation study and the case study. We conclude with some remarks in Section 4.

Methods

This section is organized as follows. In Section 2.1, we review the integrative framework that we follow in the manuscript. In Section 2.2, we introduce an improved prior model for gene–CGH associations, and in Section 2.3, we describe the model for analyzing copy number aberrations.

Integrative Bayesian hierarchical model. A Bayesian hierarchical model that integrates gene expression levels with CNVs has been recently proposed.[10] The model provides a unified approach for simultaneously inferring copy number states for all samples and identifying associations between sets of gene expression levels and copy number states. Let Y_{ig} denote the expression measurement for gene g ($g = 1,\ldots, G$) and X_{im} the observed CGH measurement, ie, the normalized $\log_2$ intensity ratio, for the mth CGH probe ($m = 1,\ldots, M$), in sample i ($i = 1,\ldots, n$). Let $\mathbf{Z} = [\mathbf{Y}, \mathbf{X}]$ indicate the matrix of all data. In our integrative framework, the observed CGH intensities, X_{im}, are treated as surrogates for the unobserved copy number states, and an HMM accounts for the measurement error in the observed intensities. Let $\boldsymbol{\xi} = [\boldsymbol{\xi}_1,\ldots, \boldsymbol{\xi}_M]$ be the matrix of the latent copy number states. We assume that the CGH probes are ordered according to their chromosomal location and that the elements of the matrix $\boldsymbol{\xi}$ take any of the four possible values,

$\xi_{im} = 1$ for copy number losses;
$\xi_{im} = 2$ for copy neutral states;
$\xi_{im} = 3$ for a single copy gain;
$\xi_{im} = 4$ for multiple copy gains.

We assume that, given the latent states, the observed CGH measurements contain no additional information on the observed gene expression levels. Furthermore, we assume independence of the gene expression measurements, conditional on the copy number states, and independence of the CGHs, given their latent states. Hence, we factorize the likelihood into two components as

$$f(\boldsymbol{Z} \mid \xi) = \prod_{i=1}^{n} \left\{ \prod_{g=1}^{G} f\left(Y_{ig} \mid \xi_i\right) \prod_{m=1}^{M} f\left(X_{im} \mid \xi_{im}\right) \right\} \quad (1)$$

where one component captures the latent structure underlying the CGH intensities and the other component models the association between the resulting copy number states and the gene expression levels. Such joint modeling reduces the bias that arises when the uncertainty in the CNV estimation process is ignored (ie, copy number calls are used as if they were the true states), by allowing for the simultaneous inference of CNVs and their association with gene expression.[10]

Modeling the association between gene expression and CNVs. The model on $\mathbf{Y}$ in the likelihood factorization (1) captures the association between the gene expression levels and the latent CNV states. A commonly used modeling approach assumes a linear regression model of the type

$$Y_{ig} = \mu_g + \xi_i \beta_g + \varepsilon_{ig} \quad (2)$$

where $\mu_1,\ldots, \mu_G$ are gene-specific intercepts, and $\varepsilon_{ig} \sim N(0, \sigma_g^2)$, with σ_g^2 being the gene-specific variance.[6,10,11]

In model (2), we find, for each gene, a parsimonious set of CGH aberrations that most likely affect the gene expression

levels. This can be seen as a variable selection problem. Let **R** be a binary matrix representing the associations, that is r_{gm} is set to 1, if β_{gm} in equation (2) is significantly different than zero, and is set to 0 otherwise. A common Bayesian approach to variable selection employs r_{gm} to define a *spike-and-slab* prior on β_{gm},

$$\pi\left(\beta_{gm} \mid r_{gm}, \sigma_g^2\right) = r_{gm} N\left(0, c_\beta^{-1}\sigma_g^2\right) + \left(1 - r_{gm}\right)\delta_0\left(\beta_{gm}\right) \quad (3)$$

with $\delta_0(\cdot)$ being a point mass at zero.[12–15] The prior model is completed with conjugate distributions on the error precision, $\sigma_g^{-2} \sim G\left(\frac{\delta}{2}, \frac{d}{2}\right)$, and on the intercepts, $\mu_g \mid \sigma_g^2 \sim N\left(0, c_\mu \sigma_g^2\right)$, with δ, d, c_μ, and c_β being hyperparameters to be set.

A key feature in the variable selection construction above is the prior distribution on the latent selection indicator r_{gm} in (3). A mixture of an independent prior, ie, a Bernoulli prior, and a dependent component accounting for dependence between adjacent DNA segments has been proposed.[10] Here, we propose a spatially informed distribution based on a probit link. Contiguous regions with the same non-neutral copy number state are likely to correspond to the same DNA aberration and therefore to jointly affect the expression level of a gene. Accordingly, a spatial prior formulation explicitly assumes that the probability of selection at location m depends on the copy number states and on the selection status of its adjacent probes at positions $\{m-1, m+1\}$. A way of achieving this is to first define a probe-specific quantity that captures information on the physical distance among probes and on the frequency of change points at position m in copy number states across all samples as

$$s_{(m-1)m} = \frac{1}{n}\sum_{i=1}^{n} \frac{e^{1-\frac{d_m}{D}} - 1}{e - 1} \mathbf{I}\left\{\xi_{im} = \xi_{i(m-1)}\right\}$$

with d_m being the distance between the adjacent probes $\{m-1, m\}$ and D the total length of the DNA fragment (eg, the length of the chromosome) under consideration. Comparable measures of similarity that incorporate physical distances between probes have been reported in the literature on copy number detection.[3,16] In this study, we propose to model the probability that the mth probe is associated with the gth gene through a latent spatial probit regression. More specifically, we assume

