Bonn-Aachen International Center for Information Technology

B-IT

Master thesis

External validation and characterisation of cancer subtypes using SBC

Author:
Camila Duitama González

First Examiner:
Prof. Dr. Holger Fröhlich

Second Examiner:
Prof. Dr. Martin Hofmann-Apitius

Advisor:
Dr. Ashar Ahmad

Submitted: October 8, 2019

Declaration of Authorship

I herewith certify that this material is my own work, that I used only those sources and resources referred to in the thesis, and that I have identified citations as such.

_____ _____

Location, Date Signature

Acknowledgments

I would like to thank my family. I owe them all I am and all I have, and it is through their nurturing and unconditional love that I thrive. I would also like to thank my friends, the old ones and the new ones, for their support and company during my time in Germany.

I am very thankful to Ashar for his patience and dedication during the most fruitful period of my masters. I deeply appreciate all the conversations we had, academic and non-academic, in which he was always respectful, encouraging and very attentive. It is often underestimated how important a supportive supervisor is in academic life. Finally, I would also like to thank Prof. Fröhlich for his support and guidance.

Abstract

The Survival Based Bayesian Clustering (SBC) model developed by Ahmad and Fröhlich (2017), infers clinically relevant cancer subtypes, by jointly clustering molecular data along with survival data. Originally, the model was tested on a a Breast Cancer (Van De Vijver et al., 2002) and a Glioblastoma Multiforme (GBM) (Verhaak et al., 2010) data set, without any further external validation. The objective of this master thesis was to perform an external validation of the SBC, a goal that entailed two major tasks: a rigorous feature engineering and selection process that improved the known predictive ability of the model, and the characterisation of the obtained clusters and corresponding signature by delving into other types of clinical and omics data such as Copy Number Variation and miRNA.

The TCGA-GBM data set was retrieved using the Bioconductor package RTCGAToolbox and after data preprocessing, appropriate normalisation and correction for sample selection bias, a combined patient cohort of 421 samples was obtained (160 patients for the training and 261 patients for the validation set). Various feature engineering and selection techniques were explored. Every SBC model fit was done using Gibbs sampling. The best feature engineering and selection approaches were the Block HSIC-Lasso model for mRNA-based selection and a Penalized Accelerated Failure Time model on a collection of oncogenic gene sets for pathway-based selection. In both cases there was an improvement of the initial Predictive C-Index (Block HSIC-Lasso feature selection = +1.5%, PAFT feature selection = +27.6%) and Recovery C-Index (Block HSIC-Lasso feature selection = +8.7%, PAFT feature selection = +5.0%).

The work done in this master thesis is a step forward in the validation of the SBC model on an external data set such as the TCGA-GBM patient cohort.

Contents

List of Figures

"*Ya no sabremos que sería de la vida sin las versatilidades del carbono, sin la afición de la molécula a copiarse, sin el alosterismo de ciertas proteínas. Qué sería de cada sol, de cada mundo, sin la improbable isotropía y corrimiento del espacio, si no fuese la materia a tal punto propensa a la espiral, propicia a la orbe, si fuera levemente otro el delicado contrapeso de sus órbitas..., y con cuántos avatares, y con qué albúres, estará a su vez esta página en secreta y simultánea resonancia.*"

— Carlos Framb, Un día en el paraíso.

1 Introduction

The background of this master thesis is the Survival Based Bayesian Clustering (SBC) (Ahmad and Fröhlich, 2017). This model infers clinically relevant cancer subtypes, by jointly clustering molecular data along with survival data. In the original paper, the emphasis was on cluster discovery along with clustering characterisation on a Breast Cancer Data Set(Van De Vijver et al., 2002) and a Glioblastoma Multiforme (GBM) Data set (Verhaak et al., 2010). GBM is the most common and aggressive primary central nervous system malignancy, and has a median survival of 15 months (Omuro and DeAngelis, 2013). It was also the first cancer type to be systematically studied by The Cancer Genome Atlas (TCGA), with the hope that a comprehensive catalogue of its molecular alterations, using multidimensional and high-resolution data sets, will be a crucial resource to understand the disease and develop better treatments (Brennan et al., 2013).

The major motivation of this master thesis is the further validation of the published SBC method on an external data set. In this project the focus was on Glioblastoma Multiforme, in particular, the SBC model was trained on samples from the Verhaak study with the intention to validate its findings on the TCGA-GBM cohort. The data collection and preprocessing, the model implementation (learning of all the model parameters and cluster memberships) as well as the evaluation and interpretation of the results through cluster characterisation will be presented in the coming chapters.

The code to this master thesis was made available in the following GitHub repository: `https://github.com/CamilaDuitama/MasterThesis`. Furthermore, a detailed outcome of the several models trained in order to achieve the final results presented in this master thesis are displayed in a more graphical website: `https://camiladuitama.github.io/MasterThesis/index.html`

In order facilitate the understanding of the following chapters, a brief introduction on the topics of precision medicine and the contributions of the SBC model to this field, as well as an overview of the current landscape of GBM in precision medicine will be presented.

1.1 Precision medicine and the contribution of the SBC to the field

Precision medicine, also known as personalised or stratified medicine, is defined by the NCI (2019) as a "form of medicine that uses information about a person's genes, proteins and environment to prevent, diagnose and treat disease". Precision medicine has the potential to reduce health care costs, increase quality of life and survival time of patients, facilitate earlier disease detection, reduce adverse effects, increase patient compliance and produce a significant shift towards prevention medicine over curative medicine (Mathur and Sutton, 2017).

More specifically, precision medicine in oncology takes molecular and cellular features of a tumour, as well as environmental conditions, to create custom-made treatments for the patients (Le Tourneau et al., 2019). Because cancer is a complex disease known for its heterogeneity (both in patients and tissue samples) (Kitano, 2002), and most of the current treatments and clinical trials do not take into account the diversity in patient population (Prados et al., 2015), *patient stratification* is presented as a strategy to classify patients into more homogeneous subgroups, using different sources of biological high-throughput data (omics data coming from genomic, transcriptomics, epigenomics, proteomics, metabolomics experiments, etc.), bio-imaging data, Electronic Healthcare Records (EHR), data from wearable sensors and mobile applications, among others (Fröhlich et al., 2018). In the face of this highly-dimensional and noisy data sets, patient stratification is usually made along with *biomarker discovery*. Biomarkers are defined characteristics *measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention*" (Group et al., 2016). Moreover, biomarker discovery in the context of patient stratification, can be understood as the selection of significant features (i.e. signature genes) that will help establish the distinction between patient subgroups, a procedure that is pivotal when working with high-dimensional data.

High-dimensional multi-modal data sets are particularly informative in the field of personalised medicine, because they capture and understand different dimensions of a patient and enhance the prediction performance of patient stratification algorithms (Fröhlich et al., 2018). In this context, multivariate statistical models, such as the SBC model, are ideal tools to analyse highly multi-dimensional data sets. SBC models multi-modal omics data along with clinical outcome data for the purpose of patient stratification(Ahmad, 2019), by inferring clinically relevant molecular subtypes of patients based on molecular data.

Initially the SBC achieved its original goal of discovering cluster of patients using

their molecular data and obtaining strong curve separability, on simulated data as well as on Breast Cancer Data Set(Van De Vijver et al., 2002) and a GBM data set (Verhaak et al., 2010). In the original paper validation was done by randomly dividing the real data sets into training and validation sets. However, no external validation of the SBC had ever been done before the elaboration of this master thesis project.

1.2 Glioblastoma Multiforme (GBM)

Gliomas, also known as primary brain tumours, are classified according to their cell of origin (Hanif et al., 2017) and are divided into two main subgroups: non-diffuse gliomas and diffuse gliomas. In the same manner, diffuse gliomas are categorised into astrocytomas (with glioblastoma as its most frequent and malignant representative), oligodendrogliomas or oligoastrocytomas (Wesseling and Capper, 2018), according to the histological information from glial cells, which are cels that provide support, isolation and surround neuronal cells. Additionally, GBM can be further classified according to the type of glial cell which is more dominant in the histological examination of the tumour sample (astrocytes, oligodendrocytes, and microglial cells (Purves et al., 2001)). Interestingly enough the Verhaak et al. (2010) classification is related to these histological subtypes.

