Bioinformatics, 33(22), 2017, 3558–3566
doi: 10.1093/bioinformatics/btx464
Advance Access Publication Date: 26 July 2017
Original Paper

OXFORD

Gene expression

Towards clinically more relevant dissection of patient heterogeneity via survival-based Bayesian clustering

Ashar Ahmad[1,*] and Holger Fröhlich[1,2]

[1]Bonn Aachen International Center for Information Technology, University of Bonn, 53127 Bonn, Germany and [2]UCB Biosciences GmbH, 40789 Monheim, Germany

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on July 20, 2016; revised on June 5, 2017; editorial decision on July 13, 2017; accepted on July 24, 2017

Abstract

Motivation: Discovery of clinically relevant disease sub-types is of prime importance in personalized medicine. Disease sub-type identification has in the past often been explored in an unsupervised machine learning paradigm which involves clustering of patients based on available-omics data, such as gene expression. A follow-up analysis involves determining the clinical relevance of the molecular sub-types such as that reflected by comparing their disease progressions. The above methodology, however, fails to guarantee the separability of the sub-types based on their subtype-specific survival curves.

Results: We propose a new algorithm, Survival-based Bayesian Clustering (SBC) which simultaneously clusters heterogeneous-omics and clinical end point data (time to event) in order to discover clinically relevant disease subtypes. For this purpose we formulate a novel Hierarchical Bayesian Graphical Model which combines a Dirichlet Process Gaussian Mixture Model with an Accelerated Failure Time model. In this way we make sure that patients are grouped in the same cluster only when they show similar characteristics with respect to molecular features across data types (e.g. gene expression, mi-RNA) as well as survival times. We extensively test our model in simulation studies and apply it to cancer patient data from the Breast Cancer dataset and The Cancer Genome Atlas repository. Notably, our method is not only able to find clinically relevant sub-groups, but is also able to predict cluster membership and survival on test data in a better way than other competing methods.

Availability and implementation: Our R-code can be accessed as https://github.com/ashar799/SBC.

Contact: ashar@bit.uni-bonn.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A key idea of personalized medicine is an individually optimized patient treatment. This idea typically implies a stratification of patients into sufficiently homogeneous sub-populations. In that context characterization of disease sub-types is of high relevance. Some early land-mark studies based on gene expression dataset (Alizadeh *et al.*, 2000; Beer *et al.*, 2002; Lapointe *et al.*, 2004; Van't Veer *et al.*, 2002) have piqued interest in this direction. Disease subtype identification with an emphasis on patient survival prediction can be approached using either the molecular-omics data alone (fully unsupervised) or based entirely on patient survival data (fully supervised) by dichotomizing the patients into predefined groups like 'low-' and 'high-risk' and then using standard discriminative analysis tools like support vector machines, linear discriminant analysis or multinomial regression (Shipp *et al.*, 2002) to make predictions. The success of the supervised approach thus critically depends on

the a priori definition of patient sub-groups and the correspondence of these groups to molecular data.

Unsupervised clustering [such as hierarchical clustering (HS)—Yeung *et al.*, 2001] on the other hand focuses on discovery of molecular separable disease sub-types without any clinically motivated a priori definition of patient sub-populations. Once the disease sub-types are established a post-hoc analysis explores differences of the sub-types with respect to the clinical outcomes. However, the aforementioned method could discover subtypes which may not be related to survival or other clinical outcomes, as is evident in Verhaak *et al.* (2010). This concern was first highlighted by Bair and Tibshirani (2004) and later by Koestler *et al.* (2010).

To address the shortcomings of traditional supervised and unsupervised approaches Raman *et al.* (2010) propose a Bayesian Infinite Mixture of Experts Model to cluster patients with respect to their survival outcomes. Their model, in addition to determining main effects of the genes, also gives an insight to their higher order interactions in different clusters. However, this is achieved at the cost of discretizing continuous variables which—of course—leads to loss of information of continuous molecular data. Furthermore, their approach may suffer from non-interpretability when patient groups with different survival outcomes have near identical molecular profiles, thus failing to provide biological explanations for survival. Bair and Tibshirani (2004) propose a semi-supervised clustering which combines both gene expression data and clinical end-point data. They first identify a set of genes that are significantly correlated with survival time (using univariate Cox regression), then subsequently apply an unsupervised clustering technique (Nearest Shrunken Centroids) with the obtained set of genes. Risk predictions are made by using the principal component scores of the above mentioned set of genes. Although successfully used in many applications, Bair and Tibshirani (2004)'0 approach also has some limitations. For example, the algorithm requires to pre-specify the number of disease subtypes, which can be difficult in practice. Furthermore, the principal components of a set of genes to predict continuous risk scores can be difficult to interpret. Finally, uni-variate gene selection can fail, if multiple genes have a joint significant effect on survival, but marginal effects are weak. Koestler *et al.* (2010) as a further development to this approach propose a Recursive Partition Mixture Model which successively fits models with varying number of clusters (K) and uses modified Bayesian Information Criterion to efficiently estimate K. Moreover the model also selects the optimal gene set M. The key idea for feature and model selection is to train the model on top-ranking genes and to check the separability of the survival curves on an independent test set. At the end that gene set M is selected which gives the lowest possible *P*-value. Although computationally attractive, the results of feature selection and cluster number determination are heavily dependent on how the whole dataset is split into training and testing.

In our work we try to overcome several of the above mentioned limitations of present techniques. More specifically our proposed 'Survival-based Bayesian Clustering (SBC)' approach has the following features:

- automated and fully Bayesian treatment of the number of clusters
- ranking of most discriminatory features
- integration of different-omics data types, as exemplified here via miRNA plus gene expression data.
- prediction of class membership and survival outcomes for patients on an independent test data.

2 Proposed approach

Our 'SBC' approach rests on the foundations of Bayesian model-based clustering and sparse bayesian survival curve estimation. We first motivate the use of these two methodologies: Although a particular appealing property of model-based clustering (here using Gaussian mixtures) is to naturally deal with uncertainty regarding cluster assignment of patients, the Bayesian framework, in addition, allows for an elegant way to circumvent the model selection problem, i.e. to decide for a particular number of clusters. More specifically, we, in our present work, build on previous work in the machine learning community on infinite Gaussian Mixture Model (GMM) Rasmussen (2000). Infinite GMMs Neal (2000) are based on a Dirichlet Process (DP) prior over parameters. The DP priors define a probabilistic model for data generation and for cluster assignments. The most important characteristic of infinite GMM is Bayesian Model Averaging which allows us to estimate the posterior distribution over the number of clusters, thus avoiding the need to compare separately fitted GMMs.