$$\pi\left(r_{gm} = 1 \mid r_{g(m-1)}, r_{g(m+1)}, \xi\right) = 1 - \Phi\left(\alpha_0 + \alpha_1 Q_m\right) \quad (4)$$

where Φ indicates the c.d.f. of a standard normal distribution, and Q_m defines a probe-level covariate that quantifies the available information as

$$Q_m = (-1)^{r_{g(m-1)}} s_{(m-1)m} + (-1)^{r_{g(m+1)}} s_{m(m+1)} \quad (5)$$

with α_0 and α_1 being hyperparameters to be set. From equations (4) and (5), some major features of the novel prior can be recognized. First of all, the probability of selection at location m depends on the adjacent probes at positions $\{m-1, m+1\}$. In particular, the probability can either increase or decrease based on the selection status of the adjacent probes, that is, whether they are included or excluded from the model. Furthermore, the amount of increase or decrease depends on the relative distance between probes as well as the frequency of change points observed at each location. For comparison, it is worth noting that the probability of selection in Ref. 10 can only increase when either $r_{g(m-1)}$ or $r_{g(m+1)}$ is selected. Due to the different weighting we propose, our model ensures better false discovery control. In addition, CNVs located in regions of persistent states of aberration are more likely to be jointly associated with the expression levels of a gene, and this effect is more likely with increased proximity of the CGH probes. Within the literature on Bayesian variable selection in linear regression models, interest is being shown to probit-like priors of type (4) as a convenient way to incorporate external information to guide the selection of the predictors.[17] The advantage of this novel formulation lies also in the interpretability of the parameters. The parameter α_0 represents a baseline intercept that can be directly set according to an a priori specified "level of significance" when there are no other covariates. For example, setting $\alpha_0 = 3$ and lacking any other covariate information, the probability of selection is 0.001 under the null distribution of no association (type 1 error). Similarly, α_1 is immediately interpretable as the regression coefficient that captures the strength of the association between adjacent probes. Also as consequence, the probe-specific quantity $s_{(m-1)m}$ has a more direct effect on the probability of selection at location m. In particular, if $s_{(m-1)m} = s_{m(m+1)} = 0$, prior (4) conveniently reduces to $1 - \Phi(\alpha_0)$, which is a Bernoulli distribution that is commonly used in Bayesian variable selection. Finally, the use of a spatial probit regression allows for the possibility of including further covariate information, if available, which can potentially drive the selection of relevant associations, eg, type of cancer, disease stage, probe methylation status, etc. The flexibility and ease of interpretation of the prior (4) and (5) result in simpler prior elicitation as well as improved performance with respect to those of previous proposals, eg,[10] as shown in Section 3.

Modeling copy number aberrations in CGH data via HMM. The model on the CGH data in (1) is defined in terms of the emission probabilities of an underlying HMM. This choice is supported by the typically persistent state observed in copy number data, meaning that copy number losses or gains at a region are often associated with an increased probability of gains and losses at neighboring regions.[3,18–20] We use a four-state HMM and assume that, conditional on the latent states, the CGH intensities are independent and normally distributed, with state-specific means and variances as

$$X_{im} \mid \left(\xi_{im} = j\right) \overset{iid}{\sim} N\left(\eta_j, \sigma_j^2\right) \quad (6)$$

where η_j and σ_j^2, respectively, represent the expected $\log_2$ ratio and the variance for CGH probes in state j ($j = 1,\ldots, 4$).[19] We assume truncated normal and gamma priors for η_j and σ_j^2, respectively. A first-order Markov model captures the dependence between states in adjacent probes as

$$P\left(\xi_{i(m+1)} \mid \xi_{i1},\ldots,\xi_{im}\right) = P\left(\xi_{i(m+1)} \mid \xi_{im}\right) = a_{\xi_{im}\xi_{i(m+1)}}$$

with $\mathbf{A} = (a_{hj})$ being the matrix of transition probabilities with strictly positive elements ($h, j = 1,\ldots, 4$) and stationary distribution, π_A. The initial state probabilities are also assumed to be given by π_A. We assume that the rows of $\mathbf{A}$ are independent, each following a Dirichlet distribution, $\pi_A \sim Dir(a_1, a_2, a_3, a_4)$, for some $a_j > 0, j = 1,\ldots, 4$.

Posterior inference. For posterior inference, we rely on an MCMC stochastic search algorithm.[10] Our primary interest lies in the estimation of the association matrix $\mathbf{R}$ and the matrix of copy number states $\boldsymbol{\xi}$. Therefore, the remaining model parameters can be integrated out, both to simplify the sampler and improve the mixing of the chain.[13,14,21] Here, in particular, once we integrate out μ_g, β_g, and σ_g^2, an MCMC algorithm can be designed as follows:

1. Update $\mathbf{R}$ using a Metropolis algorithm by randomly selecting n_g genes and proposing, for each gene, a change in its inclusion status by an add/delete/swap move.
2. Update $\boldsymbol{\xi}$ using a Metropolis–Hastings algorithm by randomly choosing a column and proposing new states for a subset of its elements using the current values of the transition matrix.
3. Update the emission distribution parameters, η_j and σ_j^2, using Gibbs sampling.
4. Update the transition probability matrix, $\mathbf{A}$, using a Metropolis algorithm.

Metropolis–Hastings stochastic search algorithms of this type have been used extensively in the Bayesian variable selection literature.[10–15] The update on $\mathbf{R}$ can be made more efficient by selecting at random a subset of the rows and then performing an add/delete or swap move for every row in the subset. Also, for the update on $\boldsymbol{\xi}$, CGH probes called in copy-neutral states in more than $n \times p_C$ samples at the current MCMC iteration (with p_C set by the user) can be disregarded, since these would not be expected to be associated with changes in mRNA transcript abundance.