Currently, the revised 2016 World Health Organization Classification of Tumor in the Central Nervous System (2016 CNS WHO) considers both histological and molecular information in its novel reclassification of gliomas, and it has stated that GBM are to be divided into three main categories (Louis et al., 2016):

- IDH-wildtype which accounts for 90% of the cases and is predominant among patients who are 55 years old or more.
- IDH-mutant which accounts for 10% of the cases and is more prevalent among younger patients.
- Glioblastoma NOS, or 'not otherwise specified' for cases where the molecular testing could not be performed or the results were inconclusive.

Not only are all of the previously mentioned GBM subtypes classified by the WHO as having a level of malignancy IV (I being the lowest level and IV the highest), but GBM also has a median survival of 14-15 months after diagnosis and accounts for 50% of all gliomas in all age groups (Hanif et al., 2017). Effective treatment for this diagnosis is still unavailable, and tumour heterogeneity, biological complexity and the difficulties in drug delivery are present challenges in the management of the disease (Prados et al., 2015).

In 2010 Verhaak et al. (2010) reported the molecular classification of GBM tumours based on gene expression profiles into four subtypes with unique Copy Number and Somatic Mutation alterations:**Proneural, Neural, Classical, and Mesenchymal**. Because Glioblastoma Multiforme (GBM) is one of the most heterogeneous tumours, it is not likely that one treatment will serve for all patients diagnosed with the disease, and this is why the genotypic profiling done by Verhaak et al. (2010) and other patient stratification techniques such as the SBC represent significant advances in the field of personalised medicine in GBM.

1.3 Motivation

It is worth emphasising that before this master thesis was done the SBC had only been evaluated on real cancer data sets which have been widely characterised in the literature (Van De Vijver et al., 2002; Verhaak et al., 2010). Additionally, performing an external validation of the SBC on historical data such as the Verhaak samples, was a reasonable decision to make, in order verify the reproducibility of the SBC model trained on the Verhaak labels in a different cohort of TCGA-GBM samples and facilitate cluster interpretability. The validation of any statistical model in an external data set is a key step in Translation Medicine, where recommendations based on the model are ultimately translated into clinical practice. Broadly speaking, statistical models need to be internally validated (as it was done by Ahmad and Fröhlich, 2017 on their paper), retrospectively validated on an external data set (as in done on this master thesis) and finally, prospectively validated in early clinical trials to be able to make it through to clinical practice. A highly cited and successful example of a model that went through the previously described steps is MamaPrint^{TM}(Van De Vijver et al., 2002), a 70-gene-signature-based prognostic test approved for clinical use by the FDA in 2007, which is able to select patients who would benefit from adjuvant chemotherapy and accurately predict the development of distant metastases. In this context, the work done in this project is vital in the effort to develop a diagnostic tool for clinical practice, that is aligned with the goals precision medicine and personalised care.

Furthermore, clustering and predictive models such as the SBC, are sensitive to the number of input features they are trained on, as increasing number of covariates might add noise to them and result in faulty model fit and reduced predictive ability. It is for this reason, that selecting a relevant-to-survival and reduced set of input features, known in the machine learning jargon as *feature selection*, is pivotal for the performance of the SBC. Performing feature selection on the large set of features (genes) that comes with a typical multidimensional mRNA expression

matrix, can be achieved by analysing the connection between the expression of genes with a clinical response (e.g. survival time), with the use of additional knowledge available such as pathway information. Moreover, using pathway or gene set data reduces the complexity of analysis, improves model interpretability and makes use of the vast amount of biological knowledge gathered by experts in the field for many years.

To summarise, the main motivations behind this master thesis project were the external validation of the SBC model and a feature engineering or selection approach that could possibly improve the known ability of the model to discover clinically relevant cancer subtypes.

1.4 Objectives

- Perform an external validation of the SBC based clusters on the Verhaak et al. (2010) study ensuring distinct survival curves for the training and validation set clusters, using an external data set from patients diagnosed with Glioblastoma Multiforme.
- Characterise the obtained clusters and the corresponding SBC signature by looking at other data types for consistent patterns such as Copy Number Variation data, miRNA data and clinical data.

2 State of the Art

2.1 Data: The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas is a large pan-cancer genomics programme established by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). It has generated over 2.5 petabytes of omics data, with the aim to catalogue and discover alterations found in different types of high-throughput data sets that can be associated to cancer through multi-dimensional analyses. It has large cohorts of over 33 types of cancer, with data ranging from gene expression, exon expression, miRNA expression, Copy Number Variation, etc.

Data obtained from the TCGA is categorised by data levels in the following manner :

- **Level I:** Raw, non-normalised data.
- **Level II:** Processed data.
- **Level III:** Segmented/interpreted data applied to individual samples.
- **Level IV:** Summarised data.

Only data of levels III and IV is freely available from publicly accessible databases without extra permissions (Tomczak et al., 2015).

2.1.1 Accessing to the TCGA Data

Due to the variability and granularity of the data sets offered by the TCGA, accessing the data has been a challenge to the research community and several tools to retrieve molecular data sets are available (Colaprico et al., 2015)

Furthermore the TCGA data processing by the Broad Institute formally ended in July of 2016, which is why the content migrated to the Genomics Data Commons (GDC) Data Portal hosted by the NCI (Mounir et al., 2019)

Genomics Data Commons (GDC)

This is an information system for storing, analysing and sharing genomic and clinical data from patients with cancer maintained by the NCI. It currently hosts several large genomic projects, including TCGA, Therapeutically Applicable Research to Generate Effective Treatments (TARGET) and the Cancer Cell Line Encyclopedia, among others (Jensen et al., 2017). The GDC provides access to multiple cancer data sets, including the TCGA, making it available via two data portals: *Harmonized* and *Legacy*. These two are described as follows:

- **Harmonized:** This data has been fully harmonized using GRCh38 (hg38) as a reference genome. By harmonized, the GDC refers to data that has gone through a process of standardisation with common pipelines that use open source sequence analysis tools, which were developed with the help of experts in cancer genomics and that are regularly evaluated and updated.

- **Legacy:** This is an unmodified collection of data that used GRCh36 (hg18) and GRCh37 (hg19) as genome reference assemblies. some previously available data types and formats are not currently available on the harmonized data and are only distributed via the GDC Legacy Archive. For instance, Methylation Data and mRNA microArray Data for the TCGA, can only be accessed through the Legacy portal.

Firebrowse

This is a web-based tool to explore and visualise cancer data generated by the Broad GDAC Firehose from the Broad Institute. The latter is a set of tools and pipelines for preprocessing and analysing various types of genomic and proteomic data (N. C. I. NCI, March 6 of 2019). Firehose provides preprocessed data to the research community, however it doesn't provide systematic access to the data, and it is not easily integrated with programming environments for post analysis.

It is important to note that the Firebrowse only serves hg18 legacy data, and that the Broad Institute no longer processes this data.

RTCGAToolbox

RTCGAToolbox (Samur, 2014) is an R based package to systematically access TCGA preprocessed data from the Firehose and to organise it for easy management and analysis.

TCGABiolinks

This is the newest of the tools previously mentioned. It systematically accesses Firehose preprocessed data and performs basic analysis and visualisation using R. TCGABiolinks integrates other existing analysis tools also written in R, allows to download older versions and provides a 'SummarizedExperiment' object, which is necessary for integration and use of other popular Bioconductor packages (Colaprico et al., 2015).

Once the TCGA data set migrated to the GDC Data Portal, this tool was updated to query the GDC portal instead.

2.1.2 TCGA Data types

Clinical Data

According to the NCI there are three subtypes of clinical data fot the TCGA dataset:

- **Clinical data:** This refers to demographic information, treatment information, and survival data among others.
- **Biospecimen data:** This contains information on how samples were processed by the Biospecimen Core Resource Centre
- **Pathology reports:** PDF files only available for certain cases.

mRNA Microarray Data

Microarray-based gene expression is a measure of the concentration of different mRNAs in a tissue. Microarray technologies work by hybridising a sample of unknown sequence to an array of immobilised DNA molecules whose sequence is known (Gershon, 2002), and were the most popular high-throughput gene expression profiling technology for several decades, until RNA-Seq technology was popularised.

TCGA provides access to normalised expression values per probe, gene or exons,using standardised software pipelines; where expression values are provided in simple tab-separated value format. For the GBM data set, there are mRNA expression files from two different platforms: HT_HG-U133A and AgilentG4502A_07 (Guo et al., 2013). The former belonged to Affymetrix (now owned by Thermofisher) and the latter to Agilent Techonolgies, but the first one had more samples which was why this data set was selected.

miRNA MicroArray Data

MicroRNA (miRNA) is an non-coding small RNA, which regulates gene expression either by binding to the 3' URT and therefore inhibiting translation or by degrading specific mRNAs. Furthermore, some miRNAs act as tumour suppressor genes, and they have been linked to a variety of cancers, including GBM (Srinivasan et al., 2011). miRNA expression data in the TCGA is Array-based and comes from the Agilent Human 8x15K miRNA Platform. All the other types of cancer present in the TCGA have miRNA sequencing of tumour samples.