In the past several attempts have been made to use DP models for clustering gene expression profiles by Medvedovic and Sivaganesan (2002), Medvedovic *et al.* (2004) and more recently by Yuan *et al.* (2011). Medvedovic *et al.* (2004) compare the performance of the infinite model to the finite model case in a simulation setting and find it advantageous to use infinite model especially in the case of high noise. Motivated by these findings and its inherent flexibility we choose Dirichlet Process Gaussian Mixture Models (DPMM) for modelling the expression profiles in our work.

One of the key innovations of this work is the additional inclusion of cluster-specific survival models in the DPMM. We use Accelerated Failure Time (AFT) model with the log-normal assumption Royston (2001) to model the survival or progression free survival times of the patients. The choice of the AFT model as opposed to the Cox Proportional Hazards model was made due to the ease of casting the AFT model in a Bayesian setting. We model the AFT as a Bayesian LASSO Park and Casella (2008) to identify potential biomarkers which are related to survival times.

Apart from cluster-specific sparse survival models two further key innovations in our approach are:

- Data Integration: We extend our mixture model to more than one data source (e.g. gene expression plus miRNA expression). As opposed to existing work our approach thus combines multi-omics and survival information to cluster patients.
- Prediction: In contrast to unsupervised clustering methods, our model can be used to make survival as well as class predictions (sub-type) of new patients. In contrast to supervised methods we do not need to know patient sub-types in advance.

3 Method details

3.1 DP mixture model

DP Mixture models belong to the broad category of Bayesian Non-Parametric methods. They allow for the inference of countable (possibly infinite) number of mixture components *K*. DP were first introduced by Antoniak (1974) and Ferguson (1973). If we assume $\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_N$ to be *N* data points drawn independently from some unknown distribution, where $\mathbf{X}_i$ can be multivariate or categorical, then the DP Prior models the density of $\mathbf{X}_i$ in the following hierarchical fashion:

$$\mathbf{X}_i|\theta_i \sim F(\theta_i)$$

$$\theta_i|G \sim G$$

$$G \sim DP(G_0, \alpha)$$

F is the conditional distribution of $\mathbf{X}_i$ which is parametrized by θ_i. G is the posterior mixture distribution which is mostly marginalized when inferring DP mixture model. G_0 is the base distribution and represents prior information about the parameter values. The parameter α is known as the concentration parameter and controls the number of clusters that we obtain from the posterior distribution. The marginalized prior representation was obtained by Blackwell and MacQueen (1973) by representing it as series of conditional distributions:

$$\theta_i|\theta_1, \ldots \theta_{i-1} \sim \frac{1}{i-1+\alpha}\sum_{j=1}^{i-1}\delta(\theta_j) + \frac{\alpha}{i-1+\alpha}G_0$$

Further details can be found in the Supplementary Material.

3.2 Hierarchical multivariate Gaussian model

As a choice for the base distribution G_0, we use a hierarchical Gaussian Model. Our Hierarchical Multivariate Gaussian mixture model (referred to in this work as DPMM) follows closely the work of Görür and Rasmussen (2010). The conjugate GMM can be described with the following sets of equations:

$$\mathbf{X}_i|(c_i = j) \sim \mathcal{N}\left(\mu_j, S_j^{-1}\right)$$

$$\left(\mu_j \,|\, S_j, \xi, \rho\right) \sim \mathcal{N}\left(\xi, (\rho S_j)^{-1}\right)$$

where $\mathbf{X}_i$ indicates a D-dimensional vector of measurements (e.g. gene expression profiles) for patient i. Furthermore, μ_j is the centre of cluster (or sub-type) j, described via a multivariate Gaussian with precision matrix S_j. The second equation constitutes a prior distribution for μ_j which itself is a normal distribution with expectation ξ and scaled precision matrix ρS_j.

We regularize the precision matrix towards a diagonal matrix W as in Bouriga and Féron (2013):

$$(S_j|\phi, W) \sim \mathcal{W}\left(\phi, (\phi W)^{-1}\right)$$

where $\mathcal{W}$ denotes a Wishart distribution with ϕ degrees of freedom. Empirical Bayes estimates as described in Görür and Rasmussen (2010) are used as priors over the hyper-parameters (ξ, W etc.). For the complete hierarchy we refer to the Supplementary Material.

3.3 Bayesian LASSO penalized AFT model

The AFT model has been extensively used to model survival times of cancer patients, modelling either the survival probabilities or time to recurrence probabilities (Wei, 1992). In its most general form an AFT model is given by:

$$\log(t_i) = \beta_0 + \beta^T\mathbf{X}_i + \epsilon_i, i = 1, \ldots N$$

where $\log(t_i)$ is the log survival time (or progression free survival), and β is the vector of regression parameters. As most likely only a subset of features is truly associated to survival, we place a Laplacian prior over β which effectively induces a L1 penalty on the regression coefficients and penalizes small effects to exact zero. Following a Bayesian approach we place a a diffuse gamma-prior on the penalty strength parameter, λ, and evaluate its posterior.The hierarchical formulation of the Bayesian Lasso Park and Casella (2008) is:

$$\log(t_i)|\beta_0, \mathbf{X}_i, \beta, \sigma^2 \sim \mathcal{N}\left(\beta_0 + \beta^T\mathbf{X}_i, \sigma^2\right),$$

$$\beta|\sigma^2, \tau_1^2, \tau_2^2 \ldots, \tau_p^2 \sim \mathcal{N}\left(0_p, \sigma^2\mathbf{D}_\tau\right),$$

$$\mathbf{D}_\tau = \mathrm{diag}\left(\tau_1^2, \ldots \tau_p^2\right)$$

Our treatment of the censored patients follows closely to that of Sha *et al.*, 2006). The key idea is to treat the censored outcomes $I_i = 0$ as yet another unknown parameter w_i and to use our probabilistic model to estimate the censored survival times. We augmented the survival times with pseudo variables w_i which are defined as follows:

$$w_i = \log(t_i) \quad \text{if} \quad \text{censoring} = \text{FALSE}$$

$$w_i > \log(t_i) \quad \text{if} \quad \text{censoring} = \text{TRUE}$$

For the case of censoring, w_i is assumed to be drawn from a left truncated normal distribution, with the left truncation at the censored survival time (Sha *et al.*, 2006).

3.4 Bayesian regularization

Our proposed **SBC** model is fully Bayesian. Within this framework, model complexity is penalized with the help of prior distributions at several places:

- We use a DP prior to favour few clusters.
- We incorporate a prior for the covariance matrix of each cluster favouring sparse diagonal matrices.
- We use a Bayesian lasso to favour sparse cluster-specific survival regression models.

