# Supplementary Material for Towards Clinically More Relevant Dissection of Patient Heterogeneity via Survival based Bayesian Clustering

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

[1] University of Bonn, Bonn Aachen International Center for Information Technology, Dahlmannstr. 2, 53127 , Bonn, GERMANY,
ashar@bit.uni-bonn.de,
[2] UCB Biosciences GmbH, Alfred-Nobelstr. 10, 40789, Monheim, Germany

## 1 Methods

### 1.1 Dirichlet Process Prior

DP can be thought of as a distribution over distributions. It is known that Dirichlet Process Prior can also be obtained by taking the limit of a finite mixture model with $K$, the number of clusters going to infinity i.e. $K \to \infty$. It was shown by [Neal, 2000] that by introducing class labels $c_i$ for each data point the old formulation of [Blackwell and MacQueen, 1973] can be re-written as following:

$$X_i|c_i, \theta_i \sim F(\theta_{c_i})$$

$$c_i|\mathbf{p} \sim Multinomial(p_1, \ldots, p_K)$$

$$\theta_i \sim G_0$$

$$\mathbf{p} \sim Dirichlet(\alpha/K, \ldots, \alpha/K)$$

After taking the limit $K \to \infty$ and integrating out the mixing proportions $\mathbf{p}$, the conditional distribution over the class labels $c_i$ can be formulated as:

$$P(c_i = c|c_1, \ldots, c_{i-1}) \to \frac{n_{i,c}}{i - 1 + \alpha}$$

$$P(c_i \neq c_j forall j < i|c_1, \ldots, c_{i-1}) \to \frac{\alpha}{i - 1 + \alpha}$$

Dirichlet Process also defines a probabilistic model on the partition of the data points which can be imagined by $\mathbf{c} := (c_1, \ldots c_N)$:

$$p(\mathbf{c}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \prod_{j=1}^{k=K} \frac{\Gamma(n_j + \alpha/K)}{\Gamma(\alpha/K)}$$

For the purpose of sampling, we can sample from the conditional prior by imagining that $i$ is the last of the $N$ observations :

$$\theta_i|\theta_1,\dots\theta_{-i} \sim \frac{1}{N-1+\alpha}\sum_{j\neq i}\delta(\theta_j) + \frac{\alpha}{N-1+\alpha}G_0$$

The parameter $\alpha$ (also know as the concentration parameter) controls the prior number of expected clusters. Following [Görür and Rasmussen, 2010] it has been given an inverse gamma prior expressing the belief that apriori we do not expect a large number of clusters.

$$p(\alpha^{-1}) \sim Gamma(0.5, 2)$$

For further details one can refer to [Neal, 2000].

### 1.2 Hierarchical Multivariate Gaussian Model

We use a conjugate Dirichlet Process Gaussian Mixture Model as it allows for the possible marginalization of the cluster-specific parameters. More concretely, the joint distribution of the mean $\mu_j$ and the precision matrix $S_j$ follows a Normal/Wishart distribution

$$(\mu_j, S_j) \sim \mathcal{NW}(\xi, \rho, \phi, \phi W)$$

The parameter $\rho$ controls the strength of the dependence between the mean $\mu_j$ and the precision $S_j$; while $\phi$ controls the dependence between the hyperprior $W$ and precision matrix $S_j$. We used the following distribution priors:

$$\rho \sim Gamma(0.25, 2)$$

$$\frac{1}{\phi - D + 1} \sim Gamma(0.5, 2/D)$$

As $\phi$ controls the degrees of freedom in the Wishart distribution, it has been constrained as $\phi > D-1$. The hyper-parameters $\xi$ and $W$ are given priors based on the empirical Bayes estimates $\mu_y$ and $\Sigma_y$ from the data:

$$W \sim diag(\mathcal{W}(D, \Sigma_y/D))$$

$$\xi \sim \mathcal{N}(\mu_y, \Sigma_y)$$

### 1.3 Bayesian LASSO penalized Accelerated Failure Time Model

The Bayesian LASSO penalty amounts to placing a Laplacian prior on the coefficient matrix $\beta$ of the following form :

$$\pi(\beta|\sigma^2) = \prod_{j=1}^{p}\frac{\lambda}{2\sigma}e^{\lambda|\beta_j|/\sigma}$$

The penalty parameter $\lambda^2$ controls the level of sparsity and is given a gamma prior:

$$p(\lambda^2) \sim Gamma(r, \delta)$$

The values for $r = 1$ and $\delta = 1.78$ were set as in [Park and Casella, 2008]. The parameter $\sigma^2$ was given an inverse-gamma prior. The R-package *blasso* was used to sample parameters.

### 1.4 Model fitting via Gibbs Sampling

The conditional distribution of a data point to belong to a new cluster is as follows:

$$p(c_i \neq c_j \forall j | \mathbf{c}_{-i}, \mu, S, \beta_0, \beta, \sigma^2, \alpha)$$

$$\propto \frac{\alpha}{N - 1 + \alpha} \int_{\mu,S,\beta_0,\beta,\sigma^2}$$

$$\mathcal{N}(w_i | \beta_0 + \beta^T \mathbf{X}_i, \sigma^2) \mathcal{N}(\mathbf{X}_i | \mu, S^{-1}) \mathbf{d}G_0(\mu, S, \beta_0, \beta, \sigma^2)$$

As there is overall dependence between the parameters of the Gaussian mixture model $(\mu, S)$ and that of the BLASSO $(\beta_0, \beta, \sigma^2)$, the above integral is not easy to solve. This makes the overall Mixture Model non-conjugate and we resort to Neal [2000]'s auxiliary variable approach to sample from the above distribution. The key idea is to able to approximate the above integral's value by drawing auxiliary parameters from the prior-distribution and considering the problem to be temporarily a finite mixture model. In our case, we found that $U = 2$ auxiliary parameters are sufficient for good convergence. The model is fitted using an alternating Gibbs update scheme for cluster-specific parameter set $(\mu_j, S_j, \beta_{0j}, \beta_j, \sigma_j^2)$ and the class labels $(c_1, \dots c_N)$ each of which can now be sampled from it's conditional distribution. For our Gibbs Sampling we use 100 burn-in iterations and 200 MCMC samples with samples being drawn every 5th iteration (thinning). To assess the convergence of our MCMC chain, we looked at the log-likelihood trace plots. To get estimates for cluster membership of patients we use the mode of marginal posterior distribution of each of the class labels from our Gibbs sampling.

### 1.5 Making model predictions

**Survival Prediction** The weights $v_{jm}(X^*)$ for the discovered clusters $j = 1 \dots C_m$ in the MCMC sample $m$ are proportional to their corresponding densities :

$$v_{jm}(X^*) \propto \frac{n_{jm}}{N - 1 + \alpha} \mathcal{N}((X^* | \mu_{jm}, S_{jm}^{-1})$$

Apart from the already discovered clusters, the latent clusters (which do not have any data point in them) also contribute to the survival prediction according to the DP, their corresponding weight is given as:

$$v_j(X^*) \propto \frac{\alpha}{N - 1 + \alpha} \int \mathcal{N}((X^* | \mu_0, S_0^{-1}) d(G_0)$$

This integral is evaluated using the auxiliary variable formulation of [Neal, 2000] with number of auxiliary variables $U = 2$. The idea behind the auxiliary variables is to replace the above *integral* with (and hence $v_j(X^*)$ by $v_u(X^*)$) a *density* using parameters drawn from the prior distribution. The auxiliary variables thus resemble clusters with no points assigned to them. The weights $v_u(X^*)$ for these auxiliary variables are calculated using the following density:

$$v_u(X^*) \propto \frac{\alpha}{N-1+\alpha}\mathcal{N}(X^*|\mu_u, S_u^{-1})$$

where we sample $(\mu_u, S_u)$, the auxiliary parameters, $(u = 1, 2)$, from the prior distribution, which is Normal-Wishart $\mathcal{NW}$ conditioned on the hyper-parameters of the multivariate Gaussian model:

$$(\mu_u, S_u) \sim \mathcal{NW}(\xi, \rho, \phi, \phi W)$$

The corresponding auxiliary parameters for the AFT model $(\beta_{0u}, \beta_u)$ that are used for survival prediction are also drawn from their prior distribution $G_{0t}$ given by the Bayesian LASSO. Together with these weights the contribution for Survival prediction from the latent classes can then be written as:

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{u=1}^{2}(\beta_{0u} + \beta_u^T X^*) * v_u(X^*)$$

**Cluster Membership** Apart from the already discovered clusters, the Dirichlet Process prior also places non-zero probability for the test point to form a new cluster. This probability is given by:

$$p(c^* = c^{new}|X^*, \theta_{1:N}^{(1:m)}, c_{1:N}^{(1:m)}) = b\frac{\alpha}{N-1+\alpha}\int \mathcal{N}(X^*|\mu_0, S_0^{-1})d(G_0)$$

where $(\mu_0, S_0)$ are drawn from their prior distribution $G_0$ given by the Hierarchical Multivariate Gaussian Model as described above. In-order to avoid solving this integral, we again use the auxiliary variable approach of Neal [2000] to approximate the above probability. The details of the auxiliary variable approach are the same as for the above section on **Survival prediction**. This means that apart from the existing classes, the patient $X^*$ can form a new cluster with the probability:

$$p(c^* = c^{new}|X^*, \theta_{1:N}^{(1:m)}, c_{1:N}^{(1:m)}) = b\frac{\alpha}{N-1+\alpha}\mathcal{N}(X^*|\mu_u, S_u^{-1})$$

for auxilary variables $u = 1, 2$.