Given the output of the MCMC, for each element of $\boldsymbol{R}$, we can estimate its marginal posterior probability of inclusion (PPI), $p(r_{gm} = 1|\text{data})$, by averaging the number of iterations where the element was set to 1. We can then select the most relevant associations by thresholding the PPIs based on some decision theoretic criterion. Finally, we can estimate each element of $\boldsymbol{\xi}$ as the modal state across the MCMC iterations.

Applications

Simulation study. In this section, we assess the performance of our model on simulated data. For comparison purposes, we follow the simulation scheme in Ref. 10, which reflects the understanding that single copy number aberrations typically affect segments of DNA, and that neighboring chromosomal locations are expected to share similar copy number states. In addition, transitions to the normal diploid state are more likely than transitions between different states of copy number aberration. Accordingly,

- we set $M = 1000$, $G = 100$, and $n = 100$;
- we initialize the matrix $\boldsymbol{\xi}$ with all elements set to 2;
- we select $L < M$ columns at random in batches of adjacent columns and generate their values using the following transition matrix,

$$\begin{pmatrix} 0.7500 & 0.1800 & 0.0500 & 0.020 \\ 0.4955 & 0.0020 & 0.4955 & 0.007 \\ 0.0200 & 0.1800 & 0.7000 & 0.010 \\ 0.0001 & 0.3028 & 0.1000 & 0.597 \end{pmatrix}$$

- we randomly select half of the remaining columns and, for each column, generate 10% of its positions according to the above transition matrix;
- we generate the elements of matrix $\mathbf{X}$ according to (6), fixing $\eta_1 = -0.65$, $\eta_2 = 0$, $\eta_3 = 0.65$, $\eta_4 = 1.5$ and $\sigma_1 = 0.1$, $\sigma_2 = 0.1$, $\sigma_3 = 0.1$, $\sigma_4 = 0.2$;
- we obtain the matrix of true associations, $\mathbf{R}$, by selecting two clusters of 20 adjacent CGH probes among the L columns previously selected from $\mathbf{X}$, and fix the corresponding values in R at 1. All other elements of R are set to 0;
- we generate the non-zero regression coefficients as $\beta \sim N(0.5, 3^2)$;
- finally, we generate the gene expression levels as $Y_{ig} = \mu_g + \boldsymbol{\xi}_i\boldsymbol{\beta}_g + \varepsilon_{ig}$, with $\mu_g \sim N(0, 0.1^2)$ and $\varepsilon_{ig} \sim N(\mathbf{0}, \sigma_\varepsilon^2)$. We consider two different settings for the random error standard deviation: $\sigma_\varepsilon = \{0.1, 0.5\}$.

For setting the hyperparameters, we follow the general guidelines in similar regression models for the specification of the priors on the parameters μ_g, β_g, and σ_g^2 and on the HMM parameters η_j, σ_j, and a_j.[13,14,19] For the hyperparameters of the probit prior (4), we set $\alpha_0 = 3$, which is equivalent to a prior probability of selecting 0.001 when $\alpha_1 = 0$, and then perform a sensitivity analysis on the choice of α_1. More specifically, we consider values of α_1 in the set {0, 0.5, 1, 1.5, 2}. The results we report were obtained by running MCMC chains with 1,000,000 iterations and a burn-in of 500,000. Using a dual-core Intel® Xeon® processor with 16 GB of memory, 2.2 GHz, our code takes approximately 2 minutes to run 10,000 iterations.

We begin the analysis of the simulation results by focusing on the inference on **R**. We compute an estimate of the Bayesian false discovery rate (FDR_B) and use a threshold on the PPIs that controls the false discovery rate at the 0.05 level.[22] In Table 1, we report the results in terms of specificity, sensitivity, FP counts, and false negative (FN) counts. Sensitivity is defined as the ratio of true positive (TP) counts over the total number of true connections, and specificity is defined as the ratio of true negative (TN) counts over the number of true missing connections. We also report the realized Bayesian *q value,* defined as $\min_{1-PPI \le k} FDR_B(k)$, and the Matthew correlation coefficient (MCC), calculated as

$$MCC = \frac{TP \times TN + FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

The results show that values of α_1 in the range [1, 1.5] lead to excellent performances, compared with both the independent prior (ie, $\alpha_1 = 0$) and higher values of α_1, particularly for the larger σ_ε value. Overall, our results are consistently better than those obtained with competing models.[10] This can be seen by the low number of FP detections for most of the parameter values in the range considered. Also, for some values of α_1, our prior (4) and (5) achieves perfect classification, with specificity and sensitivity equal to 1. To investigate the effect of the threshold on the PPIs on the selection results, in Figure 1, we report receiver operating characteristic (ROC)-type curves that display the FP counts versus the FN counts, calculated at a grid of equispaced thresholds in the interval [0.07, 1]. The plots clearly show that our results are satisfactory across different thresholds.

Table 1. Simulated data: Results on specificity, sensitivity, number of false positives and false negatives, MCC, Bayesian *q* values, and number of detections obtained for an FDR threshold of 0.05.