3.5 Model fitting via Gibbs sampling

SBC can be depicted as a graphical model, as shown in Figure 1. The hierarchical model formulation and the use of conditionally conjugate model enables the use of a Gibbs sampling based algorithm for parameter estimation. The cluster indicator variables c_is are updated using the following conditional distribution for those components which have non-zero elements i.e. $n_{-i,j} > 0$:

$$p\left(c_i = j \,|\, \mathbf{c}_{-i}, \mu_j, S_j^{-1}, \beta_{0j}, \beta_j, \sigma_j^2, \alpha\right)$$

$$\propto \frac{n_{-i,j}}{N-1+\alpha}\mathcal{N}\left(w_i \,|\, \beta_{0j} + \beta_j^T\mathbf{X}_i, \sigma_j^2\right)\mathcal{N}\left(\mathbf{X}_i \,|\, \mu_j, S_j^{-1}\right)$$

for all others combined we can sample from the conditional distribution as detailed in the Supplementary Material. As the assignment of a new cluster involves marginalization over mixture model parameters, this integral turns out to be non-tractable in our case. In order to circumvent this problem, we use the auxiliary variable method used in Algorithm 8 of Neal (2000) with the number of auxiliary variables set to two as described in Görür and Rasmussen (2010).

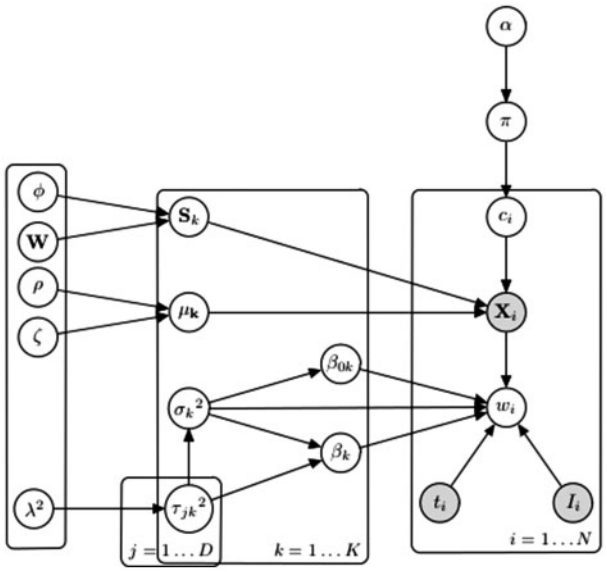

Fig. 1. Graphical Model representation for **SBC**. *D* refers to the data dimension, *N* to number of data points and *K* represents number of clusters

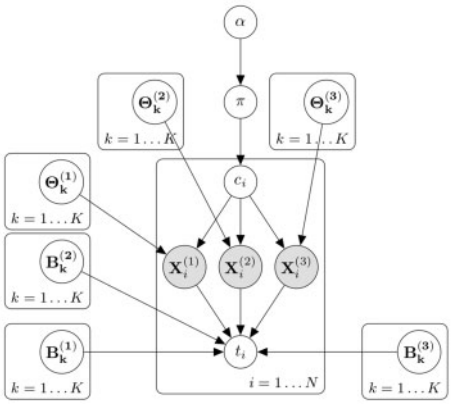

Fig. 2. Graphical Model representation for **iSBC** with $Q = 3$ data sources

3.6 Feature importance

The hierarchical formulation of the SBC allows us to define a ranking over the discriminatory ability of each feature with respect to two clusters (a and b). We define the ranking r^i of a feature i as in (Yau and Holmes, 2011):

$$r^i = \frac{\mu_a^i - \mu_b^i}{\omega^i}$$

where μ_k^i is the ith component of the mean vector μ for cluster k and ω^i is the ith diagonal element of W. When we have more than one cluster, we calculate feature importance with respect to every pair of clusters.

3.7 Data integration

Our present model can be extended to integrate more than one-omics data source. In this work, we use data sources which have continuous (Gaussian) values. These are then all modelled as described in Section 3.2. To combine several data sources ($v = 1 \dots Q$) we compare two different strategies: a) one in which we work with independently pre-filtered feature sets from each of the data source $\mathbf{X}_i^{(v)}$ and b) one in which we perform a Canonical Correlation Analysis (CCA) on the original (pre-filtered) features and then map data from each data source on the top canonical covariates ($\mathbf{X}I_i^{(v)}$).

CCA is a classical data integration method (Hotelling, 1936) which is used to extract concordant feature sets. Each of the canonical covariates is constructed to successively explain maximal correlation between linear feature combinations from two or more data sources. After we obtain the feature sets from the above mentioned two methods we assume that the complete model likelihood of feature sets of a patient given its cluster membership can be factorized as:

$$p\left(\mathbf{X}_i^{(1)}, \mathbf{X}_i^{(2)}, \dots \mathbf{X}_i^{(Q)} | c_i = j, \mu_j^{(1)}, S_j^{(1)}, \dots \mu_j^{(Q)}, S_j^{(Q)}\right)$$

$$= \prod_{v=1}^{Q} \mathcal{N}\left(\left(\mathbf{X}_i^{(v)} | \mu_j^{(v)}, S_j^{-1(v)}\right)\right)$$

where $\mathbf{X}_i^{(v)}$ denotes features of the ith patient from the vth data source with features which come either from a) or b). We further suppose a factorization of the AFT mode across data sources as:

$$\log(t_i) | \left[\mathbf{X}_i^{(1)}, \dots \mathbf{X}_i^{(Q)}, c_i = j, \beta_{0j}^{(1)}, \beta_j^{(1)}, \sigma_j^{2(1)}, \dots \beta_{0j}^{(Q)}, \beta_j^{(Q)}, \sigma_j^{2(Q)}\right]$$

$$\sim \prod_{v=1}^{Q} \mathcal{N}\left(\beta_{0j}^{(v)} + \beta_j^{T(v)} \mathbf{X}_i^{(v)}, \sigma_j^{2(v)}\right)$$

This essentially means that each data source has its own cluster-specific AFT model as described Section 3.3 and a weight that depends on the likelihood of observing the clinical endpoint with features from that data source. The cluster indicator of a patient sample c_i, as in the one data source case, is given a DP Prior. We call this approach as **integrative SBC or iSBC**. As an illustration, we have shown in Figure 2 the iSBC model with $Q = 3$. To simplify notation, all parameters of the Hierarchical Multivariate Gaussian are represented by $\mathbf{\Theta}_\mathbf{k}^{(v)}$ and those of the Bayesian LASSO penalized AFT model are denoted by $\mathbf{B}_\mathbf{k}^{(v)}$.

3.8 Making model predictions

Given a already trained SBC model with parameters $\left[\theta_{1:N}^{(m)}, c_{1:N}^{(m)}\right]$ over M MCMC samples and molecular data X^* for test patient, we would like to solve two predictions problems a) survival prediction and b) prediction of cluster membership. For the sake of simpler notation, X^* is assumed to be of one specific-omics type, but the same approach also works for multi-omics data.