## 2 Simulation

### 2.1 Data Generation

We simulated the cluster-relevant features using cluster-specific parameters (mean vector and precision matrix) employing the *MixSim* R package Melnykov et al.

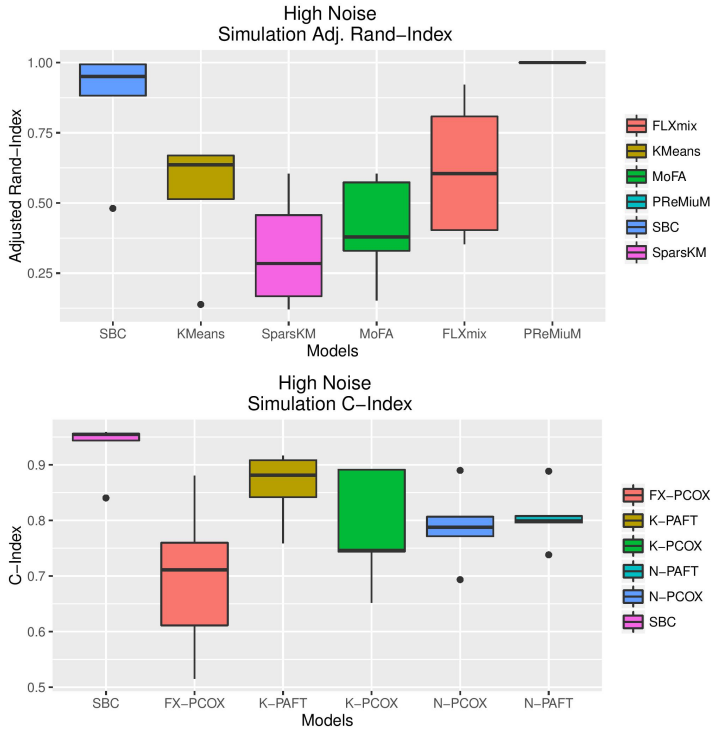


**Fig. 1.** Simulation results on the training set using SBC and the high noise scenario and D=20

[2012]. As mentioned in the main paper we have conducted the simulations in two scenarios to explore the effect of noise

1. The low noise scenario where 20 percent of the features were noise (uninformative for clustering) and there was 1 percent cluster overlap in the informative feature space (see main paper).
2. The high noise scenario where 50 percent of the features were noise (uninformative for clustering) and there was 10 percent cluster overlap in the informative feature space.

For the above two scenarios we use the cluster-specific parameters obtained from the *MixSim* R package and generate 100 points for training and 100 points for testing. In the results presented here, we simulated $K = 2$ clusters. We also used equal distribution of the data points in both the clusters (50,50). For each cluster, we then used the informative features to generate the survival times using randomly generated values for cluster-specific $(\beta_{0j}, \beta_j)$

### 2.2 Simulation results and comparisons

We initialized the model with k-means estimate by choosing $k$ with the help of silhouette plots. The superior results of our SBC model in comparison to other methods (also in high noise setting) demonstrate the need of integrating the survival times in clustering. We give a short summary of the competing clustering methods :

1. FlXmix Grun and Leisch [2008] is a curve clustering algorithm. FlexMix implements a general framework for fitting discrete mixtures of regression models. It allows the integration of Generalized linear models and penalized models. It uses an EM algorithm to estimate the parameters. We used FLXmix with *glmnet* R package for high-dimension regression. The clusters are initialized using a standard k-Means algorithm and the number of clusters are chosen based on the Bayesian Information Criterion (BIC).
2. PReMiuM Liverani et al. [2013] is a package for Bayesian clustering using a Dirichlet Process Mixture Model. It allows for both continuous/discrete variables response variables but does not deal with survival information. To implement our censored response variables, we disregarded censoring and considered the response as continuous. It also allows to make predictions. The number of clusters are discovered automatically using Dirichlet Process prior.
3. MoFA (Mixture of Factor Analyzers) McLachlan and Peel [2000] is a model-based density estimation to take into account noise in high dimensional data sets. We set the number of factors to be two and selected the number of clusters using BIC.
4. SpaseHC (Sparse Hierarchical Clustering) and Sparse-KM (Sparse K-means Clustering) are two algorithms in the R package *'sparcl'* Witten and Tibshirani [2012]. These two methods provide a principal way to deal with noisy data. The number of clusters are optimized by maximizing the average cluster silhouette width.
5. K-Means - For the case of two data sources we created a concatenated data matrix by joining the columns of the two data sources and running K-Means on the joint matrix. To choose the number of clusters we looked for clusters which maximized the average silhouette width of the clusters.

After having discovered the clustering we then fitted cluster-specific survival models using the R package *'glmnet'*.

### 2.3 Assessing the convergence of the Gibbs Sampler

In-order to assess the convergence of the MCMC sampler, we use log-likelihood trace plots. In Fig. 2 we show two such plots for the case of low and high noise scenarios with $D = 20$. In all our simulations we found 100 burn-in iterations to yield a convergent MCMC chain. We then used 200 MCMC samples for our posterior estimation (with samples being taken every 5th iteration).

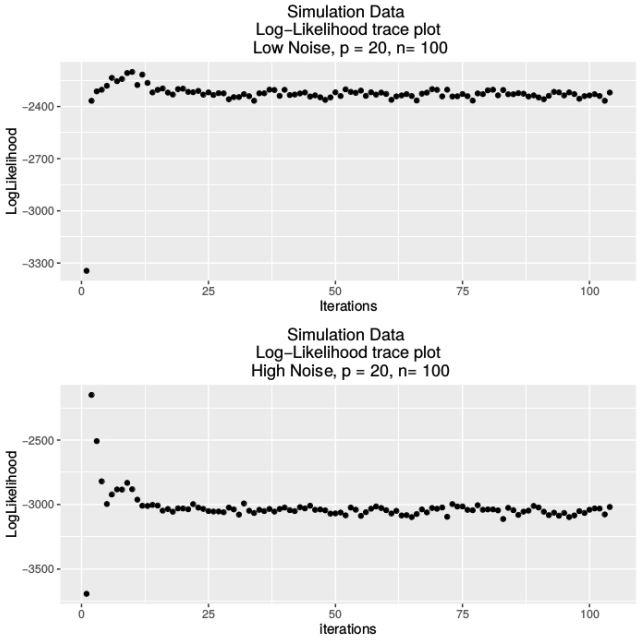


**Fig. 2.** Likelihood trace plots during the burnin period for the low and high noise scenarios

### 2.4 Effect of varying dimension $D$

We also varied the dimension (number of features) of the SBC, apart from the $D = 20$, we tested for $D = 10, 30, 50, 60$ shown in Fig.3,4,5,6. The results shown are for the training data set for 5 simulation repeats in each case. We can see that the model performance worsens as we increase the dimension, it still, however, performs better than the competing methods. The reason for which the model performance worsens on increasing the dimension is the following: as the SBC performs best when the clustering information is complementary in the molecular data and in the survival data, with increasing dimension the overall effect of the one dimensional survival information (on the data likelihood) decreases and the SBC is influenced more by the noisy molecular data at high $D$. We found that the SBC worked rather well on the range $D = 20$ to $D = 60$ and hence this range was used for the real data set to determine the SBC signature.

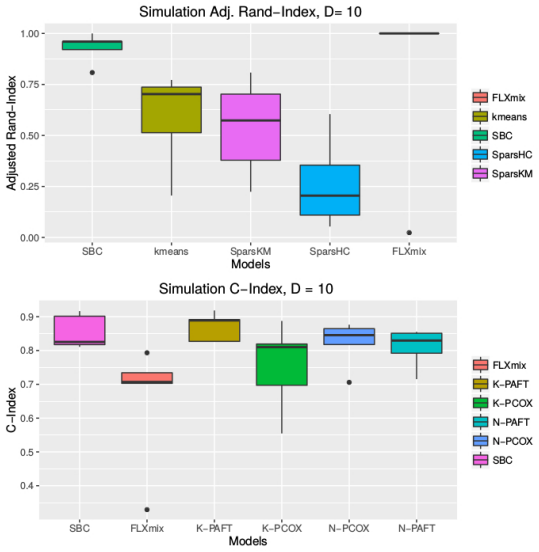


**Fig. 3.** Simulation results on the training set using SBC and the high noise scenario and D = 10.

### 2.5 Feature Importance from SBC model

As discussed above, our SBC model enables us to rank features on their ability to distinguish clusters. In our simulations we can compare the performance of the SBC to detect relevant features. From our SBC model we can get scores for the relevance of each feature which are either "relevant" or "non-relevant" and thus we can calculate the average Area Under the curve (AUC) for this