$\sigma_\varepsilon = 0.1$					
α_1 VALUE	$\alpha_1 = 0$	$\alpha_1 = 0.5$	$\alpha_1 = 1$	$\alpha_1 = 1.5$	$\alpha_1 = 2$
Specificity	0.99999	1	1	0.99998	0.99992
Sensitivity	0.85	0.95	1	0.95	1
FP/FN	1/3	0/1	0/0	2/1	8/0
MCC	0.89596	0.97467	1	0.92709	0.84512
q-value	0.03624	0.01830	0.04425	0.02619	0.04445
# of detections	18	19	20	21	28
$\sigma_\varepsilon = 0.5$					
α_1 VALUE	$\alpha_1 = 0$	$\alpha_1 = 0.5$	$\alpha_1 = 1$	$\alpha_1 = 1.5$	$\alpha_1 = 2$
Specificity	0.99998	0.99999	0.99999	1	0.99992
Sensitivity	0.6	0.7	0.95	1	1
FP/FN	2/8	1/6	1/1	0/0	8/0
MCC	0.71709	0.80826	0.94999	1	0.84512
q-value	0.03198	0.04553	0.03101	0.02579	0.04444
# of detections	14	15	20	20	28

They also highlight the consistently worse performance of the independent prior.

As for the inference on ξ, Table 2 reports the misclassification counts and corresponding percent rates obtained by considering, for each element of $\boldsymbol{\xi}$, the modal state attained at that genomic location over all MCMC iterations (after burn-in). The performances are consistently good. For the case of $\alpha_1 = 0$ and $\sigma_\varepsilon = 0.1$, our estimated means and standard deviations were $\hat{\eta} = [-0.65151, 0.00017, 0.65128, 1.49925]$ and $\hat{\sigma} = [0.10180, 0.09974, 0.10050, 0.20896]$, respectively, which are consistent with the values used to simulate the data. We obtained similar estimates in all the other cases.

Lung cancer study. We applied our Bayesian model to data from a study of lung squamous cell carcinoma, which we obtained from The Cancer Genome Atlas data portal (https://tcga-data.nci.nih.gov/tcga/). We used the level 2 (normalized signals) Agilent 415K array as the CGH data, and the Affymetrix HG-U133A array as the gene expression levels. We performed our analysis on the 131 samples that were available for both data types. We considered CGH probes belonging to chromosome 3, as it has been highly implicated in lung squamous cell carcinoma.[23,24] We further reduced the complexity of the data by filtering out genes and CGH probes that had a relatively small coefficient of variation (smaller than 1.9 and 0.35, in absolute value, for genes and CGH probes, respectively). The resulting data consisted of $G = 133$ genes and $M = 2{,}133$ CGH probes.

We ran our model using a setting similar to that adopted in the simulated example described in Section 3.1. The results we report below were obtained by setting $\alpha_1 = 1$ and $\alpha_0 = 2.32$ and by running the MCMC sampler with 500,000 iterations and a burn-in of 250,000. Figure 2 shows a heatmap of the highest PPIs of gene–CNV associations corresponding to the elements of the association matrix **R** with a PPI larger than 0.1. As expected, despite the large number of potential associations being investigated, few have relatively large PPIs. Figure 3 shows the estimated frequencies of copy number gains and losses for each of the 2,133 CGH probes considered in our analysis, as is commonly done in the literature.[25–29] In the figure, single and multiple copy gains are considered together as copy number amplifications. The estimates of the state-specific means and variances were close to their theoretical values (results not shown). Based on Figure 3, we can identify 67 probes with high-frequency (>45%) amplification and 23 probes with high-frequency (>25%) deletion. Among the identified probes, there are 36 and 13 annotated genes for amplification and deletion, respectively. Interestingly, one of those genes (DVL3) shows both high-frequency deletion and amplification, and has been recently found to be involved in lung squamous cell carcinoma.[30] Other genes detected by our method have been implicated in lung cancer, for example, EPHA6, CENTB2, and ZNF717 for high-frequency amplification, and PLD1 and ATP2C1 for high-frequency deletion.[31–34]

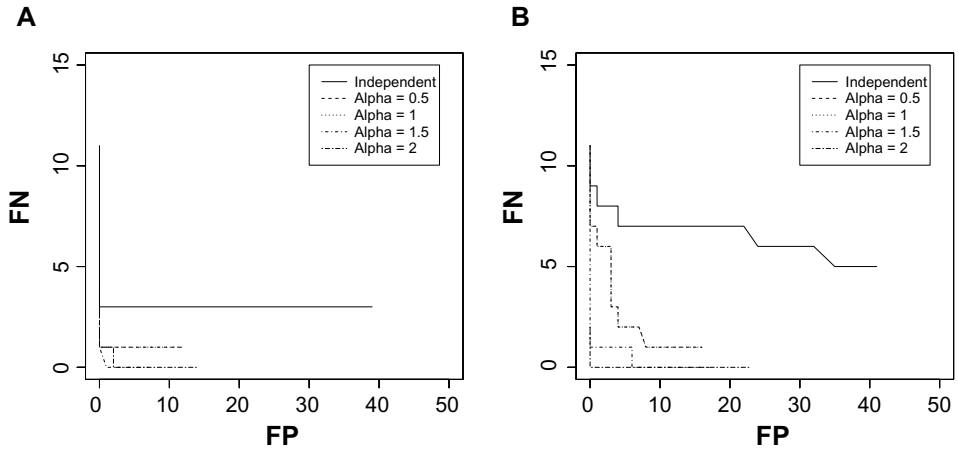

Figure 1. Simulated data: The false positive (FP) and false negative (FN) counts obtained by considering different thresholds on the marginal probabilities, for (**A**) $\sigma_\varepsilon = 0.1$ and (**B**) $\sigma_\varepsilon = 0.5$. Threshold values are calculated as a grid of equispaced points in the range [0.07, 1].