3.8.1 Survival prediction

Expectation of the survival time for a new patient according to the SBC is a weighted average over predicted survival times from each cluster:

$$\mathbb{E}\left[\log(t^*) | X^*, \theta_{1:N}^{(m)}, c_{1:N}^{(m)}\right] \approx \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{C_m} \left(\beta_{0jm} + \beta_{jm}^T X^*\right) * v_{jm}(X^*)$$

where C_m denotes the number of clusters found in MCMC sample m. Notably, each MCMC sample corresponds to one full parameter set of our model. Hence β_{0jm}, β_{jm} denote the regression parameters in our AFT model for MCMC sample m. Moreover $v_{jm}(X^*)$ is the weight that is dependent on the likelihood of X^* to belong to cluster j for the MCMC sample m. More details about the exact form of the weights $v_{jm}(X^*)$ can be found in the Supplementary Material.

3.8.2 Cluster membership

The new data point X^* is assigned a probability of belonging to the already discovered clusters for the mth MCMC sample, $c^{(m)} = 1 \ldots C_m$, by using the following conditional probability of the DP model:

$$p\left(c^* = j \mid X^*, \theta^{(m)}_{1:N}, c^{(m)}_{1:N}\right) = b\frac{n_{jm}}{N-1+\alpha}\mathcal{N}\left(X^* \mid \mu_{jm}, S^{-1}_{jm}\right)$$

where n_{jm} is the number of patients in the cluster j and μ_{jm}, S_{jm} are the corresponding cluster parameters of the Hierarchical Multivariate Gaussian model for the mth MCMC sample, b is a normalization constant.

4 Simulation study

We investigated the performance of our SBC model in various simulation settings. We simulated molecular data as multivariate Gaussians with non-trivial correlation structure and varying degrees of overlap using the *MixSim* R-package. The package provides a list of D dimensional cluster-specific mean vectors and $D \times D$ cluster-specific full covariance matrices. Since SBC—as well as competing methods—are typically applied on a pre-filtered subset of all features ($10 \leq D \leq 60$), we also investigated the robustness of our model when a certain fraction of noise features (20 and 50% of all features) were added, which did not contribute to the clustering structure. Finally, the molecular data were obtained by concatenating the relevant and noise features. For the survival data generation we used cluster-specific log-normal AFT models applied on the molecular data and then added cluster-specific Gaussian noise.

We simulated $n = 100$ data points, each for training the model and for testing it. We repeated the whole simulation process 10 times and compared our results with other competing approaches. Further details on the simulation settings can be found in the Supplementary Material.

For comparison of our SBC model with other competing models we used FLXmix Grün and Leisch (2007) k-means (kM), Mixture of Factor Analyzers McLachlan and Peel (2000), PReMiuM Liverani *et al.* (2015), sparse kM and sparse HC Witten and Tibshirani (2010). The two measures used to compare our results were the C-index Harrell *et al.* (1982) and Adjusted Rand Index. Rand Index measures the agreement between two clusterings, it ranges from 0 (no agreement) to 1 (full agreement), the adjusted Rand Index also corrects for chance groupings and can have negative values (indicating worse than chance agreement). The C-Index (or Concordance Index) is used to assess prediction performance in survival analysis and is akin to Area-Under-Curve in the classification case. To compare purely unsupervised clustering methods, such as kM against our SBC approach with respect to survival predictions on training data we applied a two-step strategy: first clustering and then fitting cluster-specific survival curves using either a lasso penalized AFT or Cox model. We call the corresponding algorithms as K-PCOX (kM clustering followed by cluster-specific penalized Cox regression), N-PCOX (Penalized Cox regression disregarding any clustering), K-PAFT (kM clustering followed by cluster-specific penalized AFT) and N-PAFT (Penalized AFT disregarding any clustering).

Our SBC algorithm generally achieves a higher adjusted Rand Index, C-index than competing methods (Fig. 3 and Supplementary Material). As expected, increasing dimension and fraction of irrelevant features had a negative influence on SBC performance, but altogether the advantage over competing methods still remained. This held true also for detecting truly relevant features.

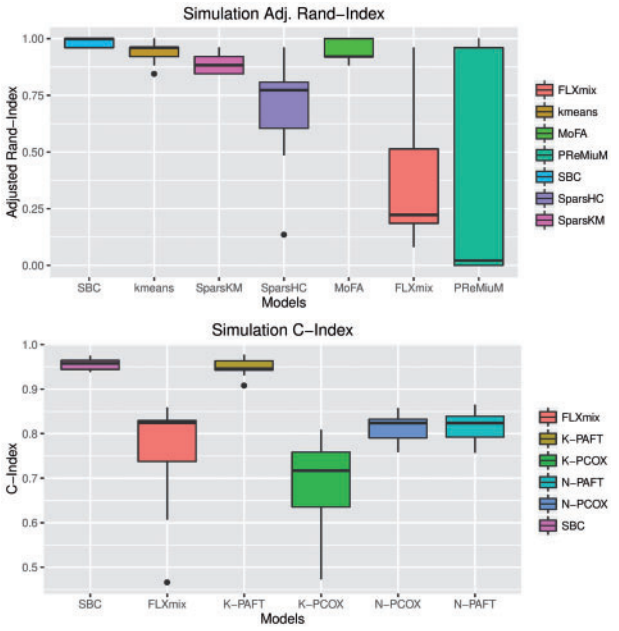

Fig. 3. Simulation data results with our model SBC on training dataset with 20% noisy features, $D = 20$

5 Real data

We apply our SBC approach on two gene expression cancer datasets and our iSBC method on a multi-omics dataset. In-order to demonstrate the predictive ability of our SBC or iSBC approach we use five times repeated 5-fold cross-validation and compare its performance with competing methods. For the biological interpretation of our method we choose to present detailed results of one randomly chosen training-testing data-split for each of the three real datasets. This example data-split divides each of the three datasets into equal training-testing partitions.