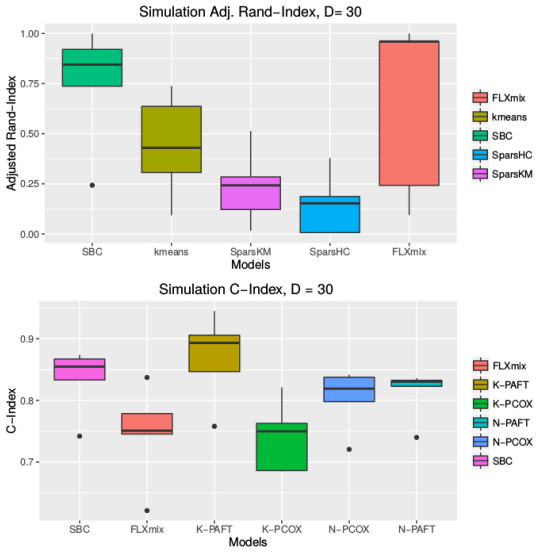


**Fig. 4.** Simulation results on the training set using SBC and the high noise scenario and D = 30.

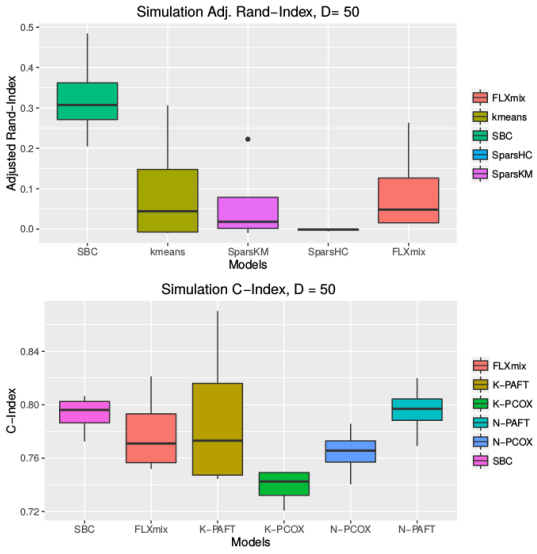


**Fig. 5.** Simulation results on the training set using SBC and the high noise scenario and D = 50.

classification. This has been shown in Fig.7 where we contrast this with penalized FlXmix [Grun and Leisch, 2008]. The results represent 5 simulation repeats and

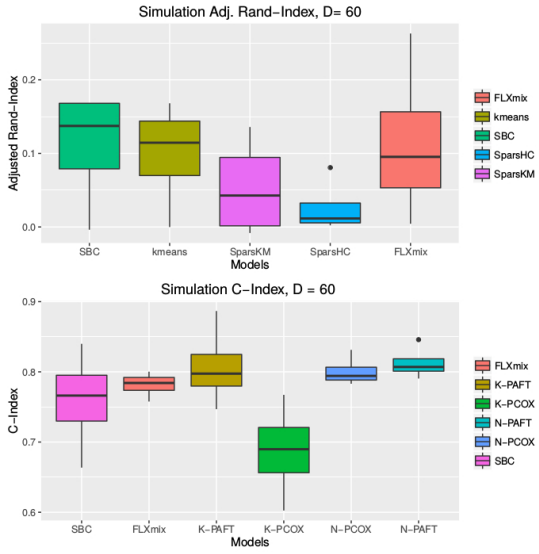


**Fig. 6.** Simulation results on the training set using SBC and the high noise scenario and D = 60.

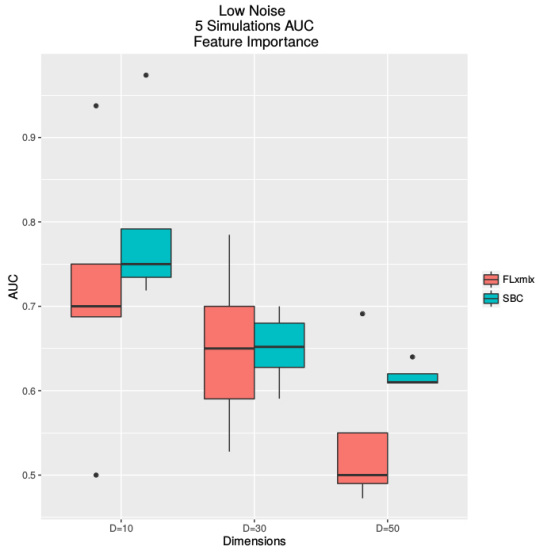


**Fig. 7.** Simulation results on the training set for detecting feature importance in the low noise scenario

we restrict ourselves to the "Low Noise scenario". One can see a similar trend to

the above simulations with increasing $D$ leading to a deterioration of the model performance.

# 3 Real Data

## 3.1 Breast Cancer Data Set

We first present here the detailed results for cross-validation study in Fig.8 and Fig.9. Depicted in these two figures is the test-statistic for the log-rank test comparing estimates of the hazard functions associated to the detected clusters. The test statistic is constructed by calculating the observed and expected number of events in each cluster at each observed time. A large value of the test statistic indicates a stronger deviance from the null hypothesis of no difference in the hazard functions of different clusters.

Extended results for Breast Cancer example data-split are also presented here. For the example data-split, we report in Table 1 our results. We use the log-likelihood trace plot to confirm the convergence of our Gibbs sampler, as can be seen in Fig.10

| METHOD (CLUSTERING OR CLASSIFICATION) | FEATURE SET (SIGNATURE) | TRAINING (T) or PREDICTION (P) | p-value (Log Rank) | C-Index |
|---|---|---|---|---|
| SBC | SBC | T | **1.7e-08** | **0.79** |
| SBC | SBC | P | **1.2e-03** | **0.70** |

**Table 1.** Breast Cancer Data Set Results on the example data-split

The results on the training data set are presented in Figure 11 where the molecular differences between the two SBC clusters are visually visible. The columns of the heat map are arranged according to the log-odds ratio of belonging to the two clusters. We also looked for enrichment of our SBC derived "Good prognosis" and "Bad Prognosis" classes with other key factors in breast cancer progression as shown in Table2. Likewise, a hyper-geometric test ($p=2.5e-05$ ) indicates a significant enrichment of the good prognosis cluster with the luminal breast cancer sub-type (Table3 ). Gene Ontology (GO) enrichment analysis of the SBC signature was carried out via a conditional hyper-geometric test (R-package GOstats [Falcon and Gentleman, 2007]). Multiple-testing correction was applied using [Benjamini and Hochberg, 1995] method to control the False Discovery rate.

## 3.2 Glioblastoma I (Verhaak et al.)

We present here the detailed results for cross-validation study in Fig.13 and Fig.14.

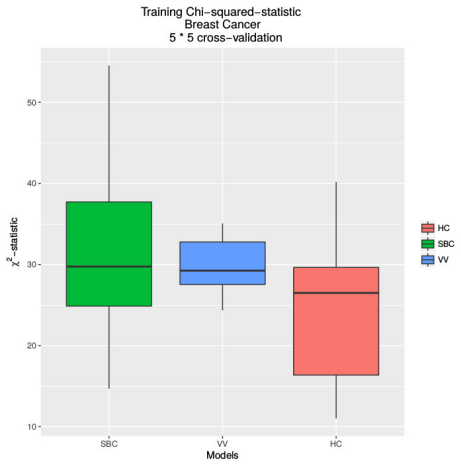


**Fig. 8.** Cross-validation results for Breast Cancer. Log-rank statistic is based on the **recovered classes** from the SBC model on the training set

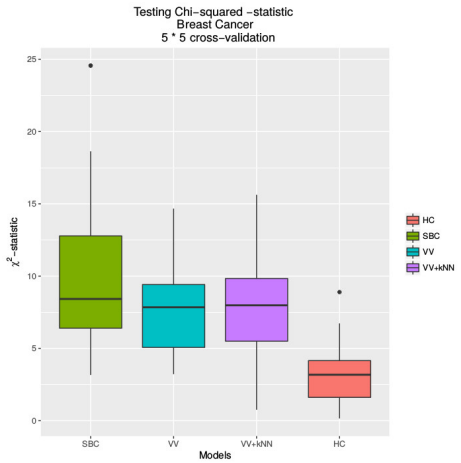


**Fig. 9.** Cross-validation results for Breast Cancer. Log-rank statistic is based on the **predicted classes** from the SBC model on the test set.

| | ER positive | ER negative |
|---|---|---|
| Bad Prognosis | 19 | 36 |
| Good Prognosis | 88 | 5 |

**Table 2.** Results on Breast Cancer Data set: Enrichment of SBC classes with ER status

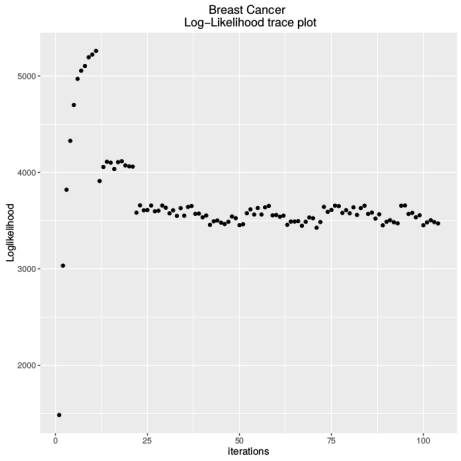


**Fig. 10.** Log-likelihood trace plots for the Breast Cancer Data Set

| | Luminal | Basal | ERBB2 | Normal |
|---|---|---|---|---|
| Bad Prognosis | 19 | 25 | 11 | 0 |
| Good Prognosis | 67 | 1 | 11 | 14 |