Our findings identify potential candidates of associations between CNVs and the transcriptional activity of target genes. In order to assess whether the identified associations have biological relevance, we performed GO analyses on the lists of selected target genes and CGH probes, by using the database for annotation visualization and integrated discovery (DAVID) tool.[35] We report the detailed results of the analyses in the Supplementary Material. Figure 4 shows some of the results from the enrichment analysis of the list of selected target genes. More specifically, the upper box of the figure (labeled mRNA) reports the four most relevant molecular functions, together with the corresponding lists of target genes. In the lower box (labeled DNA), we report the lists of CGH probes that our model found to be associated with the target genes. The estimated associations between target genes and CNVs are marked by solid lines; whereas probes appearing in multiple lists are indicated by dashed lines. In Figure 5, we report similar summaries from the gene enrichment analysis of the selected CGH probes. Specifically, in this figure, the upper box shows the molecular functions enriched in the list of CGH probes, and the lower box reports the list of target genes that our model found to be associated with the CGH probes.

The results from the GO analyses highlight the enrichment of genes that code for proteins with binding function, cell surface binding, or an extracellular matrix constituent in the selected target genes (Fig. 4), and the enrichment of genes that code for proteins in the signal transduction machinery, mainly with kinase activity, in the selected CGH probes (Fig. 5). In both cases, we identified genes as members of the ephrin family or NTRK, which have been shown to be altered in another study on lung adenocarcinoma.[36] Ephrin receptors have been shown to have an important role in tumor growth and progression in many cancers, including lung carcinoma.[37] Another relevant protein from the GO analyses is PIK3CB, phosphatidylinositol-4,5-bisphosphate 3-kinase. The PI3K/AKT1 pathway has been shown to be altered in many cancer types, and often correlates with a more aggressive form of disease.[38–41] We also found proteins of the matrix metalloproteinase family, which are often involved in the induction and promotion of cancer cell migration (MMP10 and ADAM23).[42,43] Combining this observation with the finding of alterations in genes that code for members of the fibrinogen family (FGA, FGB, and FGG), and in genes that code for proteins with surface- and matrix-binding properties, we may hypothesize a dysregulation of pathways involved in the acquisition of a migratory phenotype. Extracellular matrix remodeling plays an important role in cancer progression since it can facilitate the migration and invasion of tumor cells. The genes we found to be altered may play an important role in this context. Such a hypothesis is interesting, but will require further experimental investigation. Similar findings exist in the general literature on lung cancer in both human and mouse studies.[44–49] Finally, many of the genes we identified have

Table 2. Simulated data: Results on ξ as the number of misclassified copy number states for various values of α_1.

# MISCLASSIFICATIONS (PERCENT)	$\alpha_1 = 0$	$\alpha_1 = 0.5$	$\alpha_1 = 1$	$\alpha_1 = 1.5$	$\alpha_1 = 2$
Scenario with $\sigma_\varepsilon = 0.1$	60 (0.06%)	55 (0.055%)	62 (0.062%)	53 (0.053%)	56 (0.056%)
Scenario with $\sigma_\varepsilon = 0.5$	48 (0.048%)	54 (0.054%)	51 (0.051%)	49 (0.049%)	52 (0.052%)

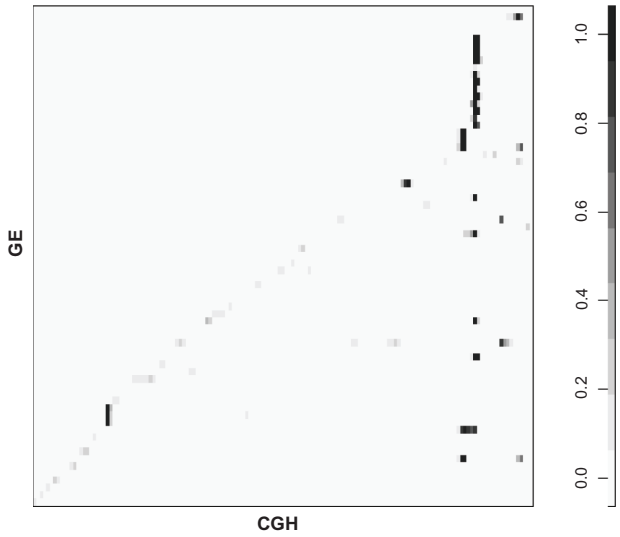

Figure 2. Case study: Heatmap of the highest PPIs of gene–CNV associations, selected by looking at the elements of the association matrix ***R*** that have a PPI greater than 0.1.

been reported in the literature on lung cancer, for example ASCL1, HLA-DQA1, and PROM1 among the gene expression probes, and EPAH3, PRKCI, and EPHB1 among the identified CGH probes.[36,50–54]

Conclusions

In this study, we have considered a recently developed Bayesian hierarchical framework for the integration of gene expression levels with CGH array measurements, collected on the same subjects. The proposed measurement error model relates the gene expression levels to latent copy number states which, in turn, are related to the observed surrogate CGH measurements via an HMM. We have investigated an alternative formulation of the spatial variable selection prior for the gene–CGH associations. Our prior exploits dependencies across adjacent DNA segments and allows for increased modeling

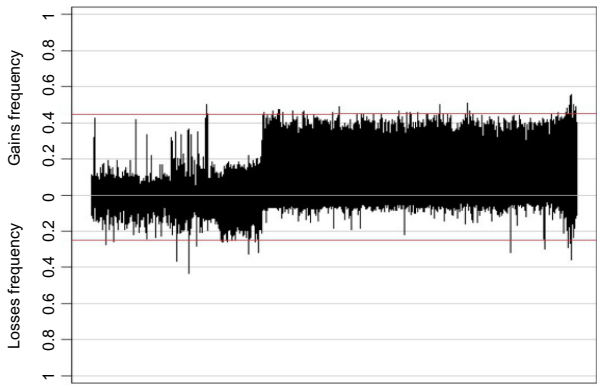

Figure 3. Case study: Frequencies of estimated gains and losses among the 131 samples for the 2,133 CGH probes considered in our analysis. Red horizontal lines correspond to the 0.25 and 0.45 thresholds on the frequencies of deletion and amplification, respectively.