5.1 Breast cancer

We used the breast cancer microarray dataset used in Van De Vijver *et al.*, 2002) and available through the seventyGeneData R-package. For the clinical endpoint we used '**time to metastatis**' along with the corresponding censoring indicator for metastatis. The authors classified the data into two groups, we call this clustering as Vijver classification (referred as VV). The 70-gene signature Van't Veer *et al.* (2002) was used to compare with our approach to stratify 295 patients in terms of our clinical end-point. In order to reduce the dimensionality of the data we pre-filtered genes according to two criteria: (i) using the most significant P-values from univariate cox-regression models and (ii) using a t-test between metastatic and non-metastatic groups. Taking the intersection of these two ranked sets we arrive at a pre-filtered list of genes which is subsequently referred to as 'SBC signaturetl Notably, the same pre-filtering was also applied to two of the competing methods to ensure fair comparison (see below). Our SBC approach outperformed the following competing methods for survival-prediction during cross-validation procedure (measured using C-Index) (see Fig. 4):

- An average linkage HC of patients on the training data (using the SBC signature) within the cross-validation procedure followed by k-nearest neighbour (k-NN) predictions for the cluster membership on the test data and survival predictions by a penalized

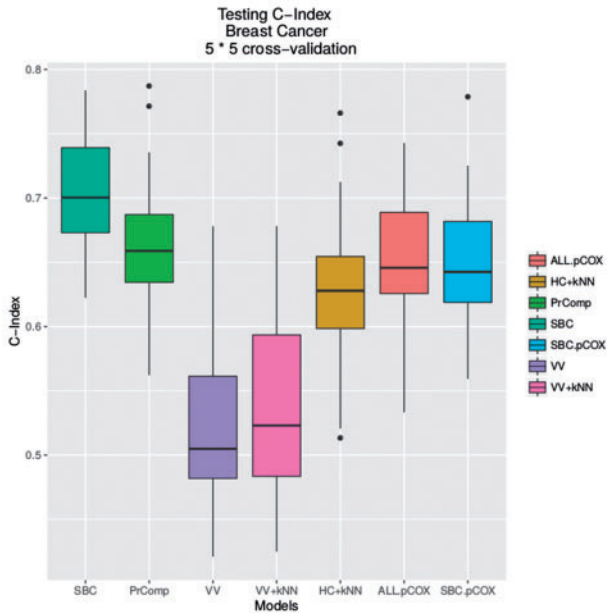

Fig. 4. Results on the breast cancer dataset. Box plots depict cross-validated C-indices for different methods

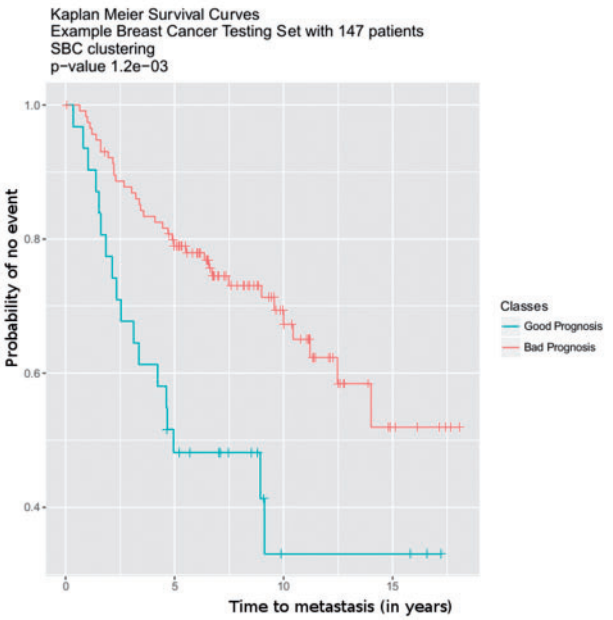

Fig. 5. Results on the Breast Cancer test dataset with the example training-testing split. Predicted classes from SBC. Crosses indicate censored outcomes. Clinical end point is time to metastasis

Cox regression model (pCOX). This approach was taken in the spirit of van't Veer *et al.* (abbreviated as HC+kNN).

- The same setting, but with original grouping of patients according to VV together with the 70 gene signature and then followed by k-NN together with pCOX (abbreviated as VV+kNN)
- Using classification by Vijver *et al.* on the training and test sets and building cluster-specific pCOX models (abbreviated as VV)
- Taking the first 20 principal components of the whole set of features on the training data, within the cross-validation procedure and using a pCOX. That means test data within the cross-validation procedure was first projected on the first 20 principal components constructed on the training data, and then survival predictions were performed via a pCOX model. (abbreviated as PrComp)
- A single L1-regularized Cox regression model (disregarding clustering) on a) the whole set of features (ALL.pCOX) and b) on the pre-filtered SBC features (SBC.pCOX)

Supplementary Figs S8 and S9 in addition indicate that SBC yields a separation of survival curves in different clusters that was at least as good as that obtained with competing stratification approaches (HC, original VV grouping). Instead of the HC, we also tried the kM clustering and the results were similar.

Next, we demonstrate the results obtained with our SBC method when training the model on a randomly chosen subset of 50% of the samples. Our SBC signature for this split comprised of 58-probe IDs. We obtained two clusters namely, 'Good Prognosis' (median time to distant metastasis 93 months) and 'Bad Prognosis' (median time to distant metastasis 47 months). These two clusters yielded two well separated survival curves ($P = 1.7e{-}08$) on the training dataset. We then used our method to predict class memberships and survival times of patients (see convergence diagnostic plot in Supplementary Material). On the testing set (the 50% of the samples not used for model training), this yielded two clusters which have significant differences in their survival curves (see Fig. 5). Further investigation of the two clusters obtained by our SBC method showed that the Bad prognosis group was significantly enriched ($P = 2.4e{-}15$, hypergeometric test) in the Estrogen Receptor negative (ER−) type. ER status has been long established as risk factor for metastatic breast cancer (Parl *et al.*, 1984). We also found significant enrichment ($P = 2.5e{-}05$, hypergeometric test) of the Good Prognosis cluster with the Luminal sub-type which has been reported to be associated with better prognosis Sørlie *et al.* (2001). Over-representation analysis of our SBC signature with respect to Gene Ontology terms revealed the significant 'Protein Methyltransferase Activity' [false discovery rate (FDR) < 0.05]. This process is indeed of known relevance for breast cancer (Kleer *et al.*, 2003).

A further ranking of the SBC genes w.r.t. their importance for clustering indicated a particular strong influence of E2F1 and TIMELESS. The gene E2F1 has been established to be related to breast cancer and is even prognostic for metastasis (Han *et al.*, 2003) while the circadian gene TIMELESS has been postulated as a risk factor for breast cancer tumorigenesis (Fu *et al.*, 2012). Another important gene according to SBC was Progestrone Receptor, whose role in breast cancer has been long known (Horwitz and McGuire, 1978). Other noteworthy genes include Reticulon 3 (RTN3), which has been associated to cell apoptosis (Lee *et al.*, 2009), and IGFBP5, which has been related to cell growth in breast cancer (Sheikh *et al.*, 1992).

5.2 Glioblastoma I (Verhaak *et al.*)

We also applied our SBC model on the Glioblastoma Multiforme (GBM) microarray data from Verhaak *et al.* (2010). The data were downloaded from https://tcga-data.nci.nih.gov/docs/publications/gbm_exp/.