**Table 3.** Results on Breast Cancer Data set:Association of SBC classes with breast cancer sub-types

Extended results for Gliobalstoma I example data-split are reported here. For the example data-split, we report in Table 4 our results. We again use the log-likelihood trace plot to assess the convergence of our Gibbs sampling iterations as can be seen in Fig.15

| METHOD (CLUSTERING or Verhaak classification) | FEATURE SET (SIGNATURE) | TRAINING (T) or Prediction (P) | p-value (Log-rank) | C-Index |
|---|---|---|---|---|
| SBC | SBC | T | **5.3e-05** | **0.68** |
| SBC | SBC | P | **3e-02** | **0.56** |

**Table 4.** Glioblastoma I data set results for example data-split

There is a significant association between clusters discovered by our SBC and the ones reported by Verhaak et al., see Table 5 (p=$3.5e-05$, $\chi^2$ test). We also note that the Best prognosis class exclusively contained samples from the Proneural Verhaak GBM class while the Good Moderate prognosis class was split between Classical and Mesenchymal sub-types.
To better understand our SBC signature we plot the feature importance of all the genes in our SBC signature to distinguish between different clusters in Fig.

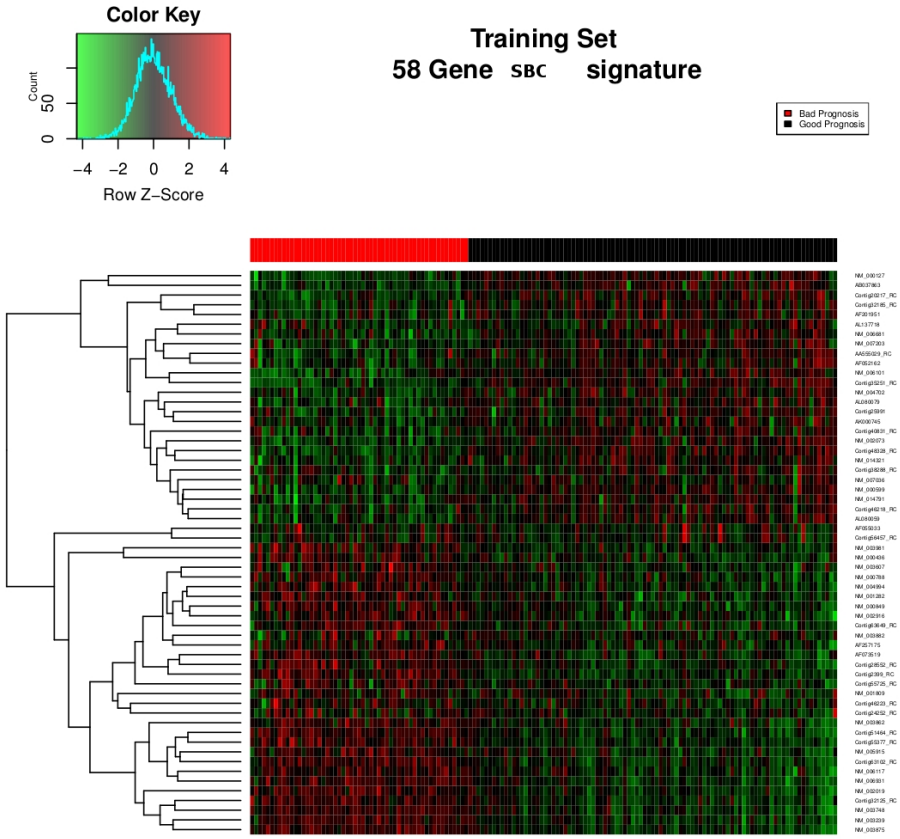


**Fig. 11.** SBC on Breast Cancer training set

| | Classical | Mesenchymal | Neural | Proneural |
|---|---|---|---|---|
| Best Prognosis | 0 | 0 | 0 | 10 |
| Worst Prognosis | 0 | 2 | 2 | 1 |
| Good Moderate Prognosis | 7 | 12 | 1 | 5 |
| Bad Moderate Prognosis | 19 | 15 | 15 | 12 |

**Table 5.** Results on Glioblastoma I: Association of SBC classes with GBM Verhaak sub-types

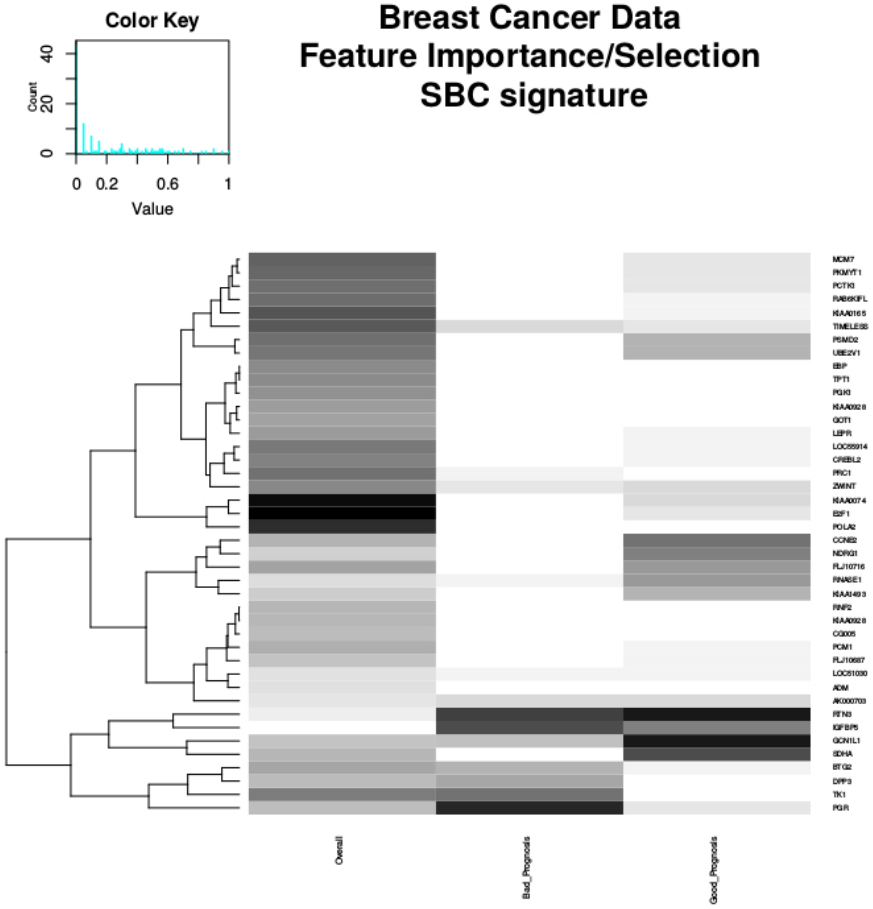


**Fig. 12.** Feature Importance and Selection from SBC on the Breast Cancer data set. The leftmost column represents importance of feature on molecular data clustering, the two right columns represent strength of association to cluster specific survival times. Darker colours imply stronger effects.

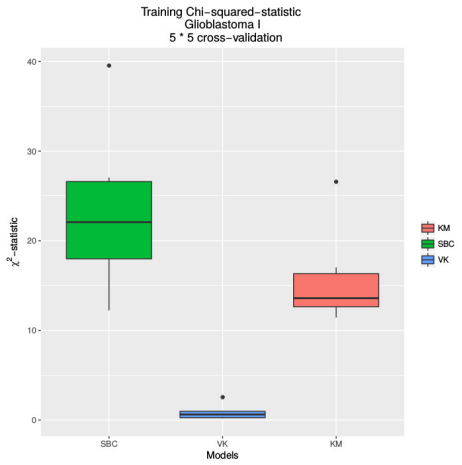


**Fig. 13.** Cross-validation results for GBM I. Log-rank statistic is based on the **recovered classes** from the SBC model on the training set.

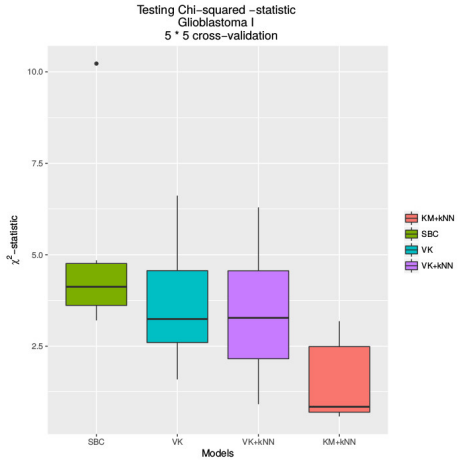


**Fig. 14.** Cross-validation results for GBM I. Log-rank statistic is based on the **predicted classes** from the SBC model on the test set.

