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)

Bayesian Disease Classification Using Copy Number Data

Subharup Guha[1], Yuan Ji[2,3] and Veerabhadran Baladandayuthapani[4]

[1]Department of Statistics, University of Missouri, Columbia, MO, USA. [2]Center for Biomedical Informatics, North Shore University Health System, Evanston, IL, USA. [3]Department of Health Studies, The University of Chicago, IL, USA. [4]Department of Biostatistics, The University of Texas, MD Anderson Cancer Center, Houston, TX, USA.

ABSTRACT: DNA copy number variations (CNVs) have been shown to be associated with cancer development and progression. The detection of these CNVs has the potential to impact the basic knowledge and treatment of many types of cancers, and can play a role in the discovery and development of molecular-based personalized cancer therapies. One of the most common types of high-resolution chromosomal microarrays is array-based comparative genomic hybridization (aCGH) methods that assay DNA CNVs across the whole genomic landscape in a single experiment. In this article we propose methods to use aCGH profiles to predict disease states. We employ a Bayesian classification model and treat disease states as outcome, and aCGH profiles as covariates in order to identify significant regions of the genome associated with disease subclasses. We propose a principled two-stage method where we first make inferences on the underlying copy number states associated with the aCGH emissions based on hidden Markov model (HMM) formulations to account for serial dependencies in neighboring probes. Subsequently, we infer associations with disease outcomes, conditional on the copy number states, using Bayesian linear variable selection procedures. The selected probes and their effects are parameters that are useful for predicting the disease categories of any additional individuals on the basis of their aCGH profiles. Using simulated datasets, we investigate the method's accuracy in detecting disease category. Our methodology is motivated by and applied to a breast cancer dataset consisting of aCGH profiles assayed on patients from multiple disease subtypes.

KEYWORDS: breast cancer, classification, Bayesian network, hidden Markov model

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

CITATION: Guha et al. Bayesian Disease Classification Using Copy Number Data. *Cancer Informatics* 2014:13(S2) 83–91 doi: 10.4137/CIN.S13785.

RECEIVED: April 15, 2014. **RESUBMITTED:** July 29, 2014. **ACCEPTED FOR PUBLICATION:** July 29, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Original Research

FUNDING: This work was supported by the National Science Foundation under award DMS-0906734 to SG. YJ's research was supported by NIH R01 CA132897. VB's research is partially supported by NIH grant R01 CA160736 and the Cancer Center Support Grant (CCSG) (P30 CA016672). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation, National Cancer Institute, or the National Institutes of Health.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: guhasu@missouri.edu

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

DNA copy number variations (CNVs) have been shown to be associated with cancer development and progression.[1] Somatic CNVs can lead to tumorigenesis. For example, loss of copy numbers for tumor suppressor genes or amplification for oncogenes both lead to cancer. The detection of these CNVs has the potential to impact the basic knowledge and treatment of many types of cancers, and can play a role in the discovery and development of molecular-based personalized cancer therapies.[2]

In early years, cytogeneticists have been limited to traditionally visually examining whole genomes with a microscope, a technique known as karyotyping or chromosome analysis. In the mid-70 s and 80s, the development and application of molecular diagnostic methods such as Southern blots, polymerase chain reaction (PCR), and fluorescence in situ hybridization (FISH) allowed clinical researchers to make many important advances in genetics, including clinical cytogenetics. However, these techniques have several limitations. First, they are very time consuming and labor intensive, and only a limited number and regions of the chromosome can be tested simultaneously. Further, because the probes are targeted to specific chromosome regions, the analysis requires

prior knowledge of an abnormality and is of limited use for screening complex karyotypes. More recently, scientists have developed techniques that integrate aspects of both traditional and molecular cytogenetic techniques called *chromosomal microarrays*.[3] These high-throughput high-resolution microarrays have allowed researchers to diagnose numerous subtle genome-wide chromosomal abnormalities that were previously undetectable and find many cytogenetic abnormalities in part or all of a single gene. Such information is useful for biologists to detect new genetic disorders and also provide a better understanding of the pathogenetic mechanisms of many chromosomal aberrations.

One of the most common types of high-resolution chromosomal microarrays are array-based comparative genomic hybridization (aCGH) methods that assay DNA CNVs across the whole genomic landscape in a single experiment.[4] With aCGH, differentially labeled test and reference samples' genomic DNAs are cohybridized to normal chromosomes, and fluorescence intensities/ratios along the length of chromosomes provide a cytogenetic representation of the relative DNA CNV across the whole genome. Whereas early aCGH arrays were mainly used in research settings, recent improvements in algorithms for aCGH data analysis as well as rapidly reducing costs now enable clinical applications of aCGH arrays, particularly in the study of cancer genomic as a diagnostic tool.[2]

In this article, we propose methods to use aCGH profiles to predict disease states. We employ a Bayesian classification model, and treat disease states as outcome and aCGH profiles as covariates – to identify significant regions of the genome associated with disease subclasses. Statistical challenges for aCGH classification include not only high dimensionality ie, large number (tens of thousands) of probes but also relatively small number of samples, more importantly, the presence of serial correlation among the features – nearby probes (by genomic location) tend to be highly correlated. Classical methods usually used for multivariate classification of high-dimensional genomic data, eg, penalized approaches (Zhu and Hastie[5] and the references there-in), do not account for the specific structure of aCGH data, as they ignore the serial dependence in the probes. To exploit the serial genomic information, typical approaches first segment the data[6] and then conduct downstream classification. Alternative methods are based on kernel-based techniques such as support vector machine (SVM),[7] and its variants exploit genomic continuity.[8] While incorporating excellent prediction capabilities, these methods do not explicitly utilize the inherent discrete nature of the latent copy number states (gain/loss/normal) in their variable selection procedures, which serves as one of the primary aims in this article.