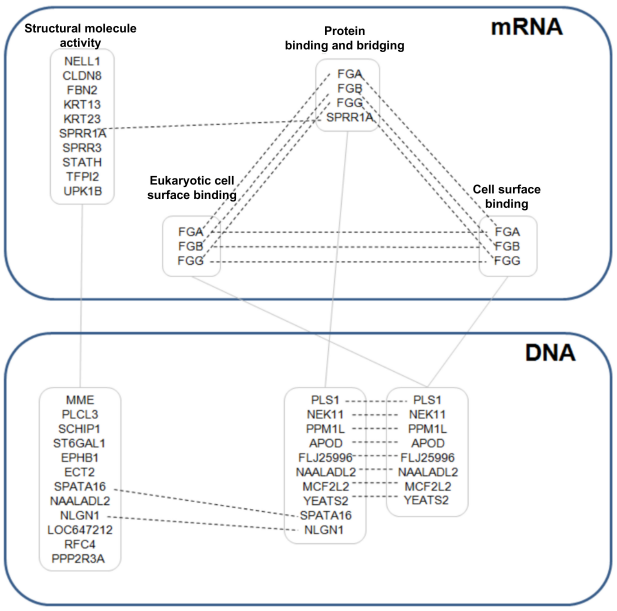

Figure 4. Case study: Schematic representation of a GO analysis of the target genes identified by our model, via thresholding the posterior probabilities of inclusion. The upper box (labeled mRNA) shows the four most enriched molecular functions together with the corresponding lists of target genes. The lower box (labeled DNA) reports the lists of CGH probes that our model found to be associated with the target genes.
Notes: The solid connecting lines (———) indicate estimated associations between target genes and CNVs; dashed lines (– – –) indicate probes that appear in multiple lists.

flexibility, which has been shown to result in easy interpretable model parameters for the purpose of prior elicitation, as well as improved performances and false discovery control on simulated data. Our HMM model considers four copy number aberration states, as commonly encountered in the literature.[19,55,56] Once the HMM states are appropriately defined, our model can easily accommodate an additional state for the loss of both copies.[3,18]

We have presented an application to data from a genomic study on lung squamous cell carcinoma. Our model has identified potential candidates of associations between CNVs and the transcriptional activity of target genes. We have assessed the biological relevance of our findings through GO analyses. These have revealed enrichments in genes that code for proteins involved in cancer, such as those of the ephrin family, phosphatidylinositol-4,5-bisphosphate 3-kinase and matrix metalloproteinase family. Among these, some are already known to be involved in lung squamous cell carcinoma, while others are interesting potential candidates for further experimental validation.

The approach we present can be extended to the analysis of RNA-Seq gene expression values. In order to appropriately take into account the nature of such data, the priors and the algorithm for posterior inference will need to be modified to accommodate the count data and a Poisson