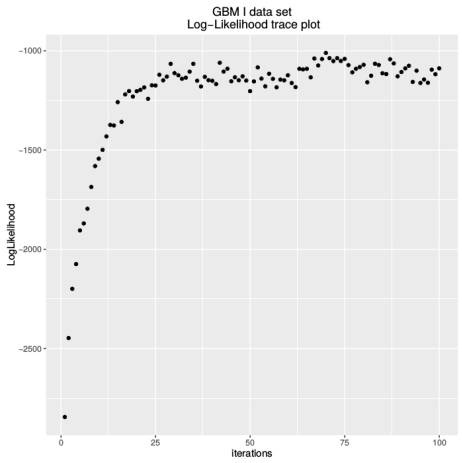


**Fig. 15.** Log-likelihood trace plots for the Glioblastoma I Set

16. Some genes which have higher contributions across all cluster comparisons (shown by a darker color in the heatmap in Fig. 16) were investigated to reveal interesting biological functions. The gene SLC25A38 which is a member of the SLC25 gene family and also plays a prominent role as a SBC signature gene has been reported to suppress cell growth in human gliomas [Waha et al., 2011].

### 3.3 Glioblastoma II (TCGA-GBM)

We present here the detailed results for cross-validation study in Fig.17 and Fig.18. Additionally, C.KM refers to K-means on the CCA processed features in Fig.17.

Extended results for the TCGA-GBM example split are also presented here. For the example data-split, we report in Table 6 our results. We again use the log-likelihood trace plot to assess the convergence of our Gibbs sampling iterations as can be seen in Fig.20. Our iSBC resulted in significantly different survival curves also on the test set (Fig.19) We conducted a cluster-specific somatic mutation enrichment analysis. For this purpose we looked for genes (which are part of our iSBC signature) which show cluster-specific somatic mutations. Somatic mutation data was only available for 23 patients out of 96 training patients. We obtained the mRNA signature from the SBC model, moreover miRNAs were also mapped to their gene targets using the 'multiMiR' package in R [Ru et al., 2014]. 57 unique genes were identified in this manner. 55 out of the 57 genes show the same pattern illustrated in table 4 where all of them show mutation exclusively in the best prognosis cluster of SBC. The interesting case is that of TP53 and PTEN genes which show a mutual exclusive behavior of somatic mutation as shown in Table 7 and Table 8.

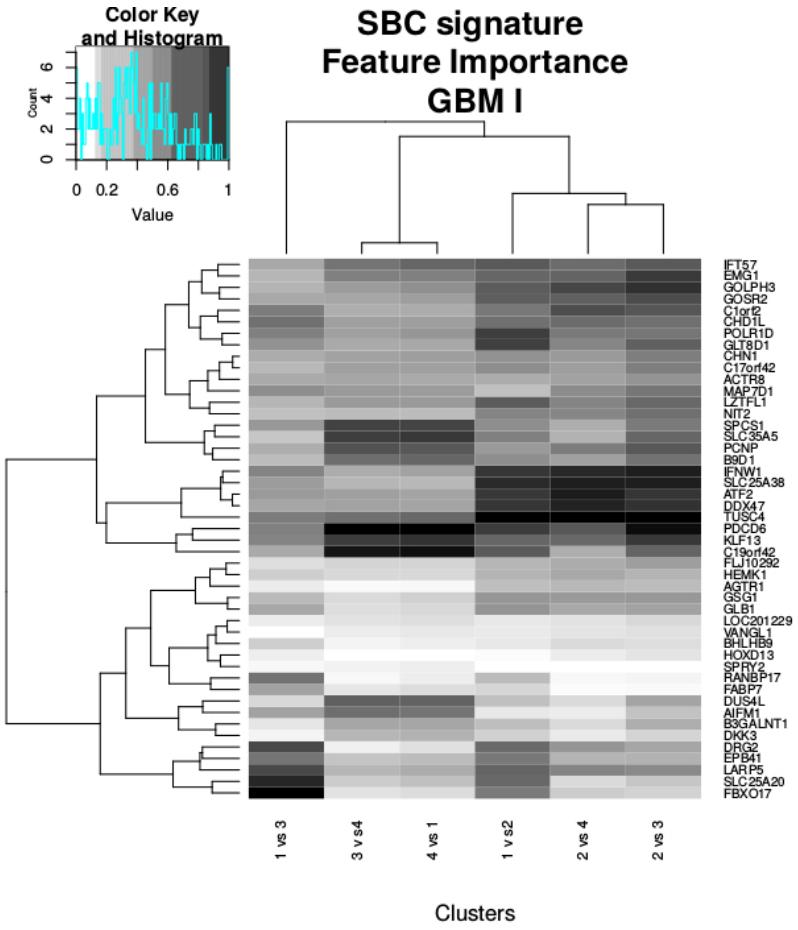


**Fig. 16.** Results on Glioblastoma I (SBC):Feature importance of the SBC signature on the GBM-Verhaak data set in discriminating respective clusters

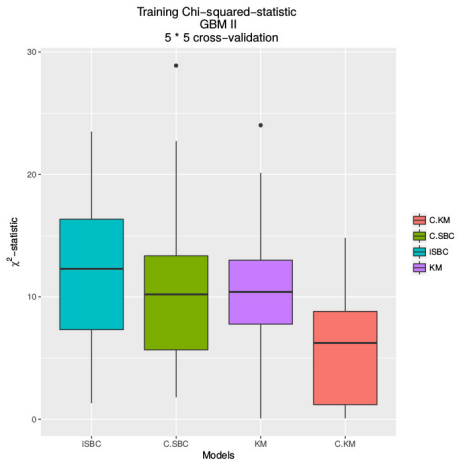


**Fig. 17.** Cross-validation results for GBM II. Log-rank statistic is based on the **recovered classes** from the iSBC model on the training set

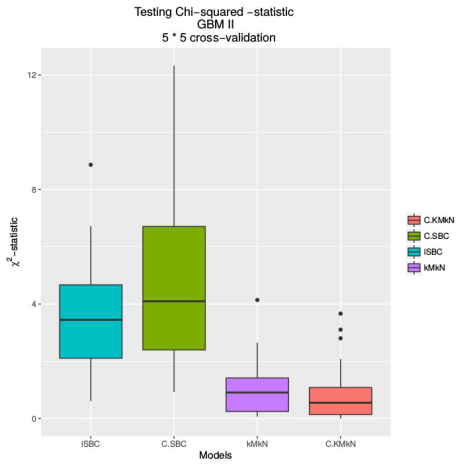


**Fig. 18.** Cross-validation results for GBM II.Log-rank statistic is based on the **predicted classes** from the iSBC model on the test set

In a similar manner as in GBM-Verhaak data set we plotted the feature importance of the mRNA and miRNA SBC signature in Fig.21 and Fig.22. We also explored features which had more contributions with respect to others (shown by darker colours in Fig.21 and Fig.22).

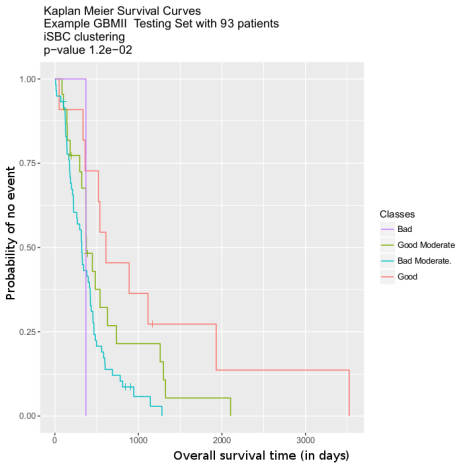


**Fig. 19.** Results on Glioblastoma II data set with example training-testing split. Predicted classes from iSBC on the test set. Crosses indicate censored outcomes. Clinical end-point is overall survival.

| METHOD | FEATURE SET (SIGNATURE) | TRAINING (T) or PREDICTION (P) | p-value Log-rank test | C-index |
|---|---|---|---|---|
| iSBC | iSBC | T | **6e-04** | **0.70** |
| iSBC | iSBC | P | **1e-02** | **0.52** |

**Table 6.** TCGA-GBM data set results for example data-split

| | Worst | Good Moderate | Bad Moderate | Best |
|---|---|---|---|---|
| Mutated | 0 | 0 | 0 | 4 |
| Non mutated | 0 | 6 | 13 | 0 |

**Table 7.** Results on Glioblastoma II (iSBC): Number of somatic mutations across iSBC defined clusters for signature genes except TP53 and PTEN

| | Worst | Good Moderate | Bad Moderate | Best |
|---|---|---|---|---|
| Mutated | 0 | 0 | 13 | 4 |
| Non mutated | 0 | 6 | 0 | 0 |

**Table 8.** Results on Glioblastoma II (iSBC):Number of somatic mutations across SBC defined clusters for TP53

| | Worst | Good Moderate | Bad Moderate | Best |
|---|---|---|---|---|
| Mutated | 0 | 6 | 13 | 0 |
| Non mutated | 0 | 0 | 0 | 4 |

**Table 9.** Results on Glioblastoma II (iSBC): Number of somatic mutations across SBC defined clusters for PTEN

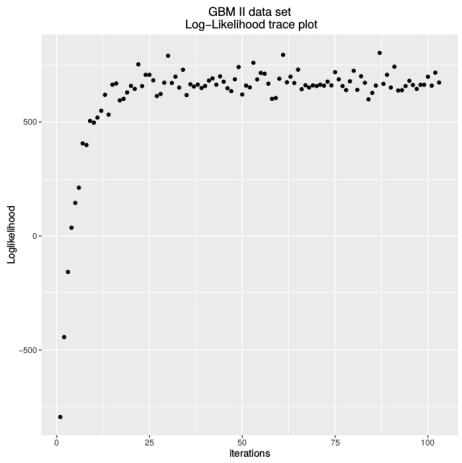


**Fig. 20.** Log-likelihood trace plots for the Glioblastoma II Set

| Data Set | Time (in minutes) | Iterations(Burn-In + Gibbs Samples) |
|---|---|---|
| Breast Cancer | 144 | (100 + 200) |
| GBM I | 83 | (100 + 200) |
| GBM II | 120 | (100+ 200) |

**Table 10.** Actual running times for SBC/iSBC on Real Data Sets

## 4 Running Times for SBC and iSBC

We present the actual running times of the SBC on the three data sets. We used 11.7 GB Intel 64 bit Xeon (R) 4x2.66 Ghz processor in Table 7.