In the Bayesian framework, several innovative variable selection strategies have been developed in various contexts, with reasonable degrees of success. Some of these approaches can be regarded as *linear variable selection methods*. These include stepwise selection,[9] penalized regression approaches such as lasso (and its variants),[10] and non-concave penalized likelihood approaches.[11] The technique applied in this paper is based on Bayesian linear variable selection approaches, including spike and slab mixture priors,[12] stochastic search variable selection,[13] Gibbs-based variable selection,[14] Bayesian model averaging,[15,16] and indicator priors.[17] The stochastic search variable selection approach of George and McCulloch[13] has been extended to multivariate settings by Brown et al.[18] and to generalized linear mixed models by Cai and Dunson.[19] Effective variable selection methods have also been developed for multinomial probit models by Sha et al.[20] and for microarray data with censored outcomes by Lee and Mallick[21] and Sha et al.[22] However, none of these approaches account for natural spatial/serial dependency in the covariates (as in our case) – which might lead to biased estimates.

In this article we propose a principled two-stage method for disease classification using covariates exhibiting serial dependence. In general, the technique is applicable to datasets having the following structure. For individuals $i = 1, \ldots, n$, we have (i) two disease categories coded as the binary response y_i and (ii) aCGH emissions $e_{i1}, \ldots, e_{ip}$ corresponding to p probes, with p typically being much larger than n. The analysis broadly consists of two stages. In Stage 1, we make inferences on underlying copy number states associated with the aCGH emissions based on hidden Markov model (HMM) formulations[23] to account for serial dependencies. Subsequently in Stage 2, we analyze the model parameters associated with the binary responses, conditional on the parameters discovered in Stage 1, using Bayesian linear variable selection procedures. In particular, we select the aCGH probes having a linear regression relationship with the disease categories. The selected probes and their effects are parameters that are useful for predicting the disease categories of any additional individuals on the basis of their aCGH emissions. Our methodology is motivated by and applied to a dataset consisting of 111 breast cancer patients[24] and falling into two disease subgroups, ER+ and triple negative (TN). There are 56 TN patients and 55 ER+ patients. For each patient, DNA copy number data were generated using Agilent 4x44K CGH arrays (available at ArrayExpress accession number E-TABM-484).

The remainder of the paper is organized as follows. Section 2 provides details of the model for the two-stage analysis. Section 3 develops the posterior inference and prediction technique based on Markov chain Monte Carlo (MCMC) methods. In Section 4, using simulated datasets, we investigate the method's accuracy in detecting disease category. Finally, Section 5 analyzes the motivating breast cancer dataset and makes test case predictions.

Model

Our modeling framework consists of two stages: In Stage 1, we model the aCGH emissions, relying on HMMs to account for the serial correlations among the emissions. Then,

in Stage 2, the relationship between the HMM parameters and the subject-specific binary responses is specified using a probit regression model and the latent indicator variables using the approaches proposed by George and McCulloch,[13] Kuo and Mallick,[17] and Brown et al.[18] We expound on each of these below.

Stage 1: relationship between aCGH emissions and latent copy number states. For subjects $i = 1, \ldots, n$ and probes $j = 1, \ldots, p$, we have the binary responses $y_1, \ldots, y_n$ representing the two disease subcategories and the set of real-valued aCGH emissions $\{e_{ij}\}$. Let $s_{ij} \in \{-1, 0, +1\}$ be a latent variable called the *copy number state*, representing a loss, no change, and gain in copy number for individual i at probe j. The copy number state is inferred using a Bayesian HMM that accounts for the serial correlations of the aCGH emissions.

Similarly to Guha et al.[23] conditional on s_{ij}, the aCGH emissions are assumed to be normally distributed:

$$e_{ij} \mid s_{ij} \overset{\text{indep}}{\sim} N(\mu_{s_{ij}}, \sigma^2_{s_{ij}}),$$

where, because of the specific biological interpretations associated with the HMM states, we assume that $\mu_{-1} < \mu_0 < \mu_{+1}$. This assumption also prevents label switching, a well-known problem with mixture models, thereby making inferences even more efficient. The latent states $s_{i1}, \ldots, s_{ip}$ are assumed to follow a three-state HMM with stationary transition probability matrix $\mathbf{A} = ((a_{ut}))_{3\times 3}$ having row sums $\sum_{t=1,2,3} a_{ut} = 1$ for $u = 1, 2, 3$. That is, $P[s_{i,j+1} = t \mid s_{ij} = u] = a_{ut}$ for $j = 1, \ldots, (n-1)$. To further facilitate inferences of the state-specific parameters, informative conjugate priors are assigned to the parameters of the normal distribution ie, μ_s and σ_s for $s \in \{-1, 0, +1\}$. Refer to Guha et al.[23] for further details about MCMC inference of the underlying copy number states of the probes for the individuals. The technique developed in that paper is applied to infer the latent copy number states (gain/loss/normal) $s_{i1}, \ldots, s_{ip}$ for subjects $i = 1, \ldots, n$ that are subsequently used in the below Stage 2.

Stage 2: relationship between disease classification and latent copy number states. In the second stage of the analysis, we model the relationship between the disease category and latent copy number states of the genomic probes for each individual. These values are copy number states inferred from analysis in Section 2.1.

Let $u^{(-)}_{ij} = I(s_{ij} = -1)$ and $u^{(+)}_{ij} = I(s_{ij} = +1)$ be indicator functions of loss and gain. To simplify the notation, for subjects $i = 1, \ldots, n$, we collectively represent the vector of $2p$ covariates as $w_{i\cdot} = (1, u^{(-)}_{i1}, u^{(+)}_{i1}, \ldots, u^{(-)}_{ip}, u^{(+)}_{ip})'$. For covariate $j = 1, \ldots, 2p$, averaging over the individuals, let $w_{\cdot j} = \sum^n_{i=1} w_{ij}/n$. Centering and scaling over the n individuals, we transform the covariates as follows:

$$v_{ij} = \begin{cases} (w_{ij} - w_{\cdot j})/\sqrt{\sum^n_{i=1}(w_{ij} - w_{\cdot j})^2} & \text{if } \sum^n_{i=1}(w_{ij} - w_{\cdot j})^2 > 0, \\ 0 & \text{otherwise} \end{cases}$$

Let Q be the set of covariates j for which $\{w_{ij}\}^n_{i=1}$ assumes at least two distinct values. That is, $Q = \{j \mid \sum^n_{i=1}(w_{ij} - w_{\cdot j})^2 > 0\}$. Because the variables v_{ij} are centered, $j \notin Q$ if and only if $v_{1j} = \ldots = v_{nj} = 0$.