We considered the '**overall survival**' as the clinical endpoint in our analysis. Overall, 196 patients (with survival information) were selected along with the original 840 gene Verhaak *et al.*, 2010) signature which we used for comparison (henceforth known as the Verhaak signature). Using only the training data we filtered features based on their *P*-values from uni-variate Cox Regression models and

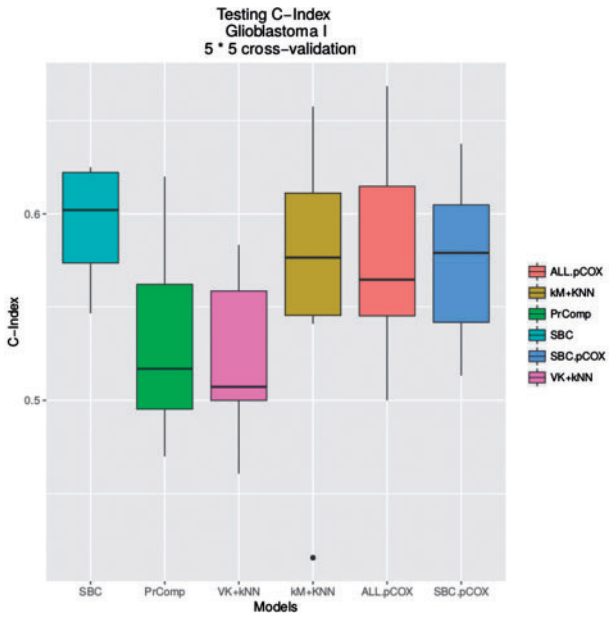

Fig. 6. Results on the Glioblastoma I dataset. Box plots depict cross-validated C-indices for different methods

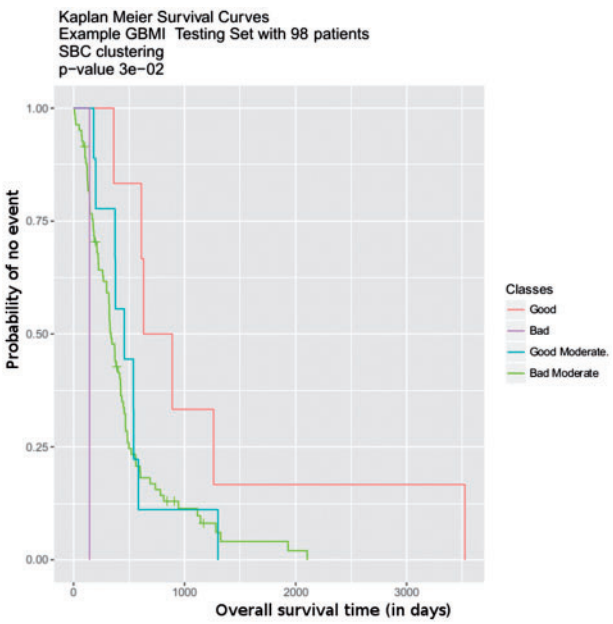

Fig. 7. Results on Glioblastoma I test dataset with example training-testing split. Predicted classes from SBC. Crosses indicate censored outcomes. Clinical end-point is overall survival

chose the top genes as our SBC signature. For the Cross-validation we used the same technique to arrive at the SBC model within each of the cross-validation loops. Our method was able to predict survival better than the following methods (see Fig. 6):

- A kM clustering of patients on the training data (using the SBC signature) and a combination of k-NN cluster assignment followed by a cluster specific penalized Cox regression. (abbreviated as kM+KNN)
- Using the original 840 gene signature of Verhaak *et al.* and their classification (VK) we trained a k-NN model for prediction. We then used this classification to build clustered pCOX models (abbreviated as VK +kNN)
- The PrComp, ALL.pCOX and SBC.pCOX, as defined earlier.

In addition, Supplementary Figures S13 and S14 indicate a better separation of survival curves with SBC than achieved by original VK stratification, VK + kNN and kM + KNN.

For our example data-split we chose top 47 genes as the SBC signature and trained our SBC model. Using that we discovered four distinct clusters (see convergence diagnostic plot in Supplementary Material) with unequal numbers of patients (10, 5, 25, 58). These clusters showed molecular differences as well as significantly different survival curves also on the test set (Fig. 7). We referred to the four clusters as 'Good', 'Good Moderate', 'Bad Moderate' and 'Worst' based on their respective mean survival times (830, 626, 380 and 180 days). Looking at the patients in the 'Best' prognosis cluster we find a high enrichment ($P = 3.5e{-}0.5$, hypergeometric test) in the 'Proneural' GBM sub-type defined by Verhaak *et al.* which has been reported in the literature to be linked with better survival (Cooper *et al.*, 2010). As in the breast cancer dataset, we again computed the feature importance of the SBC signature, one particular gene which has a higher contribution across all cluster comparisons (see Supplementary Material) is the 'Programmed cell death 6' or PDCD6 gene. It has been known for its proapoptotic function and is thought to be involved in survival pathways in cancer (Su *et al.*, 2012). Another interesting gene, which is assigned a high relevance by our method is TUSC4. TUSC4 has been established as a tumour suppressor gene regulating BRCA1 stability (Peng and Lin, 2014). BRCA1 expression has been reported as a biomarker for GBM prognosis Vassilakopoulou *et al.* (2015).

5.3 Glioblastoma II (The Cancer Genome Atlas-GBM)

We illustrate the application of our iSBC model on an alternative GBM dataset from The Cancer Genome Atlas (TCGA). We considered mRNA and miRNA expression and downloaded the data from https://tcga-data.nci.nih.gov/tcga/. 'Overall survival' was considered as the clinical end-point. 189 patients were considered, only those patients were included which were part of our earlier Glioblastoma I study. This was done so that we could compare benefits of data integration on a consistent dataset. For our iSBC method we perform the same type of pre-filtering on the training data as described before for Glioblastoma I dataset. Again we compared our two methods (iSBC and CCA pre-processed iSBC referred as C.iSBC) within a five times repeated 5-fold cross-validation procedure against:

- A combination of k-NN cluster assignment followed by a cluster specific penalized Cox regression (abbreviated as KMkN) using the SBC signature with the concatenated matrix of mRNA and miRNA expression profiles for each patient. When CCA features are used the method is referred to as C.KMkN.
- The PrComp method, as defined above but this time applied to the concatenated data matrix of gene and miRNA expression profiles.
- Single (disregarding clustering) Penalized Cox regression applied on the concatenated matrix with all the features, referred to as A.pCOX. When SBC features are used, it is referred to as B.pCOX. Although when CCA features are used we refer to it as C.pCOX.