## 5 Effect of Survival Data on SBC

### 5.1 Hierarchical Dirichlet Process Mixture Model

SBC distinguishes itself from traditional clustering algorithms for patient level microarray data by including clinical end-points as a very important source of information. In this section we explored the scenario when we ignore the clinical end-point information and use a very similar Hierarchical Dirichlet Process Mixture Model (hDPMM) (as shown graphically in Fig.23) to cluster the patients based on just their molecular profiles (gene expression). The difference between SBC and hDPMM is the absence of parameters for modelling the survival information in hDPMM. We can also make predictions from hDPMM in a very similar manner as described for SBC.
For the sake of comparability, we used the corresponding SBC signatures as the feature set for the Breast Cancer data set and Glioblastoma I data set. A survival model (penalized Cox PH) is then fitted on top the clustering obtained above to stratify the survival curves and to make predictions.

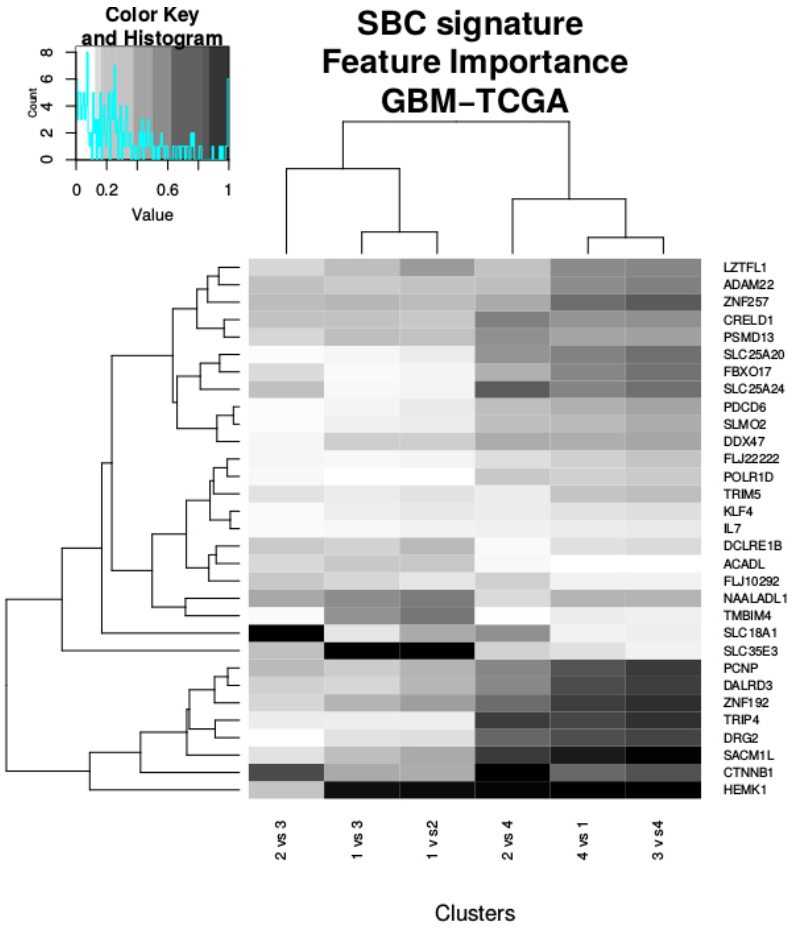


**Fig. 21.** Results on Glioblastoma II (iSBC): Feature importance of the SBC signature on TCGA-GBM gene expression in discriminating respective clusters

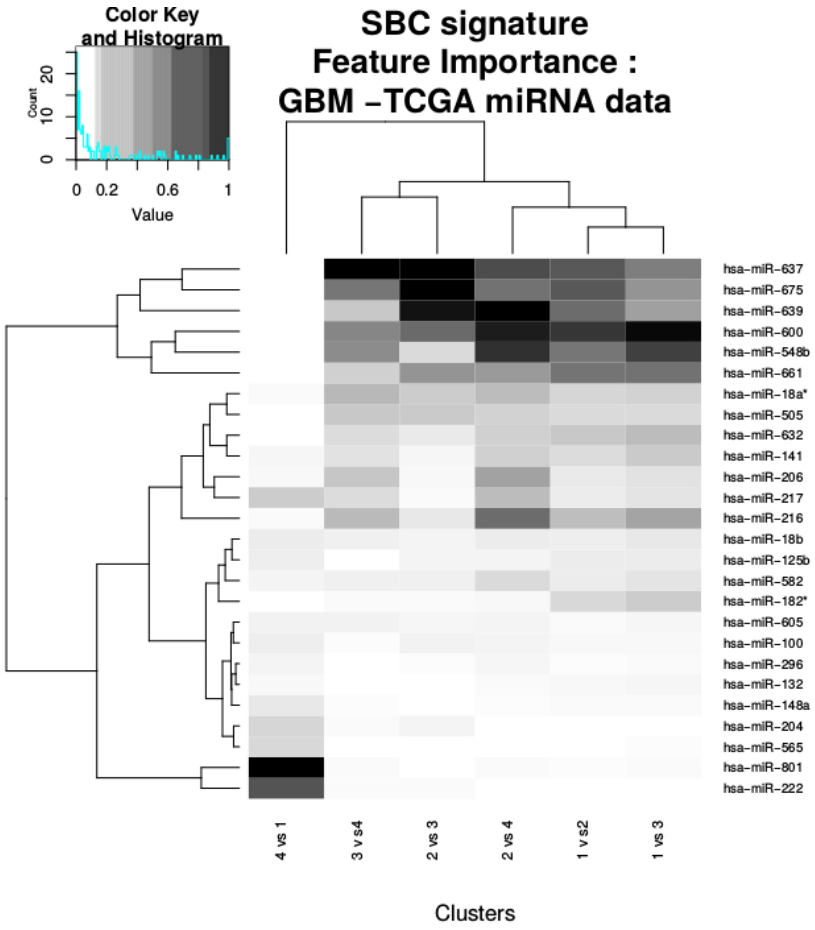


**Fig. 22.** Results on Glioblastoma II (iSBC):Feature importance of the SBC signature on TCGA-GBM mi-RNA expression in discriminating respective clusters

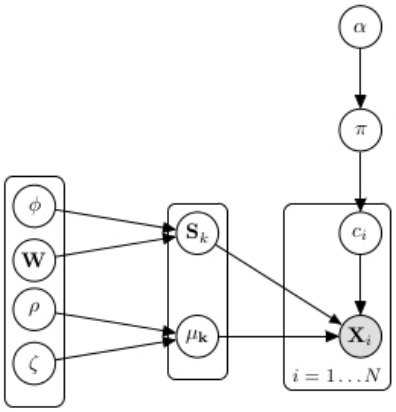


**Fig. 23.** Graphical Model representation for hDPMM

### 5.2 Breast Cancer Data Set

The hDPMM model on the breast cancer data set yields no clusters. The results are presented in Table 11 As can be seen, stratifying patients according to the SBC yields much better predictions for survival than using hDPMM. Thus we can conclude that survival information plays a vital role in obtaining the "Good prognosis" and "Bad prognosis" clusters which were obtained from our original SBC model.

| METHOD | FEATURE SET (SIGNATURE) | TRAINING (T) or PREDICTION (P) | p-value (Log Rank) | C-Index |
|---|---|---|---|---|
| SBC | SBC | T | **1.7e-08** | **0.79** |
| hDPMM | SBC | T | NA | 0.78 |
| SBC | SBC | P | **1.2e-03** | **0.70** |
| hDPMM | SBC | P | NA | 0.61 |

**Table 11.** Breast Cancer Data Set Results with hDPMM

### 5.3 Glioblastoma I Data Set

The application of the hDPMM model on the Gliobalstoma I data set yields 3 clusters, one which contains the majority of the data points. The three clusters contain 94,3 and 1 data points respectively for the training data set. For the prediction, the hDPMM places all the test points in one cluster. The results comparing hDPMM with SBC are presented in Table 12. Again, we can make a strong case that stratifying patients according to the SBC yields much better predictions for survival than using hDPMM and that the original 4 clusters obtained from the SBC model are heavily influenced by the clinical end-points.