A key assumption of our model is that probes that do not belong to Q ie for which $\{w_{ij}\}^n_{i=1}$ do not assume at least 2 distinct values, are not predictive of disease subcategory, although the probes could possibly be predictive of the disease. For this reason, we identify Q as the set of potential predictors of disease subcategory and write $q = |Q| \le 2p$. We discard all probes $j \notin Q$, relabeling the variables $\{v_{ij}: j \in Q\}$ as $\{x_{ij}: j = 1, \ldots, q\}$.

For individuals $i = 1, \ldots, n$, we assume the probit regression model proposed by Albert and Chib[25]:

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i \ge 0 \\ 0 & \text{if } z_i < 0 \end{cases} \\ z_i &= \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij} + \epsilon_i \\ \epsilon_i &\overset{indep}{\sim} N(0,1) \end{aligned} \qquad (1)$$

For the intercept β_0, we assume the prior $N(0, \tau_0^2)$. Let $\gamma = (\gamma_1, \ldots, \gamma_q)'$ be i.i.d. Bernoulli variables with $P[\gamma_j = \omega]$, where ω is expected to be relatively small and is assigned the uniform prior on (0,0.1). The remaining coefficients in (1) are independently distributed as

$$\beta_j \mid \gamma_j \overset{indep}{\sim} \begin{cases} \delta_0 & \text{if } \gamma{=}0 \\ N(0, \tau^2) & \text{if } \gamma{=}1 \end{cases}$$

where δ_0 denotes the point mass at 0. In other words, each probe is predictive of disease classification with probability ω. We assume independent exponential priors with mean 1 for τ_0^{-2} and τ^{-2}.

Gibbs Sampling Procedure

Let $\rho = 1 + \sum^q_{j=1} \gamma_j$ be the random number of variables (including the intercept β_0) that participate in the disease classification. Let $r_{ij} = z_i - \sum_{k \ne j} x_{ik}\beta_k$ for $i = 1, \ldots, n$. For a set of numbers $\{\theta_{ij}: i = 1, \ldots, n, j = 1, \ldots, q\}$, let $\boldsymbol{\theta}_j$ represent the vector $(\theta_{1j}, \ldots, \theta_{nj})'$ for probe $j = 1, \ldots, q$.

Although the Gibbs sampler is conceptually straightforward, updating of γ can be computationally intensive for large q. The step is described as follows. For probe $j = 1, \ldots, q$, let β_{-j} represent the set of regression coefficients excluding β_j. With $\boldsymbol{I}_n$ denoting the identity matrix of order n and $B_j = I_n + \tau^2 x_j x_j^T$, the posterior probability $P[\gamma_j \mid \beta_{-j}, \omega, \boldsymbol{r}_j]$ is proportional to $(1 - \omega) \cdot N_n(r_j \mid \mathbf{0}, \boldsymbol{I}_n)$ when $\gamma_j = 0$ and is proportional to $\omega N_n(r_j \mid \mathbf{0}, \boldsymbol{B}_j)$ when $\gamma_j = 1$. The density $N_n(r_j \mid \mathbf{0}, \boldsymbol{I}_n)$ can be quickly computed even in large problems. However, the density $N_n(r_j \mid \mathbf{0}, \boldsymbol{B}_j)$ involves the inversion and determinant calculation for the non-diagonal matrix $\boldsymbol{B}_j$. Because it must be

iteratively performed for every probe j, it can be computationally expensive or can at least involve large amounts of memory, when q is large. Theorem 7.1 of the Appendix exploits the structure of $\boldsymbol{B}_j$ to drastically simplify the computation. For probe $j = 1, \ldots, q$, let

$$\begin{aligned} \phi_j &= \left(\frac{1}{\sqrt{1^2+\tau^2}} - 1\right) x_j^T r_j, \\ h_j &= \phi_j x_j + r_j, \quad \text{and} \\ L_{j1} &= \sum_{i=1}^{n} h_{ij}^2. \end{aligned} \tag{2}$$

Applying Theorem 7.1, we have $det(\boldsymbol{B}_j) = 1 + \tau^2$, and $N_n(r_j|\boldsymbol{0},\boldsymbol{B}_j)$ is proportional to $\exp(-0.5 L_{j1})/\sqrt{1+\tau^2}$. The calculation is feasible even for large q.

Outline of procedure

Let $F\text{-}I(c, d)$ denote the distribution F restricted to the interval (c, d). The Gibbs sampler consists of the following steps:

- Applying Theorem 7.1, the binary indicators for probes $j = 1, \ldots, q$ are updated as follows:

$$P[\gamma_j \mid \beta_{-j}, \omega, z] \propto \begin{cases} (1-\omega)\cdot \exp(-0.5 L_{j0}) & \text{if } \gamma_j = 0 \\ \omega/\sqrt{1+\tau^2}\cdot \exp(-0.5 L_{j1}) & \text{if } \gamma_j = 1 \end{cases}$$

where $L_{j0} = \sum_{i=1}^{n} r_{ij}^2$ and L_{j1} is as defined in (2).

- Writing $x_i = (1, x_{i1}, \ldots, x_{iq})^T$ for individuals $i = 1, \ldots, n$, the subject-specific latent variables z are independently distributed as

$$z_i \mid \beta, y \sim \begin{cases} N(x_i^T\beta, 1)\cdot 1(-\infty, 0) & \text{if } y_i = 0 \\ N(x_i^T\beta, 1)\cdot 1(0, \infty) & \text{if } y_i = 1 \end{cases}$$

- Let β_I be the elements of β corresponding to the intercept and to the set of probes j for which $\gamma_j = 1$. Then $\beta_{I^c} = 0$. Vector β_I is jointly updated as

$$\beta_I \mid z, \gamma \sim N_\rho(\Sigma_I U_I^T z, \Sigma_I)$$

where U_I is an $n \times \rho$ matrix with the first column equal to a vector of n 1's and the remaining columns equal to the vectors x_j for which $\gamma_j = 1$. The variance matrix $\Sigma_I = (U_I^T U_I + \tau^{-2} I_\rho)^{-1}$.