Copy Number Data

Copy Number Data is the result of Copy Number Variation (CNV), which is the unbalanced rearrangement that increases or decreases the DNA content. This form of genome variability among individuals is composed of regions larger than 50 bp (Zarrei et al., 2015). CNV are known to be related to tumour formation and progression (Freire et al., 2008), and genes affected by CNV play an important role in oncogenesis and cancer therapy (Santarius et al., 2010).

The algorithm to analyse Copy Number Data used by both the Firebrowse and GDC is called Genomic Identification of Significant Targets in Cancer (GISTIC), which identifies *driver* CNVs by evaluating the amplitude and frequency of observed events (Mermel et al., 2011).

2.2 Statistical model: Survival Based Bayesian Clustering (SBC)

Patient stratification is an important step in personalised medicine, and has been approached in a *supervised* and *unsupervised* manner, allowing patients to be grouped into more homogeneous groups with the use of gene expression data and other kinds of omics data sets. Supervised methods involve learning models with given class labels, for example, based on a certain criterion one can divide the survival of patients into good and bad survivors. Here the goal of learning is to identify the parameters of a predictive model which allows stratification of unseen patients into exactly that: good and bad survivors. On the other hand, unsupervised methods in patient stratification model the molecular data and estimate the parameters in modelling the molecular data (for example using clustering). This modelling then uses post-hoc analysis with clinical data to establish clinical significance of the stratification. In the SBC Ahmad and Fröhlich (2017) combine the above two approaches in a sort of *semi-supervised* way: simultaneously clustering

multivariate molecular data and survival data in order to discover clinically relevant disease subtypes in a semi-supervised manner. Furthermore, the SBC model has the following unique features(Ahmad and Fröhlich, 2017):

- Automated and fully Bayesian treatment of the number of clusters.
- Ranking of most discriminatory features.
- Integration of more than one omic data type.
- Prediction of class membership and survival outcomes for patients on an independent test data.

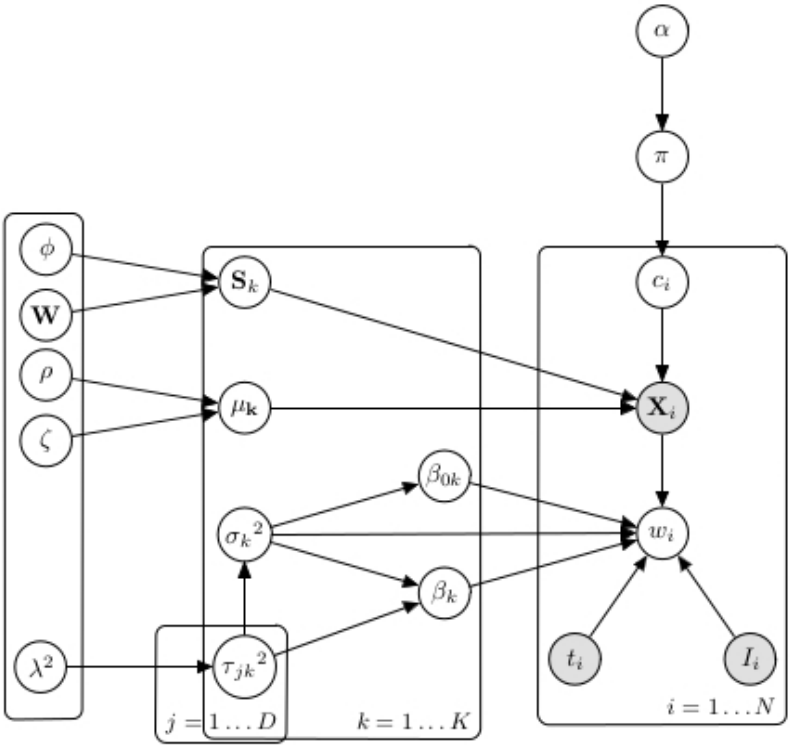


Figure 2.1: Graphical model representation for the SBC.

The overall parameters to model the joint distribution of molecular data and survival times are assumed to be drawn from Dirichlet Process Mixture Models prior. Individually, the molecular data is modelled using a Hierarchical Gaussian Model, survival data is modelled using a Penalized Accelerated Failure Time

model. Model fitting is done using Gibbs sampling. Model predictions and extraction of feature importance from the SBC if further described. A graphical representation of the model can be seen in Figure 2.1.

To understand the model structure and foundations, some concepts will be clarified in the following sections.

2.2.1 Dirichlet Process Mixture Models (DPMM)

Many probability models assume that the data obtained from an experiment comes from a collection of random variables $X_1, X_2, ...X_n$ independently drawn from some underlying probability distribution G with a probability density function g. Models that are described through a vector v of a finite number of real values are referred to as *parametric models*. However, in biology is often the case that sample size n is small (e.g. number of patients whose mRNA microarray expression data has been obtained). In these cases, fixing the number of parameters to a certain set fails to capture the true underlying probability distribution and hence there is a need to define a class of models where the vector v is assumed to be infinite, and based on the data, its finite dimensional estimate is obtained. Such probabilistic models are referred to as *non-parametric* (Ahmad, 2019).

A Dirichlet Process (DP) is a widely used random probability measure to model uncertainty about the functional form of the distribution of parameters in a model,this is, to be used as a prior distribution. DP are defined by two parameters: A positive scalar called concentration parameter α and a probability measure G_0, referred to as the base measure. G_0 is the prior guess, and α is the strength of belief in G_0 (Görür and Rasmussen, 2010).

DPMM are *non-parametric models*, that allow for the inference of possibly infinite number of clusters K. As any other method that is based on Bayesian inference, DPMM require assigning a prior distribution to all parameters in the model, which is the Dirichlet Process.

DPMM models the distribution from which X_i is being drawn as a mixture of distributions of the from $F(\theta)$. This gives the following model(Neal, 2000):

$$\begin{aligned} X_i \mid \theta_i &\sim F(\theta_i) \\ \theta_i \mid G &\sim G \\ G \mid & \; DP(G_0, \alpha) \end{aligned} \tag{2.1}$$

If i is imagined as the last of N observations, and $\delta(\theta)$ is the distribution concentrated at the single point θ, one can sample from the conditional prior in the

following manner (Neal, 2000):

$$\theta_i \mid \theta_1, ...\theta_{i-1} \sim \frac{1}{N-1+\alpha} \sum_{j \neq 1}^{i-1} \delta(\theta_j) + \frac{\alpha}{i-1+\alpha} G_0 \tag{2.2}$$

2.2.2 Hierarchical Multivariate Gaussian model

As a choice for the base distribution G_0 Ahmad and Fröhlich (2017) use a hierarchical Gaussian Model that is described by the following set of equations:

$$\begin{aligned} X_i \mid (c_i = j) &\sim \mathcal{N}(\mu_j,\, S_j^{-1}) \\ (\mu_j \mid S_j, \xi, \rho) &\sim \mathcal{N}(\xi,\, (S_j \rho)^{-1}) \end{aligned} \tag{2.3}$$

Where X_i indicates a D-dimensional vector of measurements (e.g. gene expression profiles) for patient i, 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.

2.2.3 Bayesian Lasso penalised AFT model

The Accelerated Time Failure (AFT) model is given by the following equation:

$$\log(t_i) = \beta_0 + \beta^T X_i + \epsilon_i, i = 1, 2, ...N \tag{2.4}$$

In the aforementioned equation $\log(t_i)$ is the log survival time, β is the vector of regression parameters and $_i$ is the matrix of gene expression data. Ahmad and Fröhlich (2017) place a Laplacian prior over β which effectively induces a L1 penalty on the regression coefficients and penalises small effects to exact zero.

2.2.4 Gibbs Sampling

MCMC methods are computational techniques designed to generate samples from a given probability distribution $P(\theta)$ (also called *target density*), and/or to estimate expectations of functions under this distribution. When $P(\theta)$ is not from a simple analytical form, MCMC methods are ideal to overcome the impossibility to evaluate these expectations by exact methods (MacKay and Mac Kay, 2003). Gibbs sampling is a MCMC method which assumes that even though $P(\theta)$ is too complex to draw samples from it directly, its conditional distributions $P(\theta_i \mid \{\theta_i\}_{i \neq j})$ might be tractable (of a simple analytical form). The idea in Gibbs

sampling is to generate posterior samples sweeping through each variable and fixing the remaining variables to their current values, this means, the sampling is not done on $P(\theta)$ itself, but samples are simulated by going through all the posterior conditionals, one random variable at a time (Yildirim, 2012).