Our results (Fig. 8) indicate at least as good prediction performance with our iSBC and C.iSBC methods than with competing ones

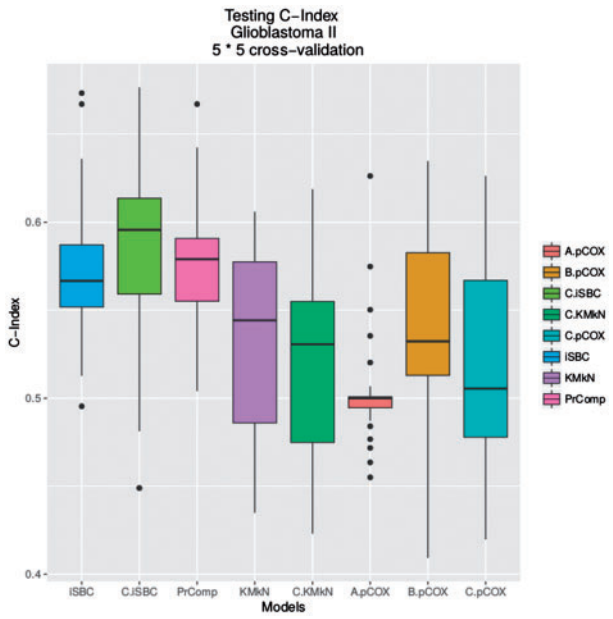

Fig. 8. Results on the Glioblastoma II dataset. Boxplots depict cross-validated C-indices for different methods

(PrComp). At the same time Supplementary Figures S17 and S18 show that our methods separated survival curves better (after predicting cluster membership of test patients) than a kM clustering approach or kM plus kNN cluster membership predictions.

Delving deeper in the example data-split we selected 31 top ranking mRNAs and the top 31 miRNA probes as our iSBC signature. We then applied our iSBC method once with and once without projecting data on the top 10 canonical covariates. The CCA preprocessing leads to slight increase in the survival prediction (Fig. 8 and Supplementary Material). In the following part we focus our discussion on the solution obtained without CCA preprocessing. Application of our iSBC approach lead to the discovery of four clusters (we call them, as before, 'Worst', 'Good Moderate', 'Bad Moderate', 'Best' based on the prognosis) of unequal number of patients (2, 27, 54, 13). The clusters from our iSBC still result in clearly separable survival curves on both training and test datasets (see Supplementary Material). We further investigated cluster-specific enrichment with respect to somatic mutations. The mutation pattern found in genes included in our model is significantly related to the iSBC derived clusters ($P = 1e-05$, χ^2-test, see Supplementary Material for further details). An interesting observation was the mutual exclusive mutation pattern of TP53 and PTEN genes among the iSBC clusters, meaning that if TP53 was found mutated in one iSBC cluster, PTEN was never mutated in that cluster and vice-versa. This mutual exclusivity has also been reported in literature (Kurose *et al.*, 2002). Over-representation analysis of the iSBC signature revealed the significant Gene Ontology term 'negative regulation of G1/S transition of mitotic cell cycle' (FDR < 0.05). This is highly interesting because cancer cells have an over-active cell cycle, leading to proliferation and hinting at possible mechanism for cancer progression. Looking at the most discriminatory features from our iSBC model (see Supplementary Material for details), we find that one important mRNA iSBC feature is 'developmentally regulated GTP-binding protein 2' or DRG2 gene which has been shown to induce apoptosis in cancer cells Jie *et al.*, 2012). Another interesting and discriminatory gene is β-catenin (CTNNB1), which is a key protein in the Wnt signaling pathway. Deregulation of the Wnt pathway has been associated with various cancers, including GBM Lee *et al.* (2016). Another discriminatory gene identified by iSBC is ADAM22, which has been shown to be under-expressed in high-grade gliomas Giovanna *et al.* (2006). An important miRNA feature miR-661 is known to activate the p53 pathway and suppresses tumour progression (Hoffman *et al.*, 2014). Furthermore we found miR-675, which has been linked to Gliomas (Shi *et al.*, 2014) while miR-637 has been shown to inhibit tumorigenesis in various cancer types (Zhang *et al.*, 2011) and is discussed as a prognostic marker in gliomas Que *et al.* (2015).

6 Conclusion

We have introduced a novel fully Bayesian clustering algorithm (SBC) which takes in clinical end-points of patients along with heterogeneous-omics data to perform two tasks in one—(i) patient subgroup identification on training data and (ii) prediction of patient sub-group and survival time on testing data. Our method was based on the motivation of discovering clusters of patients using their distinct molecular signatures and strong survival curve separability. Another important motivation was the predictive utility of our approach along with biological interpretability. We have shown with simulations and real data that our method outperforms ad-hoc algorithms like kM followed by fitting cluster-specific survival models. Furthermore, our SBC yields clearly better results than a hierarchical Gaussian DPMM without survival information, indicating the relevance of the clinical outcome in our model (see Supplementary Material). We believe the ability of SBC to identify patient-subgroups differing in survival constitutes an advantage compared witho existing approaches like Van't Veer *et al.* (2002) and Verhaak *et al.* (2010). Furthermore, SBC is principally able to take into account more than one -omics data source. Our assumed cluster specific factorization of the complete likelihood essentially weighs features inversely to their noise level. The CCA preprocessing approach explored here is a refinement in that context, which could potentially also allow for combining discrete with continuous data types, as e.g. shown in Witten *et al.* (2009). In future research we want to explore this aspect further and see, how CCA or similar latent factor approaches could be integrated better into our SBC method.

From a statistical point of view SBC is a coherent clustering scheme which groups data points based on their similarities to each other and their similarities to their (possibly) censored response variable. We have also used penalized estimation of the parameters which allows us to deal with $n < p$ problem, casting it in a Bayesian hierarchical setting. Our simulation results point to the superiority of our method in comparison to other state-of-the art techniques. On real data we have shown the ability of SBC to discover and predict hitherto unknown clusters which also show distinct progression patterns. Notably, the run time of our method for these applications varied between 1.5 and $\sim$2 h, which appears practically affordable. Of course, larger datasets are expected to require longer Gibbs sampling and thus more computation time. In practice it is thus recommended to reduce the number of features before applying SBC.

One of the key challenges for any clustering algorithm for biological data is to explore the biological underpinnings of the obtained clusters. In this regard we have found that certain sub-types from our model are particularly enriched in certain biological markers (e.g. ER status for breast cancer) and also correlate strongly with some sub-types in the well-established classification schemes for example of Sørlie *et al.* (2001), Verhaak *et al.* (2010). Altogether we

see SBC as a step towards a more clinically relevant dissection of patient heterogeneity.

Conflict of Interest: none declared.

References

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Antoniak,C.E. (1974) Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Ann. Stat.*, 1152–1174.

Bair,E. and Tibshirani,R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.*, **2**, e108.

Beer,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

Blackwell,D. and MacQueen,J.B. (1973) Ferguson distributions via pólya urn schemes. *Ann. Stat.*, 353–355.

Bouriga,M. and Féron,O. (2013) Estimation of covariance matrices based on hierarchical inverse-wishart priors. *J. Stat. Plan. Inferen.*, **143**, 795–808.