- $\tau_0^{-2} \mid \beta_0$ is distributed as gamma $(3/2, (1+\beta_0^2)/2)$.
- $\tau^{-2} \mid \beta_{-0}$ is distributed as gamma $((1+\rho)/2, (1+\sum_{j=1}^{q}\beta_j^2)/2)$.
- $\omega \mid \gamma$ is distributed as beta $(\rho, q-\rho+1) \cdot 1\ (0, 0.1)$.

Test case predictions

Suppose we have the aCGH profiles of n^* additional *test case* individuals from the same hypothetical disease population. Using the within-variable means and variances of the training sample, we transformed the aCGH profiles to obtain the covariates $x_{i^*} = (1, x_{i^*1}, \ldots, x_{i^*q})^T$ for individuals $i^* = 1, \ldots, n^*$ belonging to the test sample. Let D represent the training set data. The posterior probability that individual i^* belongs to disease category 1 is

$$\begin{aligned} P[y_{i^*} = 1 \mid D] &= \int P[y_{i^*} = 1 \mid \beta][\beta \mid D] d\beta \\ &= 1 - \int \Phi[0 \mid x_{i^*}^T\beta, 1][\beta \mid D] d\beta. \end{aligned}$$

A consistent (in simulation size) estimate of this probability is then

$$\hat{P}[y_{i^*} = 1 \mid D] = 1 - \sum_{t=1}^{M} \Phi(0 \mid x_{i^*}^T \beta^{(t)}, 1)/M$$

where $\beta = \beta^{(t)}$ is the value generated at the Mth MCMC iterate. We declare the disease category of the test case individual labeled i^* as

$$\hat{y}_{i^*} = \begin{cases} 0 & \text{if } \hat{P}[y_{i^*} = 1 \mid D] < 0.5 \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

Simulation Study

We generated a training sample consisting of $p = 2000$ aCGH profiles for $n = 100$ individuals. The individuals were regarded as random draws from a disease population where $100 \times (1 - p^*) = 25\%$ of the individuals had "disease 0" and the remaining $100 \times p^* = 75\%$ individuals had "disease 1," so that $p^* = 0.75$ represented the prior probability of disease 1 in the population.

Disease 0 was assumed to be characterized by losses ($s = -1$) from probes 201 to 400 and gains ($s = 1$) from probes 1401 to 1800. Disease 1 was characterized by losses from probes 301 to 500 and also from probes 1601 to 1800. The remaining probes were assigned a copy number state of 0. For each disease subcategory, we randomly selected 10% of the probes that were associated with the disease and randomly set their copy number states to be copy neutral, gains, or losses with equal probability. Additionally, random noise at the probe level was then added to the profiles by selecting 2% (ie, 4000) of the remaining probes and randomly changing their copy number states. These values constituted the variables s_{ij} in Stage 2 of the Section 2 model, and were assumed to be known in the simulation.

As described in Section 2, the variables were then transformed to obtain the covariates w_{ij} and v_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, 2p$. The set $Q = \{j \mid \sum_{i=1}^{n}(w_{ij} - w_{\cdot j})^2 > 0\}$ was evaluated to identify $q = 2571$ probes for which the individuals had at least two distinct values. These variables were relabeled as $\{x_{ij}; j = 1, \ldots, q\}$, and the remaining variables were discarded. The model was fit using the Gibbs sampler of Section 3. An initial set of 10,000 samples was run to allow the MCMC chain to forget its starting values. A 1-in-10 subsample of $M = 100{,}000$ additional draws was stored for posterior inferences. Figure 1

presents histograms for the marginal posteriors of the intercept β_0, standard deviations τ_0 and τ, and Bernoulli probability ω, which are used in the sequel to make predictions for the disease categories of the test case individuals.

We evaluated the success of the predictive ability of our approach by drawing 50 independent test samples of $n^* = 200$ individuals from the same hypothetical disease population and generating their aCGH profiles based on their disease categories. Exactly 50 of these 200 test case individuals had disease 0, and the remaining 150 individuals had disease 1. Using the within-variable means and variances of each training sample, we transformed the aCGH profiles to obtain the covariates $x_{i^*} = (1, x_{i^*}1, \ldots, x_{i^*}q)^T$ for individuals $i^* = 1, \ldots, n^*$ belonging to the test sample of each of the 50 datasets.

For each dataset, using the stored MCMC sample of size $M = 100{,}000$ and as described in Section 3, we computed the posterior probability of disease 1, $\hat{P}[y_{i^*} = 1 \mid D]$, for the $n^* = 200$ individuals. The estimated $\hat{y}_{i^*}$ for the $n^* = 200$ individuals were computed as in (3). These values versus the true disease categories y_{i^*} are summarized in Table 1. The graph reveals the remarkable accuracy of the proposed methodology in detecting disease category. Specifically, for all 50 datasets, the technique resulted in perfect disease prediction with no false classification.

Breast Cancer Data Analysis

We analyzed the breast cancer dataset from Andre et al.[24] which consists of $n = 111$ individuals with either disease subcategory ER+ (label "1") or TN (label "0"). There are 56 TN and 55 ER+ patients. aCGH emissions for these individuals were available on the same set of $p = 42{,}416$ probes along with the probes' locations. Specifically, the chromosome and the distance in megabases (MB) from a telomere are available for every probe.

As described in Section 2.1, we used this information to first infer the latent copy number states e_{ij} of the probes using a Bayesian HMM, where $i = 1, \ldots, 111$ and $j = 1, \ldots, 42{,}416$. Then, as described in Section 2.2, we obtained the indicator functions, u_{ij}^+ and u_{ij}^-, of gain and loss. These indicator variables were transformed to obtain the covariates w_{ij} and v_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, 84{,}832$. The set $Q = \{j \mid \sum_{i=1}^{n}(w_{ij} - w_{\cdot j})^2 > 0\}$ was evaluated to identify $q = 5{,}543$ covariates having at least two distinct values for the 111 individuals. These variables were relabeled as $\{x_{ij}: j = 1, \ldots, 5{,}543\}$ and retained as potential regressors. The remaining variables were discarded because they were unlikely to be associated with the subcategory classification.