In the SBC, Gibbs sampling is used for parameter estimation. The cluster indicator variables c_i are updated using the following conditional distribution for those components which have non-zero elements:

$$\begin{aligned} p(c_i = j \mid c_{-i}, \mu_j, X_j^{-1}, \beta_{oj}, \beta_j, \sigma_j^2, \alpha) \\ \propto \frac{n_{-i,j}}{N-1+\alpha} \mathcal{N}(w_i \mid \beta oj + \beta_i^T X_i, \sigma_j^2) \mathcal{N}(X_i \mid \mu_j, S_j^{-1}) \end{aligned} \tag{2.5}$$

On the other hand, the conditional distribution of a data point to belong to a new cluster is:

$$\begin{aligned} p(c_i \neq c_j \forall j \mid c_{-i}, \mu, S, \beta_0, \beta, \sigma^2, \alpha) \\ \propto \frac{\alpha}{N-1+\alpha} \\ \int \mathcal{N}(w_i \mid \beta 0 + \beta^T X_i, \sigma^2) \mathcal{N}(X_i \mid \mu, S^{-1}) dG_(\mu, S, \beta_0, \beta, \sigma^2) \end{aligned} \tag{2.6}$$

Equation 2.6 cannot be analytically obtained. See Ahmad and Fröhlich (2017) for details on the use of auxiliary variables to sample for the density function given.

2.2.5 Model predictions

Once the SBC model is trained, with m MCMC samples and molecular data X^*a test patient, *survival prediction* is a weighted average over the predicted survival times from each cluster:

$$\mathbb{E}[log(t^*) \mid X^*, \theta_{1:N}^{(m)}, c_{1:N}^{(m)}] \approx \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{C_m} (\beta_{0jm} + \beta_{jm}^T X^*) * v_{jm}(X^*) \tag{2.7}$$

C_m is the number of clusters found in the MCMC sample m, β_{0jm} and β_{jm} are the regression parameters of the AFT model for the sample m, and $v_{jm}(X^*)$ is the weight that accounts for the likelihood of X^* of belonging to cluster j for the sample m.

On the other hand, *cluster membership* is assigned by using the following conditional probability as the probability of a new data point X^* of belonging to the

previously discovered clusters for the m^{th} MCMC sample $c^{(m)} = 1, ..., C_m$:

$$p(c^* = j \mid X^*, \theta_{1:N}^{(m)}, c_{1:N}^{(m)}) = b\frac{n_{jm}}{N - 1 + \alpha}\mathcal{N}(X^* \mid \mu_{jm}, S_{jm}^{-1}) \tag{2.8}$$

In this expression, n_{jm} are the number of patients in cluster j and μ_{jm}, S_{jm} are the cluster parameters of the hierarchical Multivariate Gaussian model for the m MCMC sample, while b is a normalisation constant (Ahmad, 2019).

2.2.6 Feature importance

The SBC defines a ranking on the discriminatory ability of each feature by doing pairwise comparisons between clusters a and b. The ranking r^i of a feature i is defined as (Ahmad, 2019):

$$r^i = \frac{\mu_a^i - \mu_b^i}{\omega^i} \tag{2.9}$$

μ_k^i is the i^{th} component of the mean vector μ for cluster and ω^i is the i^{th} diagonal element of the matrix W, which is a hyperprior in the form of a diagonal matrix from the Dirichlet Process Mixture Models

3 Methodology

3.1 Data preparation

3.1.1 Data obtention

The goal of obtaining the data from the TCGA study was to include two sets of patients: the older Verhaak samples and the newer TCGA-GBM cohort samples. As the original SBC publication looked into microarray mRNA expression, for this master project only patients for whom this kind of data was primarily available were preferred. Subsequently, the largest sample size that had available mRNA microarray, Copy Number Variation, miRNA microarray and clinical data was selected. The package *(RTCGAToolbox)* (Samur, 2014) was used to systematically access the TCGA-GBM Legacy files from the 2016-01-2018 batch. Despite being aware that *(RTCGAToolbox)* queries the static unmodified collection of data from the Broad GDAC Firehose, which is now currently being handled and updated by the Genomics Data Commons, this package was preferred over the others for the following reasons:

- Simplicity to use.
- It can be systematically accessed using R, which is the programming language in which the SBC was implemented, so using this package facilitated the merging of different parts of the project.
- It offered access to the Legacy TCGA data set, which is the only data set with available mRNA microarray data that included the Verhaak samples.
- It included the largest sample size with the most diverse data sets (CNV,Clinical, miRNA, etc.)
- The data sets it accesses were presented in the most convenient format after the standardised preprocessing all samples were subjected to by the Broad Institute pipelines, a format that would facilitate the succeeding cluster validation and interpretation.

The code used to download the data is available here: `https://github.com/CamilaDuitama/MasterThesis/blob/master/DownloadTCGA.R`

3.1.2 Data processing and sample selection

The code for this section is available in the following Jupyter Notebook: `https://github.com/CamilaDuitama/MasterThesis/blob/master/Data%20preparation-RTCGA.ipynb`. Data processing consisted of:

1. Unification of patient and genomic trait identifiers. Each TCGA data sample is identified by a barcode that describes characteristics such as the project name, Tissue Source Site (TSS), participant who donated the sample, sample type, vial or order of sample in a sequence of samples, portion, plate and analyte (see GDC Documentation for more details). Some of the data matrices accessed in the data obtention part had longer barcodes than others because they retained more specific information regarding the data source. In order to facilitate manipulation all the barcodes were limited to the project, TSS and participant identifier (a barcode of 12 characters in total, see Figure 3.1)

2. Verification of data completeness and adequate format. This required looking for missing data, checking if the data frame shape was correct and examining the distribution of the different clinical variables and omics features among patients.

3. Inspection of Verhaak samples and Verhaak signature genes in the TCGA-GBM data set. This required finding the TCGA barcode for the Verhaak samples and contrasting it with the accessed data sets. Furthermore, an estimation of the number of data types available per patient had to be calculated.

4. Initial visualisation through PCA plots, histograms and boxplots.

5. Sample splitting into training and validation set. As it was previously mentioned, the training set was constituted by the available Verhaak samples in the TCGA-GBM, and the rest of the samples were used for validation.

6. Correction for sample selection bias. The default assumption in many learning scenarios is that the training and validation sets are independently and identically drawn from the same distribution, but when the distributions on this sets do not match there is a *sample selection bias* or *covariate shift*, which occurs often with the gene expression profile studies that use mRNA

microarray data (Huang et al., 2007), such as the the TCGA-GBM. One approach to try to control for this bias in the data processing and sample selection part, was by evaluating each of the clinical variables available for the samples, and assessing if there was a difference between training and validation sets evident in a boxplot. After this, only the Karnofsky Performance Score presented considerable differences between the sets and therefore certain members of the validation set were excluded, as it will be later explained. Another measure taken to try to control for sample selection bias was by scaling the data sets.

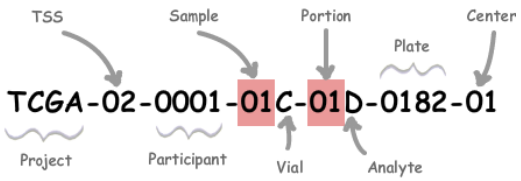


Figure 3.1: Example of a TCGA barcode with the highest number of identifiers Image taken from https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/

A total of 496 patients had mRNA microarray, miRNA microarray, Copy Number Variation and Clinical data available. As the original SBC publication used the Verhaak samples (160) as training samples, they were also taken as training samples for the master thesis. The others (336) we treated as validation cohort. However, the Karnofsky Index, an important clinical feature associated with survival was on average higher in the training than in the validation set. Therefore, patients with a Karnofsky Index lower than 70 were removed from the validation set (see Figure 3.2). The set presented in Table 3.1 was used for the subsequent steps of feature processing, model implementation and interpretation.