Cooper,L.A. *et al.* (2010) The proneural molecular signature is enriched in oligodendrogliomas and predicts improved survival among diffuse gliomas. *PloS One*, **5**, e12548.

Ferguson,T.S. (1973) A bayesian analysis of some nonparametric problems. *Ann. Stat.*, 209–230.

Fu,A. *et al.* (2012) Genetic and epigenetic associations of circadian gene timeless and breast cancer risk. *Mol. Carcinogenesis*, **51**, 923–929.

Giovanna,M. *et al.* (2006) Adam22, expressed in normal brain but not in high-grade gliomas, inhibits cellular proliferation via the disintegrin domain. *Neurosurgery*, **58**, 179–186.

Görür,D. and Rasmussen,C.E. (2010) Dirichlet process gaussian mixture models: Choice of the base distribution. *J. Comp. Sci. Technol.*, **25**, 653–664.

Grün,B. and Leisch,F. (2007) Fitting finite mixtures of generalized linear regressions in R. *Comput. Stat. Data Anal.*, **51**, 5247–5252.

Han,S. *et al.* (2003) E2f1 expression is related with the poor survival of lymph node-positive breast cancer patients treated with fluorouracil, doxorubicin and cyclophosphamide. *Breast Cancer Res. Treat.*, **82**, 11–16.

Harrell,F.E. *et al.* (1982) Evaluating the yield of medical tests. *Jama*, **247**, 2543–2546.

Hoffman,Y. *et al.* (2014) mir-661 downregulates both mdm2 and mdm4 to activate p53. *Cell Death Differ.*, **21**, 302–309.

Horwitz,K.B. and McGuire,W. (1978) Estrogen control of progesterone receptor in human breast cancer: correlation with nuclear processing of estrogen receptor. *J. Biol Chem.*, **253**, 2223–2228.

Hotelling,H. (1936) Relations between two sets of variates. *Biometrika*, **28**, 321–377.

Jie,C. *et al.* (2012) Skp1-cullin1-f-box (scf)-mediated drg2 degradation facilitated chemotherapeutic drugs induced apoptosis in hepatocellular carcinoma cells. *Biochem. Biophys. Res. Commun.*, **420**, 651–655.

Kleer,C.G. *et al.* (2003) Ezh2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc. Natl. Acad. Sci. USA*, **100**, 11606–11611.

Koestler,D.C. *et al.* (2010) Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. *Bioinformatics*, **26**, 2578–2585.

Kurose,K. *et al.* (2002) Frequent somatic mutations in pten and tp53 are mutually exclusive in the stroma of breast carcinomas. *Nat. Genet.*, **32**, 355–357.

Lapointe,J. *et al.* (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. USA*, **101**, 811–816.

Lee,J.T. *et al.* (2009) Over-expression of reticulon 3 (rtn3) enhances trail-mediated apoptosis via up-regulation of death receptor 5 (dr5) and down-regulation of c-flip. *Cancer Lett.*, **279**, 185–192.

Lee,Y. *et al.* (2016) Wnt signaling in glioblastoma and therapeutic opportunities. *Lab. Invest.*, **96**, 137–150.

Liverani,S. *et al.* (2015) PReMiuM: An R package for profile regression mixture models using Dirichlet processes. *J. STAT. SOFTW.*, **64**, 1.

McLachlan,G. and Peel,D. (2000) Mixtures of factor analyzers. *Finite Mixture Models*, 238–256.

Medvedovic,M. and Sivaganesan,S. (2002) Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, **18**, 1194–1206.

Medvedovic,M. *et al.* (2004) Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, **20**, 1222–1232.

Neal,R.M. (2000) Markov chain sampling methods for dirichlet process mixture models. *J. Comput. Graph. Stat.*, **9**, 249–265.

Park,T. and Casella,G. (2008) The Bayesian lasso. *J. Am. Stat. Assoc.*, **103**, 681–686.

Parl,F.F. *et al.* (1984) Prognostic significance of estrogen receptor status in breast cancer in relation to tumor stage, axillary node metastasis, and histopathologic grading. *Cancer*, **54**, 2237–2242.

Peng,Y. and Lin,S.-Y. (2014) Tusc4 functions as tumor suppressor by regulating brca1 stability and functions. *Cancer Res.*, **74(Suppl. 19)**, 1573–1573.

Que,T. *et al.* (2015) Decreased miRNA-637 is an unfavorable prognosis marker and promotes glioma cell growth, migration and invasion via direct targeting Akt1. *Oncogene*, **34**, 4952.

Raman,S. *et al.* (2010) Infinite mixture-of-experts model for sparse survival regression with application to breast cancer. *BMC Bioinformatics*, **11**, 1.

Rasmussen,C.E. (2000) The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems*, pp. 554–560.

Royston,P. (2001) The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. *Stat. Neerland.*, **55**, 89–104.

Sha,N. *et al.* (2006) Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, **22**, 2262–2268.

Sheikh,M.S. *et al.* (1992) Identification of the insulin-like growth factor binding proteins 5 and 6 (igfbp-5 and 6) in human breast cancer cells. *Biochem. Biophys. Res. Commun.*, **183**, 1003–1010.

Shi,Y. *et al.* (2014) Long non-coding rna h19 promotes glioma cell invasion by deriving mir-675. *PLoS One*, **9**, e86295.

Shipp,M.A. *et al.* (2002) Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, **8**, 68–74.

Sørlie,T. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA*, **98**, 10869–10874.

Su,D. *et al.* (2012) Pdcd6 is an independent predictor of progression free survival in epithelial ovarian cancer. *J. Transl. Med.*, **10**, 1.

Van De Vijver,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

Van't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Vassilakopoulou,M. *et al.* (2015) Geno-21brca1 protein expression predicts survival in glioblastoma patients from a nrg oncology/rtog cohort. *Neuro-Oncology*, **17**(Suppl. 5), v96.

Verhaak,R.G. *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell*, **17**, 98–110.

Wei,L. (1992) The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Stat. Med.*, **11**, 1871–1879.

Witten,D.M. and Tibshirani,R. (2010) A framework for feature selection in clustering. *J. Amer. Statist. Assoc.*, **105**, 713–726.

Witten,D.M. *et al.* (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534.

Yau,C. and Holmes,C. (2011) Hierarchical bayesian nonparametric mixture models for clustering with variable relevance determination. *Bayesian Anal. (Online)*, **6**, 329.

Yeung,K.Y. *et al.* (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.

Yuan,Y. *et al.* (2011) Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.*, **7**, e1002227.

Zhang,J-F. *et al.* (2011) Primate-specific microrna-637 inhibits tumorigenesis in hepatocellular carcinoma by disrupting signal transducer and activator of transcription 3 signaling. *Hepatology*, **54**, 2137–2148.