To investigate the reliability of the proposed method of these actual datasets, we performed 50 independent

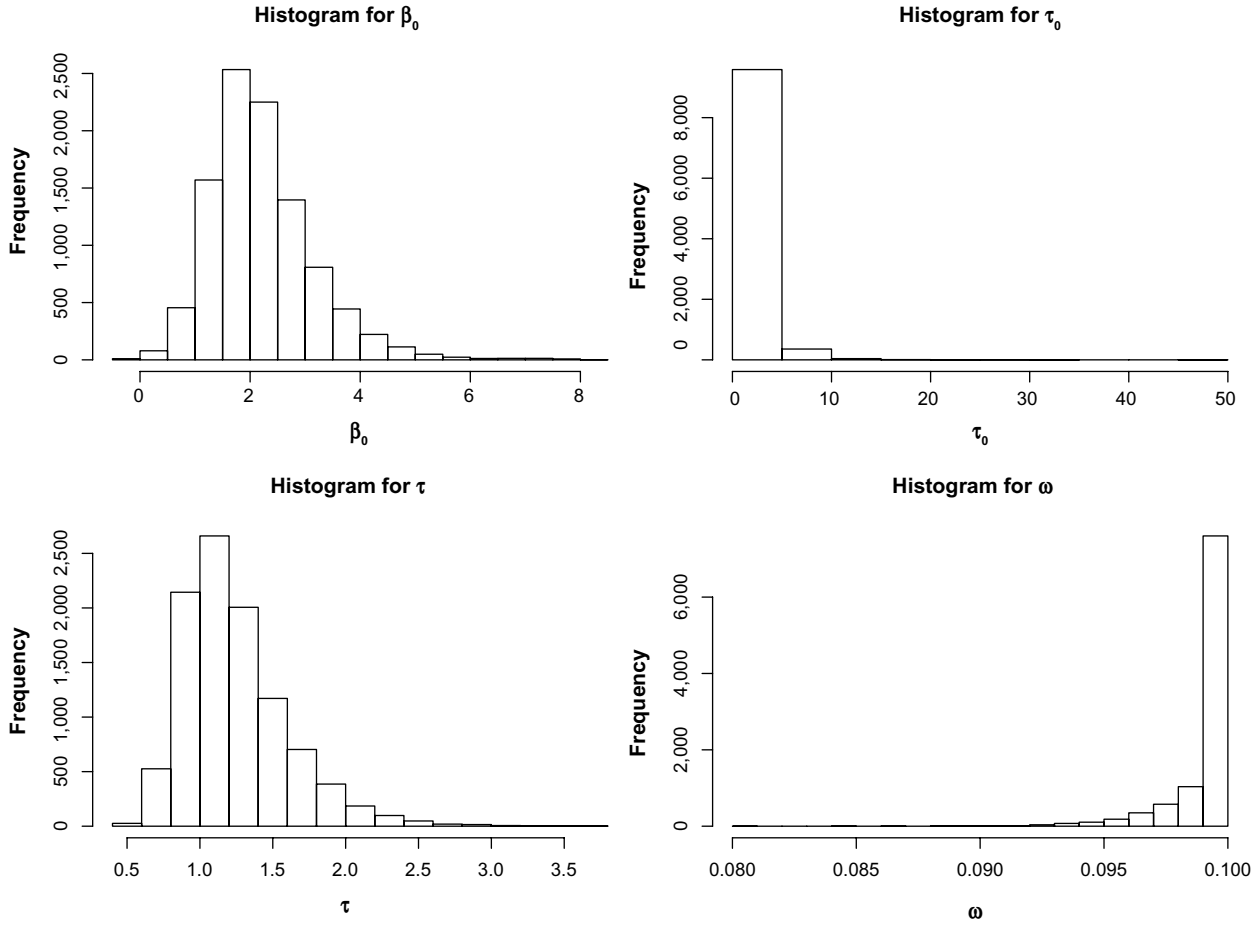

Figure 1. Histogram of selected model parameters for the simulation study.

Table 1. For the 200 individuals belonging to the 50 test samples of the simulation study, the estimated disease category versus the true category averaged over the 50 test samples. Perfect classification was obtained for each dataset. As a result, the standard errors shown in parenthesis are all zero.

	ESTIMATED	
	$\hat{y}_{i^*} = 0$	$\hat{y}_{i^*} = 1$
Truth		
$y_{i^*} = 0$	50 (0)	0 (0)
$y_{i^*} = 1$	0 (0)	150 (0)

replications of the following steps. (i) We randomly split the data into training and test sets in a 4:1 ratio. (ii) We analyzed the disease subcategories and the $q = 5{,}543$ covariates of the 89 training set individuals using the Bayesian probit regression model with likelihood function (1). The model was fit using the Gibbs sampler of Section 3. An initial set of 10,000 samples was run to allow the MCMC chain to overcome its initial values. A 1-in-10 subsample of $M = 100{,}000$ additional draws was stored for posterior inferences. (iii) As described in Section 3, we used the $q = 5{,}543$ covariates of the 22 test case individuals to predict their disease subcategories. These predictions were compared with the actual disease subcategories of these 22 individuals to compute the classification error rate for the specific training–test case random split. An average of the 50 independent estimates in Step (iii) yielded a simulation-based estimate of the classification error rate for the proposed method. This was estimated to be 22.55% with a standard error of 1.16%.

The significant probes (covariates) that were found to be predictive of disease subtype are plotted in Figures 2–4. We assumed a posterior probability threshold of $\delta = 0.15$ that yielded 500 markers along the entire genome predictive of the disease classification. Figure 2 plots a bar graph of the chromosomal breakdown of these markers. As can be seen, most of the significant markers are located on chromosomes 5, 12, 16, and 17. The corresponding karyograms Figures 3 and 4 show the breakdown on the markers by chromosomal locations for negative (red) and positive (green) associations with the disease states, respectively.

Our results are promising based on the locations of selected markers. As noted, most markers are on chromosomes 5, 12, 16, and 17. It has been shown that chromosome 5q deletions are the most frequent aberration in breast tumors from *BRCA1* mutation carriers. The deletions in 5q occur at high frequencies on putative tumor suppressor genes such as *XRCC4*, *RAD50*, *RASA1*, *APC*, and *PPP2R2B*.[26] Chromosome 16q has been a target region for the detection of biomarkers for breast cancer.[24] We identified a high concentration of biomarkers in 16q as well. In addition, our flagged biomarkers on chromosome 17 are also convincing, since chromosome 17 is the host for the most famous breast cancer gene *BRCA1* as well as *ER*. Interestingly, little is known about the association of CNVs on chromosome 12 with subgroups of breast cancer. Our findings on chromosome 12 could be potentially new discoveries that might warrant further functional validation.