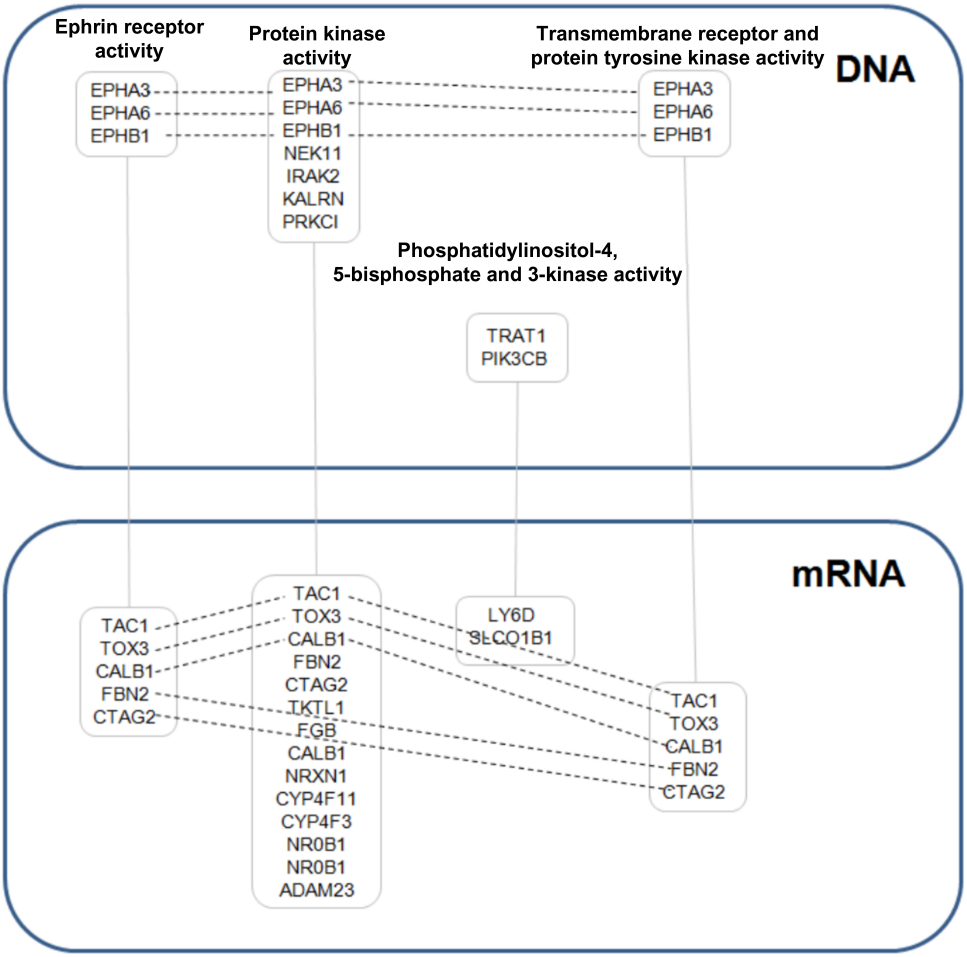

Figure 5. Case study: Schematic representation of a GO analysis of the CGH probes identified by our model, via thresholding the posterior probabilities of inclusion. The upper box (labeled DNA) shows the four most enriched molecular functions together with the corresponding lists of CGH probes. The lower box (labeled mRNA) reports the lists of target genes that our model found to be associated with the CGH probes.
Notes: Solid connecting lines (———) indicate the estimated associations between target genes and CNVs; dashed lines (– – –) indicate probes that appear in multiple lists.

regression model. This represents an interesting avenue for future work.

Author Contributions

Conceived and designed the experiments: No experiments. Analyzed the data: AC, MG, MV. Wrote the first draft of the manuscript: AC, MG, MV. Contributed to the writing of the manuscript: AC, MG, MV. Agree with manuscript results and conclusions: AC, MG, MV. Jointly developed the structure and arguments for the paper: AC, MG, MV. Made critical revisions and approved final version: AC, MG, MV. All authors reviewed and approved of the final manuscript.

Supplementary Data

GO Analyses. The supplementary material shows details on the GO analyses described in the paper.

REFERENCES

1. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature*. 2006;444:444–54.
2. Colella S, Yau C, Taylor JM, et al. QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*. 2007;35(6):2013–25.
3. 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*. 2007;17(11):1665–74.
4. Venkatraman E, Olshen A. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*. 2007;23:657–63.
5. Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007;315:848–53.
6. Monni S, Tadesse M. A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Anal*. 2009;4(3):413–36.
7. Breheny P, Chalise P, Batzler A, Wang L, Fridley BL. Genetic association studies of copy-number variation: should assignment of copy number states precede testing? *PLoS One*. 2012;7(4):e34262.
8. Barnes C, Plagnol V, Fitzgerald T, et al. A robust statistical method for case-control association testing with copy number variation. *Nat Genet*. 2008;40:1245–52.
9. Subirana I, Diaz-Uriarte R, Lucas G, Gonzalez J. CNVassoc: association analysis of CNV data using R. *BMC Med Genomics*. 2011;4:47.
10. 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. *Annals of Applied Statistics*. 8(1):148–75.
11. Richardson S, Bottolo L, Rosenthal J. Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Stat*. 2010;9:539–69.
12. George E, McCulloch R. Approaches for Bayesian variable selection. *Stat Sin*. 1997;7:339–73.
13. Brown P, Vannucci M, Fearn T. Multivariate Bayesian variable selection and prediction. *J R Stat Soc Series B Stat Methodol*. 1998;60:627–41.