| METHOD | FEATURE SET (SIGNATURE) | TRAINING (T) or PREDICTION (P) | p-value (Log Rank) | C-Index |
|---|---|---|---|---|
| SBC | SBC | T | **5.3e-05** | **0.68** |
| hDPMM | SBC | T | 0.90 | 0.73 |
| SBC | SBC | P | **3e-02** | **0.56** |
| hDPMM | SBC | P | NA | 0.50 |

**Table 12.** Glioblastoma I Data Set Results with hDPMM

# 6 Effect of CCA pre-processing on iSBC

## 6.1 Glioblastoma II Data Set

Here we report the results for CCA pre-processed iSBC on the example training-testing split for GBMII data. We used a non-penalized CCA approach to pre-process the Glioblastoma II Data set with the initial set of 31 mRNA and 31 miRNAs. We chose the top 10 Canonical Correlates and obtained the corresponding transformed data matrices for mRNA and miRNA each now containing 10 features each. The results on this transformed Glioblastoma II data set are shown in Table 13 which show marginal benefits on the prediction performance of the integrative SBC model using features derived from applying CCA to our SBC signature data sets. We still get four clusters as before with 57, 18, 15 and 6 data points respectively. Furthermore, we obtain feature importance for our new set of features (as shown in Fig. 26 and Fig.27) and plot the factor loading matrices in Fig.24 and Fig.25. The corresponding correlation values for the features are also shown.

| METHOD | FEATURE SET (SIGNATURE) | TRAINING (T) or PREDICTION (P) | p-value (Log Rank) | C-Index |
|---|---|---|---|---|
| iSBC | SBC | T | **6e-04** | **0.70** |
| iSBC | CCA | T | 1e-03 | 0.68 |
| iSBC | SBC | P | **1e-02** | **0.52** |
| iSBC | CCA | P | 1e-02 | 0.54 |

**Table 13.** Glioblastoma II Data Set Results with new feature sets derived from CCA

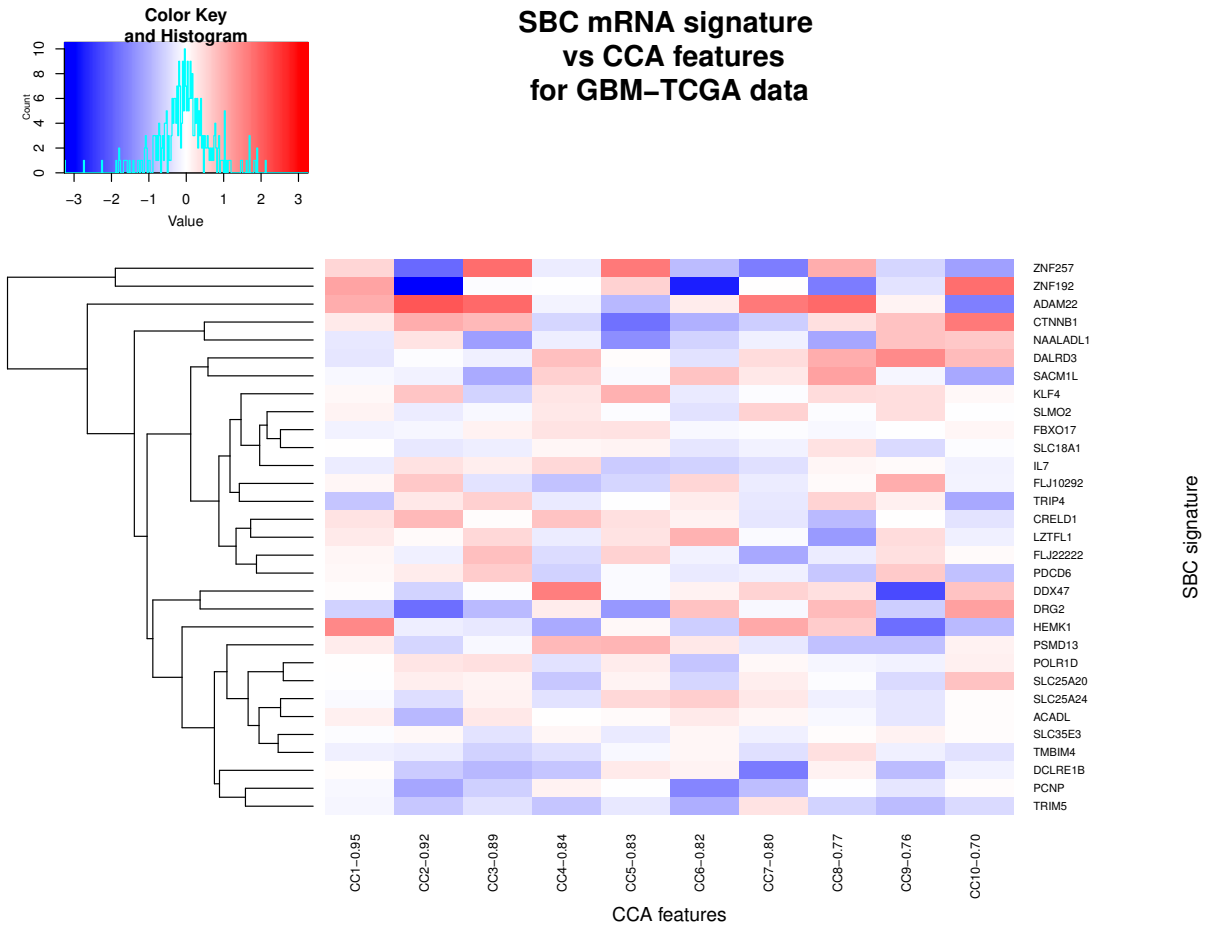


**Fig. 24.** Factor Loading Matrix between CCA features and the original SBC mRNA signature. Canonical covariates are named as CC1-xx to CC10-xx, where xx indicates the respective canonical correlation

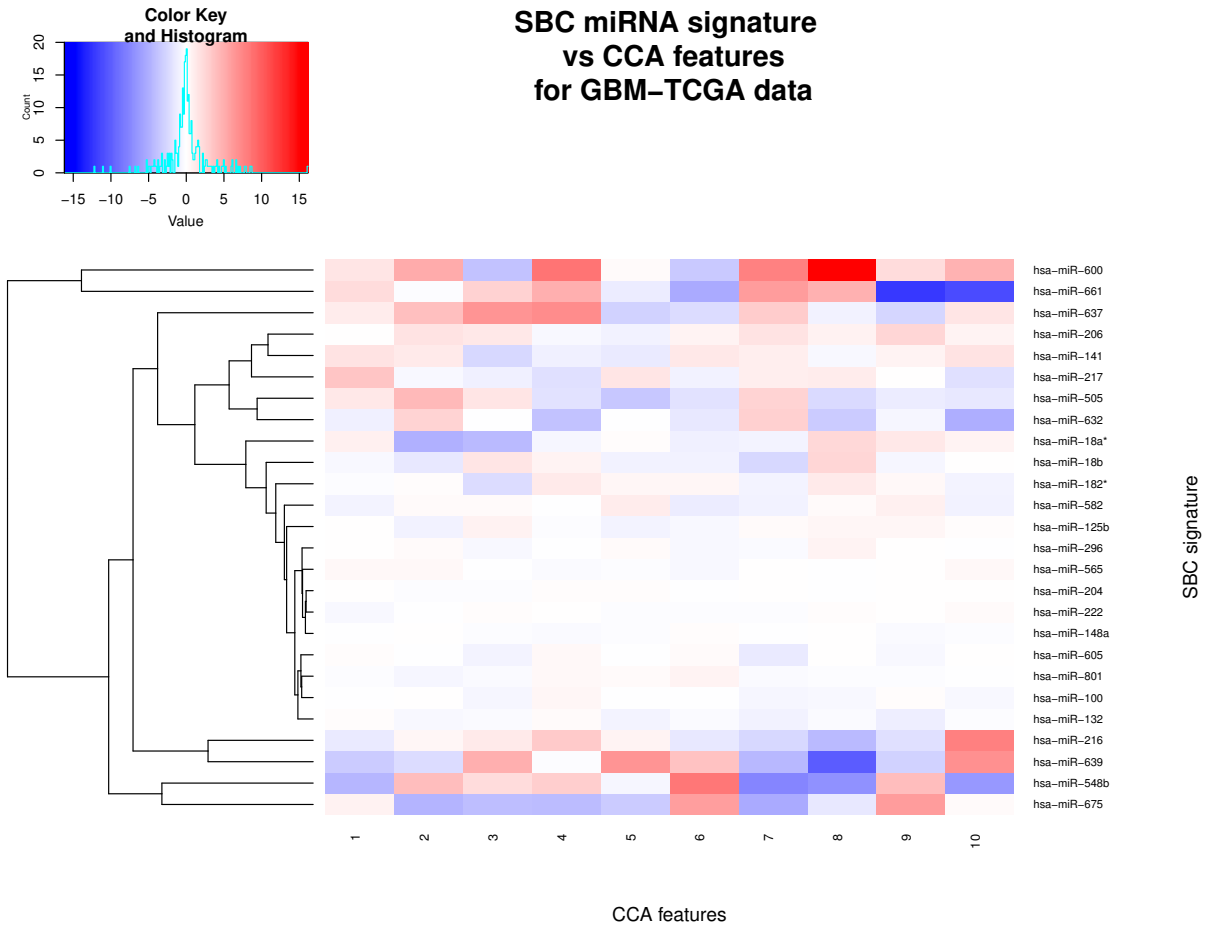


**Fig. 25.** Factor Loading Matrix between CCA features and the original SBC miRNA signature. Canonical covariates are named as CC1-xx to CC10-xx, where xx indicates the respective canonical correlation.