Conclusions and Discussion

The detection of CNVs in aCGH methods is important for the treatment of many types of cancers, especially in the development of molecular-based personalized cancer therapies. We propose a framework for the prediction of disease types using aCGH profiles. We employ a Bayesian classification model and treat disease states as outcome and aCGH profiles as covariates in order to identify significant regions of the genome associated with disease subclasses. Specifically, we propose a principled two-stage method using the covariates exhibiting serial dependence. Stage 1 makes inferences on the underlying copy number states associated with the aCGH emissions based on HMM formulation. Using Bayesian linear variable selection procedures, Stage 2 detects the model parameters associated with the binary responses, conditional on the parameters of Stage 1.

The selected probes and their effects are parameters that are useful for predicting the disease categories of any additional individuals on the basis of their copy number profiles. A simulation study demonstrates the method's remarkable accuracy in detecting disease category. The methodology is applied to a breast cancer dataset, and we find several markers that are associated with disease subtype using the copy number profiles. Some of these discoveries confirm existing literature, and novel associations could be potential targets for future validation studies.

Our methods are general and could be potentially applied to SNP arrays as well that yield copy number profiles. A nice generalization of the method would be to incorporate genotype information (eg, allelic frequencies) in the models (especially, Stage 1) that could lead to more refined estimation of the latent copy number states. Furthermore, current technologies enable collection of multiplatform data on matched patient samples such as mRNA expression (eg, The Cancer Genome Atlas (TCGA)) that can be leveraged to provide a more detailed understanding of the biological mechanisms involved in cancer development and progression. We leave these tasks for future consideration.

Appendix

Theorem 7.1: Let $x = (x_1, \ldots, x_n)'$ be a vector such that $x^T x = 1$. Define the matrices $\boldsymbol{A} = xx^T$ and $\boldsymbol{B} = \boldsymbol{I}_n + \tau^2 \boldsymbol{A}$. Then the determinant of matrix $\boldsymbol{B}$ is $1 + \tau^2$. Given $r \in \mathrm{R}^n$, define the vector $h = (h_1, \ldots, h_n)^T = \phi\, x + r$ and scalar $\phi = (1/\sqrt{1+\tau^2} - 1) x^T r$. Let $L = \sum_{i=1}^{n} h_i^2$. Then the n-variate normal density

$$N_n(r \mid 0, B) = \frac{1}{(2\pi)^{n/2}\sqrt{1+\tau^2}} \exp\left(-\frac{1}{2}L\right).$$

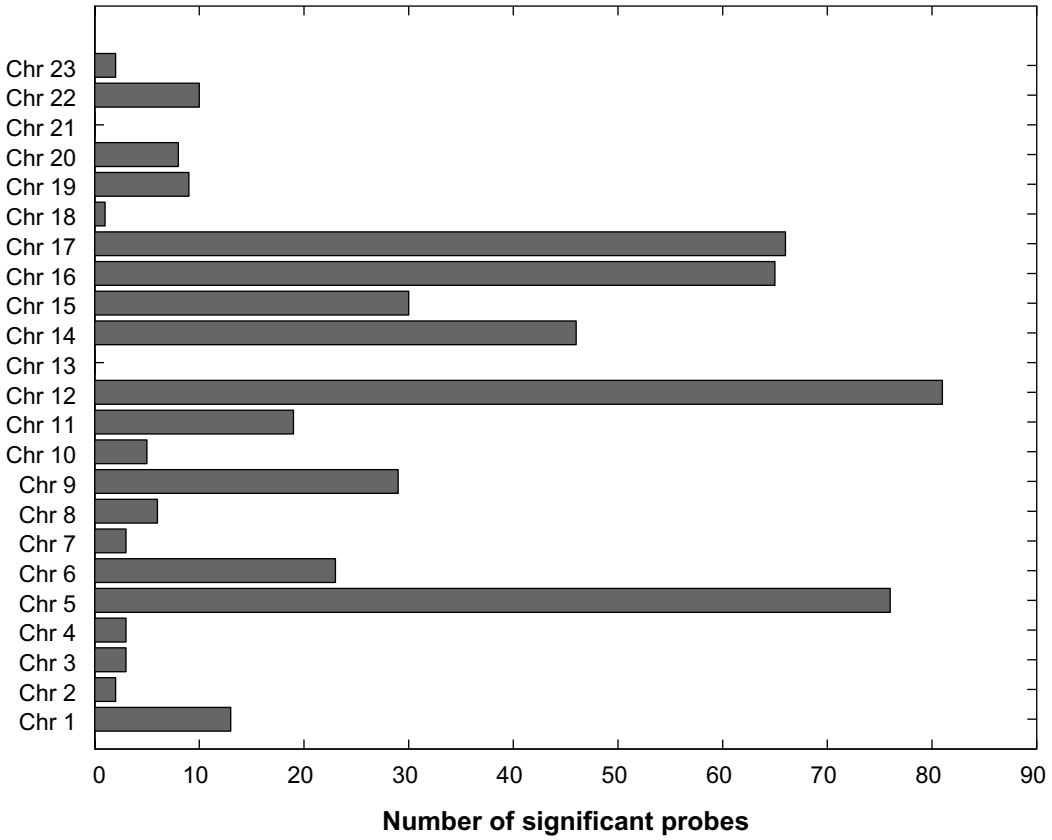

Figure 2. Number of significant markers broken down for each chromosome.