As mentioned, sample selection bias occurs when the distributions of the training and validations sets do not match. This is a problematic situation because it breaks the assumption that states that both the training and validation sets come from the same distribution. In order to control for this, different scaling techniques such as moment matching, ComBat (W. E. Johnson et al., 2007) and Z-score independent scaling were evaluated on the mRNA microarray data before proceeding with the data modelling with the SBC, by doing a PCA plot inspection of the gene expression data with the training and validation labels before and after every scaling procedure. The most successful method was Z-score independent scaling of each of the sets, which means, this method gave the most homogeneous distribution of the patient samples in the PCA plot, where training and validation clusters were almost indistinguishable. Figure 3.3 shows the PCA plot before and

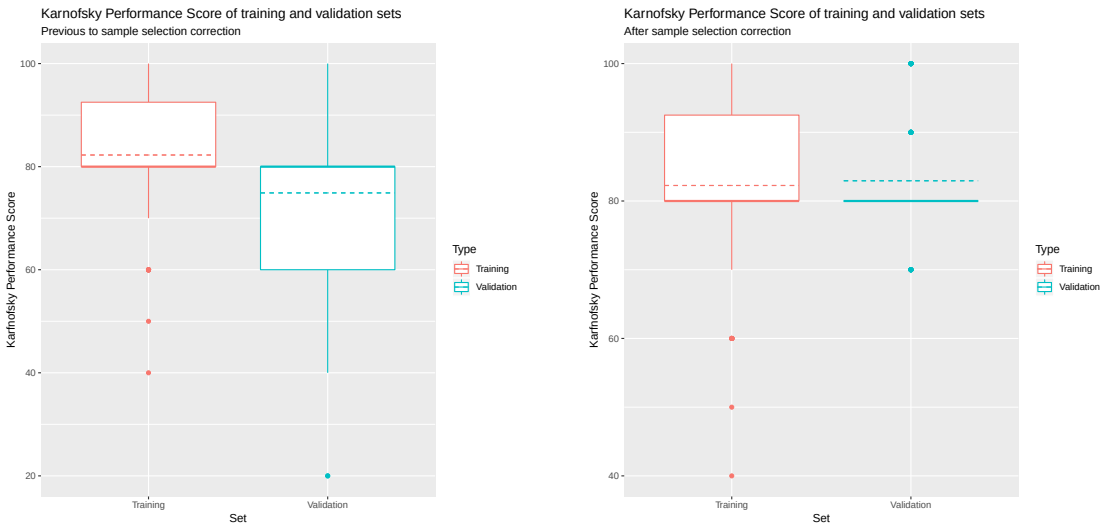


Figure 3.2: Boxplots of the clinical variable Karnofsky Performance Score for the training and validation sets before and after performing the sample selection correction on the validation set. The dashed lines correspond to the mean value of the Karnofsky Performance Score on each of the sets

after such scaling method. As it can be seen, before Z-score scaling a certain portion of the validation samples (blue dots) was segregated in the left hand side of the plot, and after Z-score scaling, the PCA plot shows a more homogeneous distribution of training and validation samples, where they aren't part of two clearly distinguishable clusters. This serves as evidence to say that the input data for the section will not have a very different distribution between train and validation sets.

Type	Sample size	Average Karnofsky Index	Average age (years)	♂/♀ [1]	Censoring [2]
Training set	160	80.00	56.0 ± 14.7	1.58	$\frac{6}{154}$
Validation set	261	80.00	58.8 ± 14.8	1.56	$\frac{48}{213}$

Table 3.1: Final sample selection and summary of main clinical features

[1] Proportion of male over female patients
[2] Proportion of patients who survived over patients who died

Figure 3.3: Effect of Z-score independent scaling on the mRNA microarray data visible on the PCA plot with training and validations labels. These are the PCA plot of the gene expression data for the training and validation sets before and after Z-score scaling of each set. The data visible in the plot on the right was later used in the procedure described in section 3.2

3.2 Data modelling using SBC

For every model trained, the sample splitting described on the Table 3.1 was preserved, taking as input data for the model both the years to death (or days to last follow-up if the patient survived) as the clinical end-point, and the mRNA expression data. Initially, the model was trained only on the TCGA-GBM mRNA expression matrix, i.e., no further feature engineering or selection (apart from the SBC signature calculation embedded in the model) was performed on the data [3]. The rationale behind this was:

- To have an initial idea of the performance of the model if the pipeline described by Ahmad and Fröhlich (2017) in their paper and their published code were used.
- To have a point of reference in the comparison of the results obtained before and after the feature selection and engineering process evaluated in this masters (additional to the SBC signature calculation embedded in the model),

[3] It is important to emphasise that on this model run as well as on all the other models trained in this master thesis, the sample splitting is described by Table 3.1 after a correction of the Karnofsky Index on the training patients, the data sets were independently scaled with Z-score scaling and the model parameters were initialised with K-means clustering

and assess the contribution or effect of this methods in the predictive performance of the SBC.

Initialisation is fundamental for Bayesian models such as the SBC, mainly because the results you get with this methods are highly sensitive to the initial set of parameters set up before the training phase begins. Because it is possible that the Gibbs sampler gets trapped in local minima, a careful selection of a good initial set of parameters might avoid this situation. In this project, K-means and Flexmix clustering were tested as initialisation methods for the SBC parameters. The tests to evaluate which initialisation method gave the best results were done on the Karnosfky corrected and independently scaled TCGA-GBM data sets as described in Section 3.1.1 (see Table 3.1), without any additional feature engineering and selection methods. K-means initialisation gave a higher predictive performance and improved survival curve separation. This process and its intermediate results were kept out of the written thesis in order to facilitate interpretation. Nonetheless, every model tested along with the visualisation of its intermediate results are available on the website and GitHub repository referenced in this document. To emphasise, *every model presented in this thesis was initialised with the K-means algorithm.*

3.2.1 Feature engineering and selection

Clustering algorithms are in general sensitive to the pre-filtered subset of features used to train them. The SBC itself had a decreased performance when it was deliberately trained on an increased fraction of irrelevant features in a set of simulated data (Ahmad and Fröhlich, 2017). This served as a motivation to look into pre-filtered predictor sets as input features for the SBC, by delving into different feature engineering and selection methods. The hope was to improve the survival curve separation and the Predictive C-Index, i.e., the predictive performance of the model when compared to the original SBC implementation by Ahmad and Fröhlich (2017).

Feature engineering is used in this master thesis as described by M. Kuhn and K. Johnson (2019): "the process of creating representation of data that increase the effectiveness of a model". This process might lead to a different number of predictors than the one that was contained in the original data and it necessarily involves the creation of new features, which might not all be necessarily relevant to the outcome. On the other hand, *feature selection* is the *"reduction of the number of predictors as far as possible without compromising predictive performance"*.

Consequently, two feature engineering/selection approaches on the training set were taken to implement the model in the TCGA-GBM data set: The use of mRNA