14. Sha N, Vannucci M, Tadesse MG, et al. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*. 2004;60(3):812–19.
15. Stingo F, Chen Y, Vannucci M, Barrier M, Mirkes P. A Bayesian graphical modelling approach to microRNA regulatory. *Ann Appl Stat*. 2010;4(4):2024–48.
16. Marioni JC, Thorne NP, Tavare S. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*. 2006;22(9):1144–6.
17. Quintana MA, Conti DV. Integrative variable selection via Bayesian model uncertainty. *Stat Med*. 2013;32:4938–53.
18. Wang K, Chen Z, Tadesse MG, et al. Modeling genetic inheritance of copy number variations. *Nucleic Acids Res*. 2008;36(21):e138.
19. Guha S, Li Y, Neuberg D. Bayesian hidden Markov modelling of array CGH data. *J Am Stat Assoc*. 2008;103(482):485–97.
20. Yau C, Papaspiliopoulos O, Roberts GO, Holmes C. Bayesian non-parametric hidden Markov models with application to the analysis of copy-number-variation in mammalian genomes. *J R Stat Soc Series B Stat Methodol*. 2011;73(1):37–57.
21. 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*. 2011;5(3):1978–2002.
22. Newton M, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*. 2004;5(2):155–76.
23. Bass AJ, Watanabe H, Mermel CH, et al. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nat Genet*. 2009;41(11):1238–42.
24. Stead LF, Berri S, Wood HM, et al. The transcriptional consequences of somatic amplifications, deletions, and rearrangements in a human lung squamous cell carcinoma. *Neoplasia*. 2012;4(4):2024–48.
25. Broët P, Tan P, Alifano M, Camilleri-Broët S, Richardson S. Finding exclusively deleted or amplified genomic areas in lung adenocarcinomas using a novel chromosomal pattern analysis. *BMC Med Genomics*. 2009;2:43.
26. Beroukhim R, Getz G, Nghiemphu L, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci USA*. 2007;104(50):20007–12.
27. Diskin SJ, Eck T, Greshock J, et al. STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res*. 2006;16(9):1149–58.
28. Klijn C, Holstege H, de Ridder J, et al. Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res*. 2008;36(2):e13–6.
29. Chin K, DeVries S, Fridlyand J, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*. 2006;10(6):529–41.
30. Wang J, Qian J, Hoeksema MD, et al. Integrative genomics analysis identifies candidate drivers at 3q26–29 amplicon in squamous cell carcinoma of the lung. *Clin Cancer Res*. 2013;19(20):5580–90.
31. Rudin CM, Durinck S, Stawiski EW, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat Genet*. 2012;44(10):1111–6.
32. Lockwood WW, Wilson IM, Coe BP, et al. Divergent genomic and epigenomic landscapes of lung cancer subtypes underscore the selection of different oncogenic pathways during tumor development. *PLoS One*. 2012;7(5):e37775.
33. Liu P, Morrison C, Wang L, et al. Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis*. 2012;33(7):1270–6.
34. Ahn MJ, Park SY, Kim WK, et al. A single nucleotide polymorphism in the phospholipase D1 gene is associated with risk of non-small cell lung cancer. *Int J Biomed Sci*. 2012;8(2):121–8.
35. Dennis G, Sherman B, Hosack D, et al. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*. 2003;4:3.
36. Ding L, Getz G, Wheeler DA, et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*. 2008;455(7216):1069–75.
37. Giaginis C, Tsoukalas N, Bournakis E, et al. Ephrin (Eph) receptor A1, A4, A5 and A7 expression in human non-small cell lung carcinoma: associations with clinicopathological parameters, tumor proliferative capacity and patients' survival. *BMC Clin Pathol*. 2014;14(1):8.
38. Cui W, Cai Y, Wang W, et al. Frequent copy number variations of PI3K/AKT pathway and aberrant protein expressions of PI3K subunits are associated with inferior survival in diffuse large B cell lymphoma. *J Transl Med*. 2014;12:10.
39. Drilon A, Rekhtman N, Ladanyi M, Paik P. Squamous-cell carcinomas of the lung: emerging biology, controversies, and the promise of targeted therapy. *Lancet Oncol*. 2012;13(10):e418–26.
40. Burris IIIH. Overcoming acquired resistance to anticancer therapy: focus on the PI3K/AKT/mTOR pathway. *Cancer Chemother Pharmacol*. 2013;71(4):829–42.
41. Pal I, Mandal M. PI3K and Akt as molecular targets for cancer therapy: current clinical outcomes. *Acta Pharmacol Sin*. 2012;33(12):1441–58.
42. Zucker S, Cao J, Chen W. Critical appraisal of the use of matrix metalloproteinase inhibitors in cancer treatment. *Oncogene*. 2000;19(56):6642–50.
43. Gialeli C, Theocharis A, Karamanos N. Roles of matrix metalloproteinases in cancer progression and their pharmacological targeting. *FEBS J*. 2011;278(1):16–27.
44. Li X, Tai H. Increased expression of matrix metalloproteinases mediates thromboxane A2-induced invasion in lung cancer cells. *Curr Cancer Drug Targets*. 2012;12(6):703–15.
45. Justilien V, Regala RP, Tseng IC, et al. Matrix metalloproteinase-10 is required for lung cancer stem cell maintenance, tumor initiation and metastatic potential. *PLoS One*. 2012;7(4):e35040.
46. Yadav V, Awasthi KSV. Preclinical evaluation of 4-[3,5-bis(2-chlorobenzylidene)-4-oxo-piperidine-1-yl]-4-oxo-2-butenoic acid, in a mouse model of lung cancer xenograft. *Br J Pharmacol*. 2013;170(7):1436–48.
47. Hu C, Lv H, Pan G, et al. The expression of ADAM23 and its correlation with promoter methylation in non-small-cell lung carcinoma. *Int J Exp Pathol*. 2011;92(5):333–9.
48. Wang H, Meyer C, Fei T, Wang G, Zhang F, Liu X. A systematic approach identifies FOXA1 as a key factor in the loss of epithelial traits during the epithelial-to-mesenchymal transition in lung cancer. *BMC Genomics*. 2013;14:680.
49. Choi J, Liu H, Song H, Park J, Yun J. Plasma marker proteins associated with the progression of lung cancer in obese mice fed a high-fat diet. *Proteomics*. 2012;12(12):1999–2013.
50. Hou J, Lambers M, den Hamer B, et al. Expression profiling-based subtyping identifies novel non-small cell lung cancer subgroups and implicates putative resistance to pemetrexed therapy. *J Thorac Oncol*. 2012;7(1):105–14.
51. Kohno T, Kunitoh H, Mimaki S, et al. Contribution of the TP53, OGG1, CHRNA3, and HLA-DQA1 genes to the risk for lung squamous cell carcinoma. *J Thorac Oncol*. 2011;6(4):813–17.
52. He Y, Li Y, Qiu Z, et al. Identification and validation of PROM1 and CRTC2 mutations in lung cancer patients. *Mol Cancer*. 2014;13(1):19.
53. Justilien V, Walsh M, Ali S, Thompson E, Murray N, Fields A. The PRKCI and SOX2 oncogenes are coamplified and cooperate to activate hedgehog signaling in lung squamous cell carcinoma. *Cancer Cell*. 2014;25(2):139–51.
54. Dmitriev A, Kashuba V, Haraldson K, et al. Genetic and epigenetic analysis of non-small cell lung cancer with noti-microarrays. *Epigenetics*. 2012;7(5):502–13.
55. Baladandayuthapani V, Ji Y, Talluri R, Nieto-Barajas L, Morris J. Bayesian random segmentation models to identify shared copy number aberrations for array CGH data. *J Am Stat Assoc*. 2010;105(492):1358–75.
56. Shah SP, Xuan X, DeLeeuw RJ, et al. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bionformatics*. 2006;22(14):e431–9.