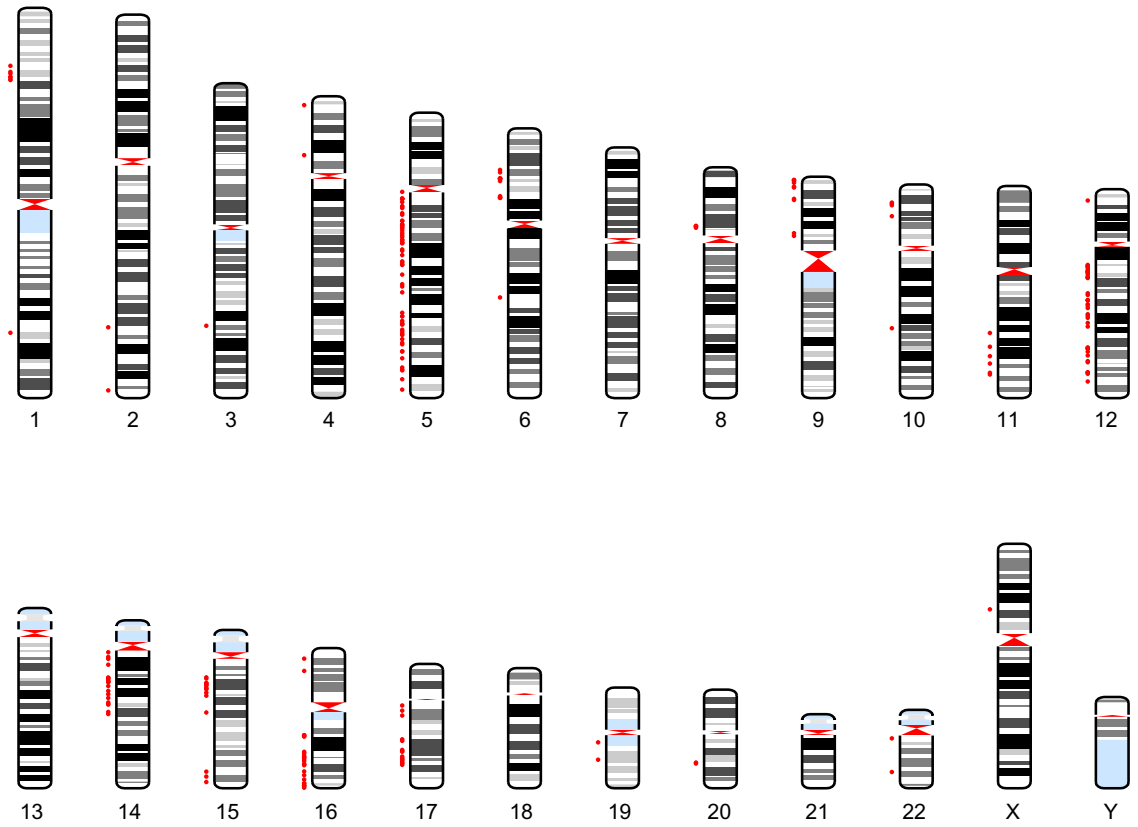

Figure 3. Human karyogram with significant locations. This figure is a karyogram that depicts the significant probes identified using our approach. The red color corresponds to negative regression coefficients.

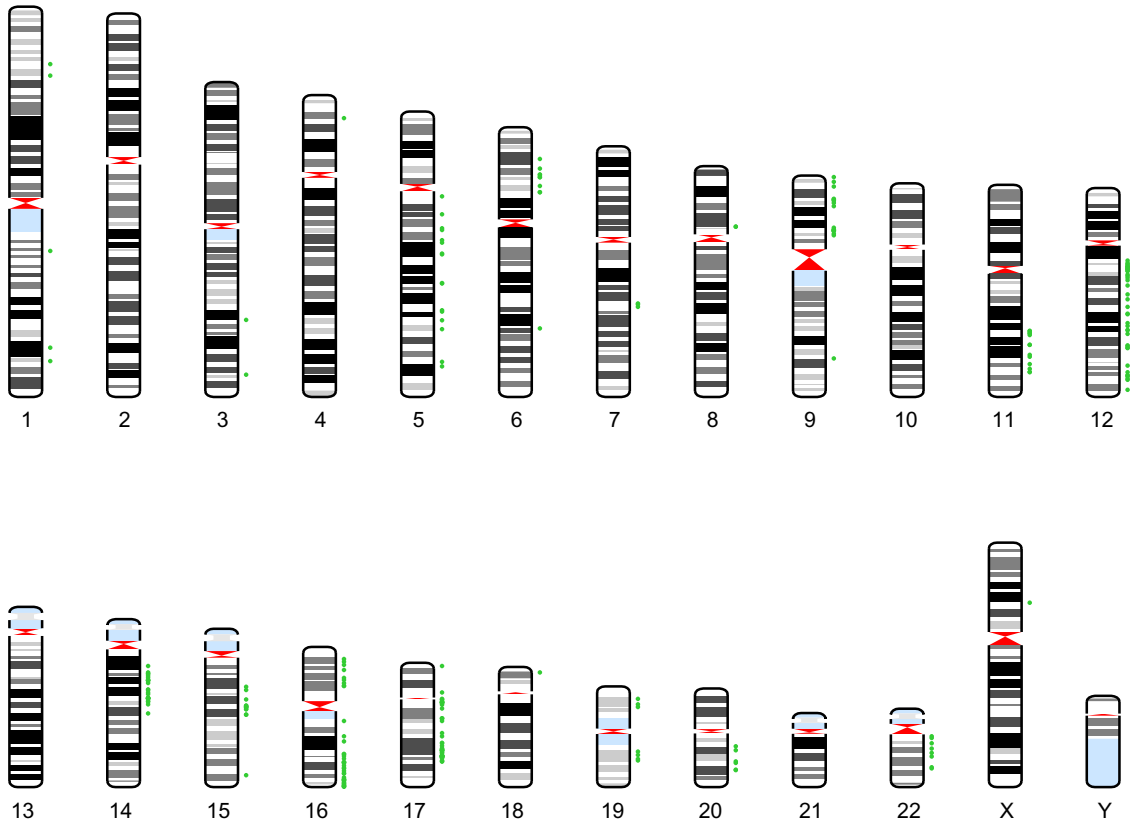

Figure 4. Human karyogram with significant locations. This figure is a karyogram that depicts the significant probes identified using our approach. The green color corresponds to positive regression coefficients.