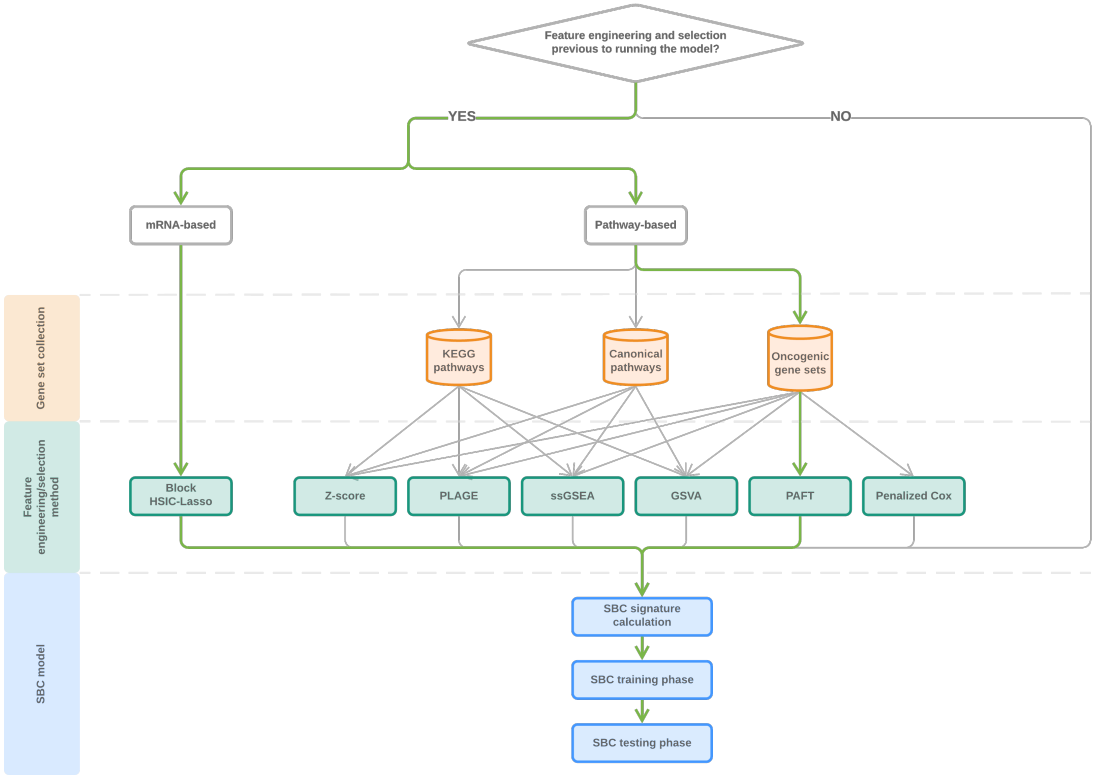


Figure 3.4: Flowchart of the pipeline used for the TCGA-GBM data modelling using SBC. Initially, no feature engineering or selection was done apart from the SBC signature calculation embedded in the model. Subsequently, different feature engineering and selection techniques on the mRNA expression data (both mRNA and pathway-based) were evaluated to improve the predictive ability of the model. The two paths highlighted in green yielded the best results, and the results of these two approaches are presented in detail in the next chapter.

microarray expression data as input features for the model (referred to as *mRNA-based feature selection*), and the use Gene Set Enrichment or aggregation scores and use the resultant summarised scores as input features (referred to as *pathway-based feature selection*). According to the aforementioned definitions, the pathway-based methods involved both feature engineering and selection procedures, because they resulted in a new set reduced of features that were used as an input for the model (which no longer were the original mRNA gene expression values), whereas the mRNA-based methods were only a feature selection procedure (after which the SBC received a subset of the original mRNA gene expression values). Both methods will be explained in the next sections.

mRNA-based feature selection

Block HSIC-Lasso In the machine learning field, biomarker discovery has been treated as a feature selection problem for which a varied set of linear and non-linear feature selectors for high-dimensional data sets have been proposed. The former does not always accurately represent biological data, and the latter group of feature selection methods such as SpAM(Ravikumar et al., 2009), MRMR(C. Ding and Peng, 2005) and HSIC-Lasso(M. Yamada et al., 2014) were developed with the aim to capture non-linear relationships between a class label and the input features.

HSIC-Lasso uses the Hilbert-Schmidt Independence Criterion (HSIC), a non-linear association measure of the dependence between two random variables X and Y, which in our case are the gene expression values and the time-to-event data (days to death, and days to last followup). HSIC is defined as:

$$\begin{aligned} \mathrm{HSIC}(X,Y) = \mathbb{E}_{x,x',y,y'}[K(x,x^{'})L(y,y^{'})] + \\ \mathbb{E}_{x,x'}[K(x,x^{'})]\,\mathbb{E}_{y,y'}[L(y,y^{'})] - \\ 2\,\mathbb{E}_{x,y}[\mathbb{E}_{x}^{'}[K(x,x^{'})\,\mathbb{E}_{y}^{'}[L(y,y^{'}]] \end{aligned} \tag{3.1}$$

Where $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $L : y \times y \to \mathbb{R}$ are positive definite kernels, and $\mathbb{E}_{x,x',y,y'}$ denotes the expectation over independent pairs (x, y) and (x', y') drawn from $p(x, y)$. HSIC is equal to 0 if X and Y are independent, and non-negative otherwise (Climente-González et al., 2019). HSIC Lasso is presented in

the following form:

$$\min_{\alpha \epsilon \mathbb{R}^d} \| \bar{\mathbf{L}} - \sum_{k=1}^{d} \alpha_k \bar{\mathbf{K}}^{(k)} \|^2_{Frob} = \frac{1}{2}\text{HSIC}(\mathbf{y}, \mathbf{y}) - \sum_{k=1}^{d} \alpha_k \text{HSIC}(\mathbf{u}_k, \mathbf{y}) + \frac{1}{2} \sum_{(} k, l = 1)^d \alpha_k \alpha_l \text{HSIC}(\mathbf{u}_k \mathbf{u}_l) \tag{3.2}$$

A state-of-the-art method that was used for this master project was Block HSIC-Lasso (Climente-González et al., 2019), an algorithm based on HSIC-Lasso that has been successfully tried on mRNA microarray data and solves the computational problems presented by the original method, while still selecting a small set of independent features that are highly dependent on the phenotype. The authors recommend using Block HSIC-Lasso when a small set of non-redundant, good predicting biomarkers is needed, which is exactly the case for the input features of the SBC. Block HSIC-Lasso was used in this project from the Python package *pyHSICLasso* with a Gaussian Kernel and a block size B of 20. The class labels used was the time-to-event data from the Clinical data set (without the censoring information), the input was the mRNA microarray expression matrix and the output was a a reduced mRNA microarray expression matrix, i.e., a matrix with only the subset of genes that Block HSIC-Lasso selected as features that correlated with the survival.

Pathway-based feature engineering

The motivation to produce a set of pre-filtered features using a pathway-based feature engineering and selection method, was to obtain a set of predictors that correlated with survival, increased the predictive ability of the SBC but also improved model interpretability by making use of the annotated biological knowledge that has been accumulated by the scientific community over the years. To achieve this, the idea was to use a group of pathway aggregation scores in group of gene sets, and obtain a matrix of aggregated scores of i samples by j gene sets. This required first selecting a list of gene sets and a set of methods to be assessed, which is why a brief description of the Molecular Signatures Database (mSigDB) and its gene set collections as well as a description pathway aggregation methods implemented in this project are described in the following paragraphs.

Molecular Signatures Database (mSigDB): The Molecular Signatures Database is a collection of annotated gene sets, and it is one of the largest and most popular repositories of gene sets (Liberzon et al., 2011). The latest version of mSigDB

contains 7 collections of grouped genes, which are constantly reviewed, curated and manually annotated. Three mSigDB gene sets were accessed in R through the CRAN package *(mSigDB)*.

The following gene sets were tested for the feature engineering process previous to the use of the SBC on the TCGA-GBM data set, because the curated group of genes were thought to be helpful for cluster interpretation and could contain a priori biological information relevant for cluster validation:

- **Canonical Pathways**: The C2 collection of mSigDB contains a subcollection called Canonical Pathways which has 1329 curated gene sets from databases such as BioCarta, KEGG, Matrisome Project, Reactome, among others (Broad Institute, August 7 of 2019[a]).

- **KEGG pathways**: The subcollection of Canonical pathways contains 186 gene sets derived to the KEGG pathway database.

- **Oncogenic Gene sets**: The collection C6 contains 189 gene sets of cellular pathways which tend to be dis-regulated in cancer were generated from microarray data from NCBI GEO, from experiments which involved perturbation of known cancer genes or curated from scientific publications (Broad Institute, August 7 of 2019[b]).

Pathway aggregation scores: With the advent of high-throughput mRNA experiments, GSE methods appeared as a way to facilitate the interpretation and assessment of biological activity. Their aim is to establish if there are deferentially expressed gene sets within a list of genes which is associated with a certain group, phenotype or biological condition. Most GSE analyses begin with a ranked gene list, followed with the mapping of such genes into predefined gene sets, and end with a gene expression statistic summarised into a single enrichment score for each gene set (Hänzelmann et al., 2013). Following the comparisons made by Hänzelmann et al. (2013) and using the GSE score methods available in their Bioconductor package *(gsva)*, four unsupervised GSE methods that calculate single sample pathways summaries of expression were tested in the feature engineering process:

- **Z-score**:

  If each gene set $\gamma = i, ..., k$ has a calculates z-score $Z_i, ..., Z_k$ for each gene inside the gene set, the summarised gene set score per gene set γ is defined

as (E. Lee et al., 2008):

$$Z_\gamma = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \tag{3.3}$$

- **Pathway level Analysis of Gene Expression (PLAGE)**:

  This method performs a Singular Value Decomposition (SVD) of the gene expression matrix of each gene set (i.e. the expression matrix contains only the genes within each gene set at a time), and computes a "metagene"' which are the coefficients of the first right-singular vector. This "metagene" is interpreted as the summarised gene set score for each gene set (Tarca et al., 2013; Tomfohr et al., 2005).