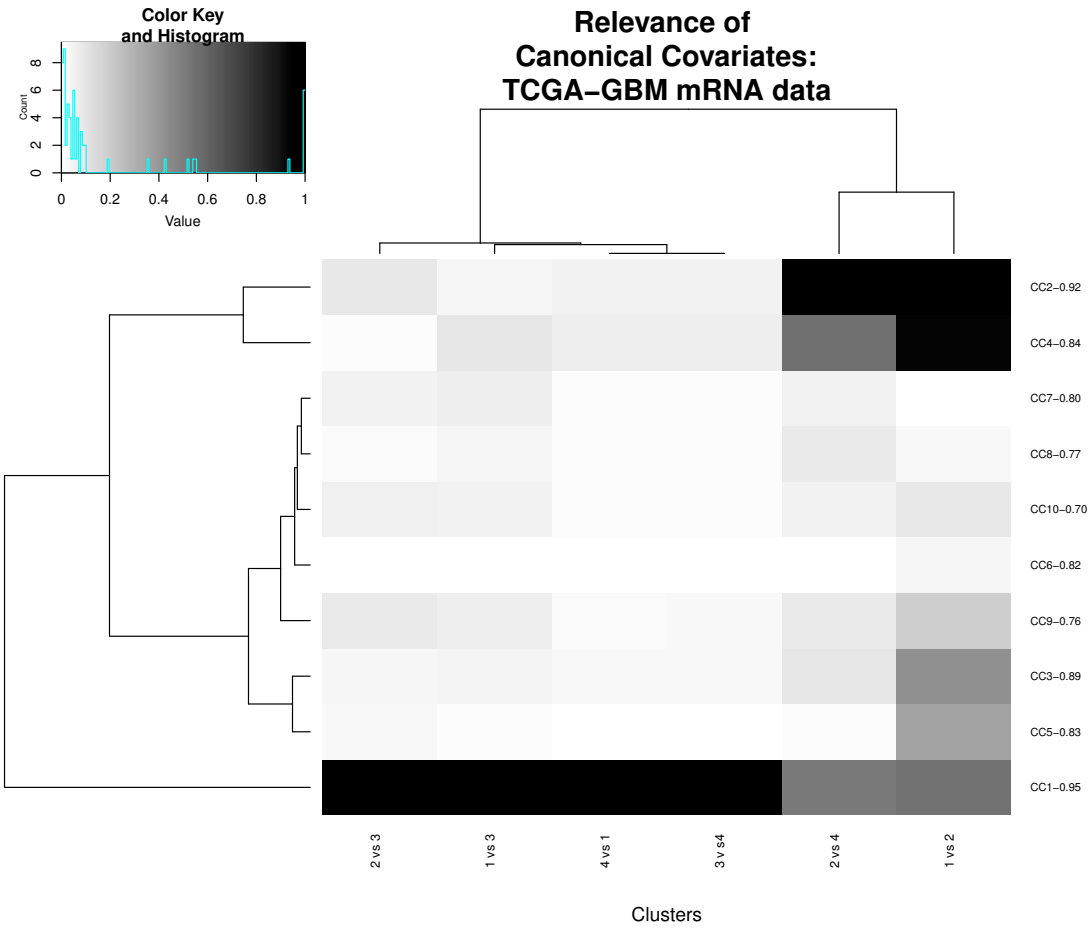


**Fig. 26.** Feature Importance of the new CCA features derived from the mRNA-SBC signature.Canonical covariates are named as CC1-xx to CC10-xx, where xx indicates the respective canonical correlation

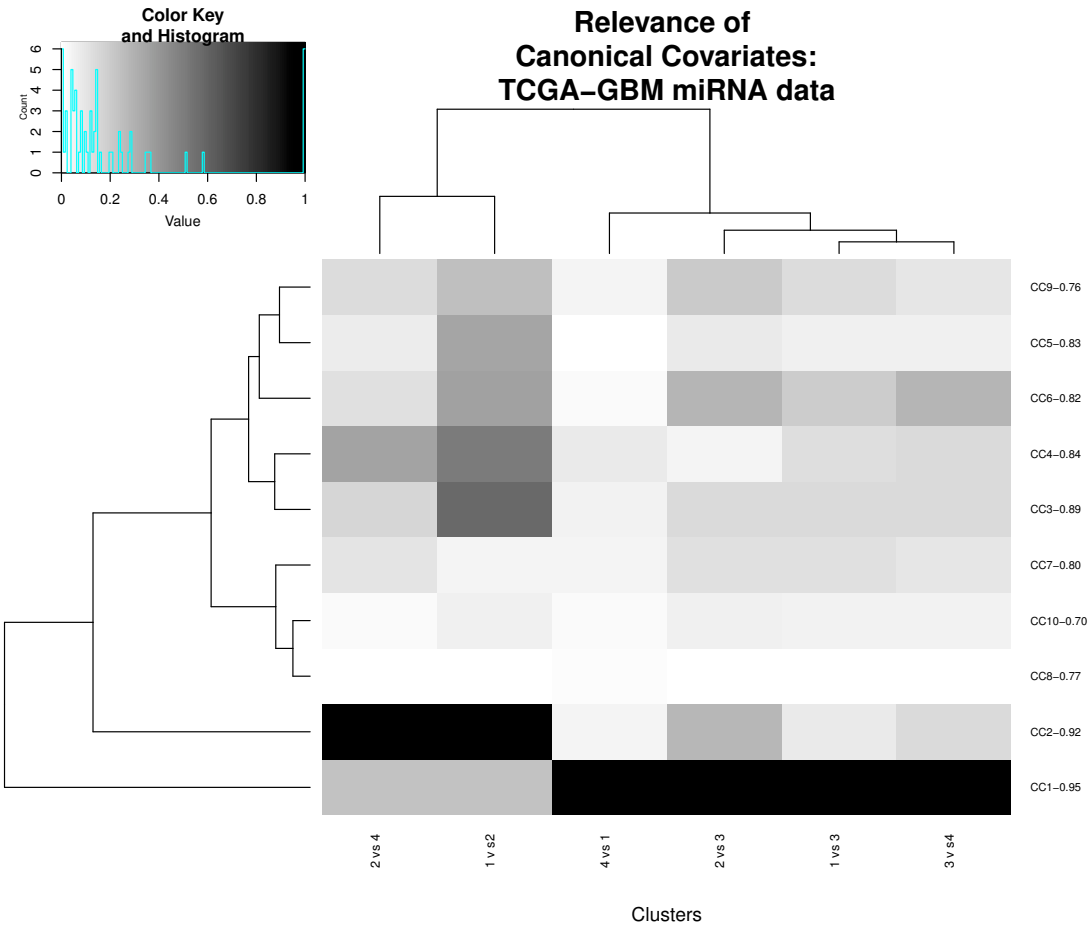


**Fig. 27.** Feature Importance of the new CCA features derived from the miRNA-SBC signature. Canonical covariates are named as CC1-xx to CC10-xx, where xx indicates the respective canonical correlation

# Bibliography


Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

D. Blackwell and J. B. MacQueen. Ferguson distributions via pólya urn schemes. *The annals of statistics*, pages 353–355, 1973.

S. Falcon and R. Gentleman. Using gostats to test gene lists for go term association. *Bioinformatics*, 23(2):257–258, 2007.

D. Görür and C. E. Rasmussen. Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664, 2010.

B. Grun and F. Leisch. Flexmix version 2: finite mixtures with concomitant variables and varying and constant parameters. 2008.

S. Liverani, D. I. Hastie, L. Azizi, M. Papathomas, and S. Richardson. Premium: an r package for profile regression mixture models using dirichlet processes. *arXiv preprint arXiv:1303.2836*, 2013.

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

V. Melnykov, W.-C. Chen, and R. Maitra. Mixsim: an r package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software*, 51(12):1–25, 2012.

R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Y. Ru, K. J. Kechris, B. Tabakoff, P. Hoffman, R. A. Radcliffe, R. Bowler, S. Mahaffey, S. Rossi, G. A. Calin, L. Bemis, et al. The multimir r package and database: integration of microrna–target interactions along with their disease and drug associations. *Nucleic acids research*, 42(17):e133–e133, 2014.

A. Waha, J. Felsberg, M. Simon, W. Hartmann, T. Pietsch, and A. Waha. A member of the slc25 family is epigenetically inactivated in human gliomas and suppresses cell growth in vitro. *Cancer Research*, 71(8 Supplement):4797–4797, 2011.

D. M. Witten and R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 2012.