Proof. Since $\boldsymbol{A} = \boldsymbol{x}\boldsymbol{x}^T$ has rank 1 and $\boldsymbol{x}^T\boldsymbol{x} = 1$, the eigenvalues of $\boldsymbol{A}$ consist of a single 1 and $(n - 1)$ number of 0's. Furthermore, the eigenvector corresponding to eigenvalue 1 must be $\boldsymbol{x}$. Let $\boldsymbol{\Lambda}_A$ be the diagonal matrix of the eigenvalues, and $\boldsymbol{P}$ be the matrix of eigenvectors of $\boldsymbol{A}$. Then $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}_A\boldsymbol{P}^T$.

Since $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}_n$ and $\boldsymbol{B} = \boldsymbol{I}_n + \tau^2\boldsymbol{A}$, $\boldsymbol{B}$ has the same eigenvectors as $\boldsymbol{A}$ and its eigenvalues are $1 + \tau^2$ and $(n - 1)$ number of 1's. The product of these eigenvalues is

$$\det(B) = 1 + \tau^2. \tag{4}$$

Matrix $\boldsymbol{B}^{-1/2}$ has the same eigenvectors as $\boldsymbol{B}$ and its eigenvalues are $1/\sqrt{1+\tau^2}$ and $(n - 1)$ number of 1's. Thus, $\boldsymbol{\Lambda}_{B^{-1/2}} = (1/\sqrt{1+\tau^2} - 1)\boldsymbol{\Lambda}_A + \boldsymbol{I}_n$ and

$$\begin{aligned} B^{-1/2} &= P\left(\left(\frac{1}{\sqrt{1+\tau^2}} - 1\right)\boldsymbol{\Lambda}_A + \boldsymbol{I}_n\right)\boldsymbol{P}^T \\ &= \left(\frac{1}{\sqrt{1+\tau^2}} - 1\right)P\boldsymbol{\Lambda}_A\boldsymbol{P}^T + \boldsymbol{P}\boldsymbol{P}^T \\ &= \left(\frac{1}{\sqrt{1+\tau^2}} - 1\right)\boldsymbol{A} + \boldsymbol{I}_n \end{aligned}$$

Given $r \in \mathrm{R}^n$, we have

$$\boldsymbol{B}^{-1/2}r = \left(\frac{1}{\sqrt{1+\tau^2} - 1}\right)\boldsymbol{A}r + r. \tag{5}$$

We obtain the result on substituting (4) and (5) in the n-variate normal density.

Author Contributions

Conceived and designed the experiments: SG, YJ, VB. Analyzed the data: SG, VB. Wrote the first draft of the manuscript: SG, YJ, VB. Contributed to the writing of the manuscript: SG, YJ, VB. Agree with manuscript results and conclusions: SG, YJ, VB. Jointly developed the structure and arguments for the paper: SG, YJ, VB. Made critical revisions and approved final version: SG, YJ, VB. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. *Nat Genet*. 2005;37(suppl):S11–7.
2. Chin L, Hahn WC, Getz G, Meyerson M. Making sense of cancer genomic data. *Genes Dev*. 2011;25:534–55.
3. Vissers LE, de Vries BB, Veltman JA. Genomic microarrays in mental retardation: from copy number variation to gene, from research to diagnosis. *J Med Genet*. 2010;47:289–97.
4. Kallioniemi A, Kallioniemi OP, Sudar D, et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*. 1992;258:818–21.

5. Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. *Biostatistics*. 2004;5:427–43.
6. Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*. 2005;21:4084–91.
7. Liu J, Ranka S, Kahveci T. Classification and feature selection algorithms for multi-class CGH data. *Bioinformatics*. 2008;24:86–95.
8. Rapaport F, Barillot E, Vert JP. Classification of arrayCGH data using fused SVM. *Bioinformatics*. 2008;24:i375–82.
9. Peduzzi PN, Hardy RJ, Holford TR. A stepwise variable selection procedure for nonlinear regression models. *Biometrics*. 1980;36:511–6.
10. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997;16:385–95.
11. Fan J, Li R. Variable selection for Coxs proportional hazards model and frailty model. *Ann Stat*. 2002;30:74–99.
12. Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. *J Am Stat Assoc*. 1988;83:1023–36.
13. George E, McCulloch R. Variable selection via Gibbs sampling. *J Am Stat Assoc*. 1993;88:881–9.
14. Dellaportas P, Forster JJ, Ntzoufras I. *Bayesian Variable Selection using the Gibbs Sampling*. New York: Marcel Dekker, Inc; 1982:273–86.
15. Madigan D, Raftery A. Model selection and accounting for model uncertainty in graphical models using Occams window. *J Am Stat Assoc*. 1994;89:1535–46.
16. Volinsky C, Madigan D, Raftery AE, Kronmal RA. Bayesian model averaging in proportional hazard models: assessing the risk of stroke. *Appl Stat*. 1997;46:433–48.
17. Kuo L, Mallick B. Bayesian semiparametric inference for the accelerated failure time model. *Can J Stat*. 1997;25:457–72.
18. Brown PJ, Vannucci M, Fearn T. Multivariate Bayesian variable selection and prediction. *J R Stat Soc Series B Stat Methodol*. 1998;60:627–41.
19. Cai B, Dunson D. Bayesian covariance selection in generalized linear mixed models. *Biometrics*. 2006;62:446–57.
20. Sha N, Vannucci M, Tadesse MG, et al. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*. 2004;60:812–19.
21. Lee K, Mallick B. Bayesian methods for variable selection in survival models with application to DNA microarray data. *Sankhya*. 2004;66:756–78.
22. Sha N, Tadesse MG, Vannucci M. Bayesian variable selection for the analysis of microarray data with censored outcome. *Bioinformatics*. 2006;22:2262–8.
23. Guha S, Li Y, Neuberg D. Bayesian Hidden Markov Modeling of Array CGH Data. *J Am Stat Assoc*. 2008;103:485–97.
24. Andre F, Job B, Dessen P, et al. Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. *Clin Cancer Res*. 2009;15:441–51.
25. Albert JH, Chib S (1993). Bayesian analysis of binary and polychotomous response data. *J Am Stat Assoc*. 1993;88:669–79.
26. Johannsdottir H, Jonsson G, Johannesdottir G, et al. Chromosome 5 imbalance mapping in breast tumors from BRCA1 and BRCA2 mutation carriers and sporadic breast tumors. *Int J Cancer*. 2006;119:1052–60.