- **Single Sample Gene Set Enrichment Analysis (ssGSEA):**

  This method is presented as a variation of the traditional GSEA, where instead of ranking genes based on their correlation score with their class label, genes are ranked based on their absolute expression (i.e. enrichment scores are calculated independent of the phenotype, and therefore the method becomes unsupervised). ssGSEA computes a summarised gene set score as the "normalised difference in empirical cumulative distribution functions of gene expression ranks inside and outside the gene set" (Hänzelmann et al., 2013).

- **Gene Set Variation Analysis (GSVA)**:

  GSVA was also presented as an alternative to the limitations that the traditional Gene Set Enrichment Analysis (GSEA) had when dealing with more complex experimental designs than pairwise comparisons between two well defined groups, which is always the case with the information present in the TCGA data set. GSVA evaluates if a gene i is highly or lowly expressed in a sample j from an input matrix of p genes by n samples, by performing a kernel estimation of the cumulative density function for each gene expression profile $x_i = x_{i1}, ..., x_{in}$, and producing an expression-level statistic denoted by z_{ij}. z_{ij} is later converted to ranks for each for each sample j, and then normalised to make the ranks symmetric around zero. Next, the Kolmogorov-Smirnov-like rank statistic is calculated for every gene set. The GSVA method offers two approaches for turning the obtained Kolgomorov-Smirnov statistic into an enrichment statistic: the classical maximum deviation method (Subramanian et al., 2005) that produces a distribution of

enrichment scores that is bimodal for each gene set, and a normalised enrichment statistic proposed by Hänzelmann et al. (2013) that is unimodal and approximately normal. The output of the algorithm is a matrix of pathway enrichment scores of size m gene sets by n samples.

Additional to the four traditional methods previously mentioned, two other ad-hoc "supervised" pathway activity score methods were also evaluated:

- **PAFT**: A Penalized Accelerated Failure Time model (see Equation 2.4) was fit on the mRNA expression matrix of every Oncogenic gene set, in order to get a vector of linear predictors of length equal to the number of patients on the training set. The matrix built with the concatenated linear predictors vectors, was used as matrix of input features for the training set. The previously described model (which was fit on the training data set only), was later used to obtain a matrix of linear predictors on the gene expression data of the validation set, and such matrix of linear predictor was used on the testing phase of the SBC. The PAFT was fit ignoring the censoring information and using *(glmnet)* R package.

- **Penalized Cox Proportional Hazard Model**: The Cox regression model (Cox, 1972) is the most popular regression analysis method for survival data (Gui and Li, 2005), and has been implemented in its regularised form by the package *(glmnet)* (Friedman et al., 2010).The Cox model is described by the following equation:

$$h_i(t) = h_0(t)e^{x_i^\top \beta} \tag{3.4}$$

  where $h_i(t)$ is the hazard for patient i at times t, $h_0(t)$ is a shared baseline hazard and β is a vector of predictors (Friedman et al., 2010). For this feature engineering approach, the same process as described for the AFT based feature aggregation method was used, using a Penalized Cox model instead of a Penalized Accelerated Failure Time.

3.2.2 SBC signature calculation

The SBC performs an internal feature selection referred to as SBC signature calculation, where selected features are filtered as the top genes of the ranked p-values from an univariate Cox regression model evaluated on the training set only (Ahmad and Fröhlich, 2017). SBC signature calculation occurs before the training and testing phase begin (see Figure 3.4, section in blue) and it was performed *before every SBC model trained in this project.*

The way the SBC signature calculation was integrated with the mRNA-based feature selection was by first performing the feature selection method as described in Section 3.2.1 and then calculating the gene signature. On the other hand, the pathway-based feature engineering process and the SBC signature calculation were integrated as follows:

1. Three gene set collections (Canonical pathways, KEGG pathways or Oncogenic gene sets) were downloaded from the Molecular Signatures Database (mSigDB).

2. Four GSE score methods (Z-score, PLAGE, GSVA and ssGSEA) were simultaneously tested on each of the three gene set collections (KEGG pathways, Oncogenic gene sets, Canonical pathways), in order to obtain one matrix of aggregated scores per gene set collection (12 in total).

3. The SBC gene signature was calculated as implemented in the original model, using as input the previously obtained enrichment scores matrices of i patients by j gene sets. The result were 12 SBC signatures.

4. The GSE method that yielded the SBC signature with the highest -log(p-val) was selected for the next steps, while the other 3 GSE methods are left out and no model is trained using them.

5. The question to answer now is which of the gene set collections yields the lowest p-values after SBC signature calculation. Once more, by plotting the -log(pval) of the three matrices from the previous step, the gene set collection with the most significant p-values is selected.

6. Along with the best performing GSE method from above, two ad-hoc "supervised" pathway activity score methods were evaluated (PAFT and Penalised Cox). They are evaluated in the selected gene set collection from the previous step as described in the section 3.2.1. The matrix of aggregated scores that results after implementing this methods, is used as input to calculate the SBC gene signature, train and test the SBC.

The results obtained after step 3 can be seen in Figure 3.5, where the log transformed p-values plotted represent the distribution of the univariate association of SBC signature pathways calculated using four different GSE methods (Z-score, PLAGE, GSVA and ssGSEA) on 3 different gene set collections (Oncogenic gene sets, KEGG pathways and Canonical pathways). ssGSEA was successfully selected as the GSE method that produced the most significantly correlated aggregated

scores with survival, regardless of the gene set collection being evaluated. Furthermore, the oncogenic gene sets collection had the most significant p-value of all the methods, regardless of the GSE technique evaluated.

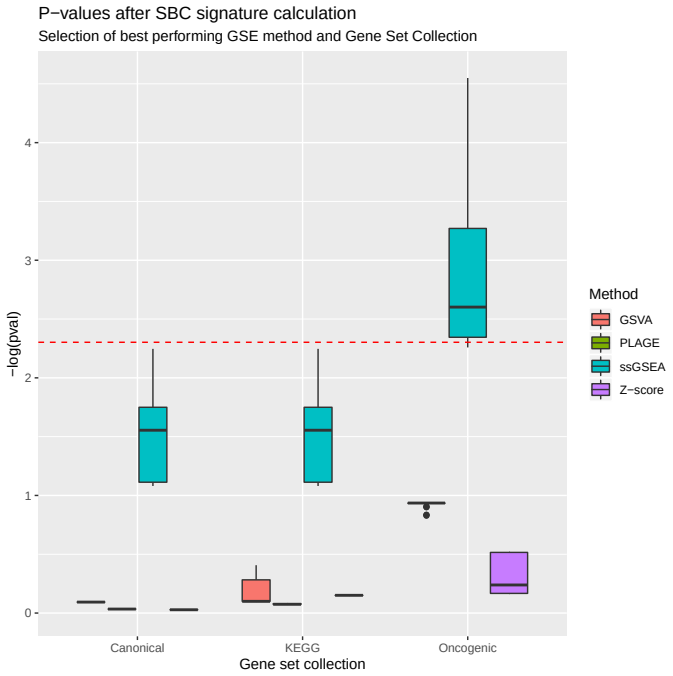


Figure 3.5: Comparison of the p-values per GSE method per gene set after SBC signature calculation. The red dashed line corresponds to a threshold of a p-value of 0.1. The p-values plotted have been -log transformed, i.e, the higher the bar the more significant the result is.

3.2.3 Model implementation

The code with the implemented model in R was obtained from the GitHub repository `https://github.com/ashar799/SBC`. However, a modification of the pipeline was done for the completion of this master project. A more reproducible script with the previously described feature engineering and selection methodologies was developed in the form of an R notebook from which a rendered document with plots and visualisations after every main step of the process is obtained per run. This script allows for the tuning of several model feature engineering/selection steps, as well as hyperparameters such as the p-value threshold for the selection of the SBC signature, the number of burnin iterations and the frequency of the thinning of the MCMC sampling, so that the code can be run on a cluster using shell commands. The full code used in the model implementation for every model trained as well as the visualisations produced after every step of the process can

be seen in the website `https://camiladuitama.github.io/MasterThesis/`

Every script consists of 3 main sections: **feature engineering and selection**, **SBC signature calculation**, **training phase** and **validation phase**. For the Gibbs Sampling 150 burn-in iterations were used, 20 MCMC samples were produced with samples being drawn every 5th iteration (thinning). Log-likelihood trace plots were used to assess the convergence of the MCMC chain. After training the model (learning all model parameters, plus cluster memberships), the validation data set was then assigned a cluster membership. Moreover, recovery and predicted C-Indices were calculated and plotted.

3.3 Multi-omics characterisation of the SBC results

The idea behind using other omics data sets was to see if the discovered (training set) and predicted (validation set) SBC clusters had distinct expression patterns visible in other data types, apart from the mRNA microarray data used to train the model. The data types used for the cluster characterisation and interpretation were Copy Number Variation and miRNA data.

The format of the Copy Number Variation data was a matrix of discrete values with patients as rows and gene names as columns, in which genes with no amplification or deletion had 0 as an entry, genes with amplification had positive values and genes with deletions had negative values. CNV data, was used to find corresponding correlations between the SBC labels using a Pearson's Chi-square test per feature gene and performing a multiple hypothesis test correction (FDR < 0.05).

Finally, the mRNA and miRNA microarray data sets were subjected to differential expression analyses using the package *(limma)*(Law et al., 2014). With the use of the SBC labels obtained after model training and testing, two type of comparisons were performed:

- Pairwise comparisons between the discovered SBC clusters, looking for differentially expressed genes between pairs of clusters (FDR< 0.05).

- A combination of all pairwise comparison into one F-test (equivalent to one-way ANOVA for each gene, except that the residual mean squares have been moderated between genes) in order to find the most deferentially expressed genes (FDR< 0.05) between all the SBC clusters. With this comparison, the idea was to find which genes are significantly up or down-regulated in any way between the discovered or predicted subtypes.

3.4 Clinical characterisation of the SBC results

Correlations between all the available clinical variables (see Table 3.2) were tested with a Pearson's Chi-square test, and significance was accepted if the p-value was below 0.05. This was done in order to achieve better cluster interpretability, by evaluating if a certain feature was significantly different between the SBC clusters. Moreover, the idea to examine if there are possible correlations between other clinical variables and the SBC clusters, was to try to detect possible confounding factors that are leading the SBC to discover trivial subtypes. Table 3.2 gives an overview of the clinical variables tested, including the clinical variable name as it appeared on the dataframe retrieved from the TCGA-GBM cohort. The explanation of what each of the clinical variable names means and the possible values they can take was taken from the documentation in the GDC Data Portal (see the Data Dictionary Viewer for further explanation)

Association between the SBC predicted labels and the known Verhaak labels for the training set was accepted if the Fisher's exact test had a p-value below 0.05. The intention behind this test was to evaluate if the known Verhaak subtypes could be associated with the SBC when tested in the TCGA-GBM cohort.

Clinical variable name	**Explanation**	**Values**
years_of_birth	Age	Integer
days_of_death	Number of days between the date used for index and the date from a person's date of death.	Integer
days_of_last_followup	Number of days between the date of last follow up to the date of initial pathological diagnosis.	Integer
gender	Text designations that identify gender. Identification of gender is based upon self-report.	-Female -Male -Unknown -Unspecified -Not reported
date_of_initial_pathologic_diagnosis	Year in which the first pathological diagnosis was made.	Integer
radiation_therapy	Whether the patient received radiation therapy or not.	Yes/No
karnofsky_performance_score	Text term used to describe the classification used of the functional capabilities of a person.	-Values from 0-100, with 10 points intervals -Unknown -Not Reported
race	An arbitrary classification of a taxonomic group that is a division of a species. The provided values are based on the categories defined by the U.S. Office of Management and Business and used by the U.S. Census Bureau.	-White -American indian or alaska native -Black or african american -Asian-native hawaiian or other pacific islander -Other -Unknown -Not reported -Not allowed to collect
ethnicity	Classification based on social groupings that are characterised by a distinctive social and cultural tradition maintained from generation to generation, a common history and origin and a sense of identification with the group.	-Hispanic or latino -Not hispanic or latino -Unknown -Not reported -Not allowed to collect

Table 3.2: Description of clinical variables for the TCGA-GBM

4 Results

As mentioned in the previous chapter, the major focus on modelling data with the SBC (and its subsequent validation) is the feature engineering and selection process. The results presented in the following sections are therefore centered around this part of the model implementation. Nonetheless, data preprocessing is also important in any validation study, as it helps avoiding sample selection bias or unwanted batch effects. In order to address these concerns, a comparison of the clinical characteristics of the training and validation samples was done, and as a result, the size of the validation set was reduced to only the Karnofsky matched validated samples. Additionally, training and validation cohorts were independently scaled with Z-score scaling, and the initialisation of the parameters of the SBC model was done using k-means clustering, as in the original implementation of the SBC (Ahmad and Fröhlich, 2017).

The feature engineering and selection methods tested in this project were divided into two categories: pathway-based feature engineering in which a set of new aggregated scores was used as input for the model, and mRNA-based feature selection methods in which a subset of the gene expression matrices was used as input for the SBC. The process has a graphical representation in Figure 3.4, and the summarised results can be seen on Table 4.1. The SBC was trained 5 times: Model 1 was implemented exactly as described by Ahmad and Fröhlich (2017) (this original method included a form of feature selection called SBC signature calculation which uses only training data to filter a subset of genes based on their p-values from univariate Cox regression models), models 2-6 used a pathway-based feature engineering technique previous to the SBC model implementation, and model 7 used a mRNA-based pathway engineering technique. It is important to reiterate that the running hyperparameters and input data for every model were always the same.

The results for the two best performing feature engineering and scaling methods (highlighted in green in Figure 3.4) are presented in this chapter: The SBC model trained after mRNA-based Block HSIC-Lasso feature selection, and the SBC model trained after the PAFT pathway-based feature selection. In order to select the best performing models for the upcoming cluster and signature characterisation, the following criteria were used:

Number	Feature engineering/selection method previous to the SBC	Description	Gene set	Feature aggregation/selection method	Input to the SBC	K training[1]	K validation[2]	P-value training	P-value testing	Predictive C-Index	Recovery C-Index
1	None	Original SBC implementation	None	SBC signature calculation only	mRNA expression values	4	4	3.197e-06	1.704e-03	0.478	0.683
2	Pathway-based	ssGSEA on KEGG pathways	KEGG Pathways	ssGSEA	Gene set aggregated scores	4	3	2.205e-01	6.616e-02	0.509	0.469
3		ssGSEA methods on Oncogenic gene sets	Oncogenic gene set			4	4	3.958e-04	2.667e-02	0.5010	0.634
4		ssGSEA on the Canonical Pathways	Canonical pathways			4	2	6.808e-03	1.157e-01	0.524	0.604
5		Penalised-Cox based feature aggregation	Oncogenic gene set	Penalised Cox		5	3	<5e-10	7.452e-02	0.530	0.966
6		PAFT based feature aggregation		PAFT		4	4	<5e-10	1.806e-02	0.528	0.959
7	mRNA-based	Block HSIC-Lasso with survival data	None	Block HSIC-Lasso	mRNA expression values	4	4	1.078e-05	4.701e-02	0.565	0.698

Table 4.1: Consolidated results for the models trained
The two best performing models are highlighted in green.

[1] Number of clusters in the training set (discovered clusters)
[2] Number of clusters in the validation set (predicted clusters)

- Equal number of clusters on the training and the validation sets. If the number of discovered clusters was different from the number of predicted clusters such model would be discarded.

- Using only the models whose number of SBC clusters on the training and validation sets were the same, one point is assigned to the model every time one of the following statements is true:
  - Lowest p-value (< 0.05) in the Kaplan Meier curves for the training set in the group.
  - Most significant p-value (< 0.05) in the Kaplan Meier curves for the validation set in the group.
  - Highest possible values for the Predictive C-Index in the group.
  - Highest possible values for the Predictive C-Index in the group.

- The two models with the highest score are selected for posterior analysis.

Number	Description	=K [3]	Lowest p-value training	Lowest p-value validation	Highest Predictive C-Index	Highest Recovery C-Index	Total
2	ssGSEA on KEGG pathways						0
3	ssGSEA methods on Oncogenic gene sets	✓					1
4	ssGSEA on the Canonical Pathways						0
5	Penalised-Cox in Oncogenic gene sets						0
6	PAFT in Oncogenic gene sets	✓	✓	✓		✓	4
7	Block HSIC-Lasso with survival data	✓			✓		2

Table 4.2: Scores for the selection of the best performing models

After assigning a score to the models 2-7, as previously described, the Table 4.2 shows the results for each of the models that used a form of feature engineering or selection previous to the SBC implementation. The Block HSIC-Lasso feature selection model (Model number 6, Predictive C-Index = 0.565, Recovery C-Index = 0.698) had a higher Predictive C-Index and a lower Recovery C-Index than the PAFT feature engineering model (Model number 7, Predictive C-Index = 0.528, Recovery C-Index = 0.959). Models 6 and 7 were selected for cluster and signature characterisation and further analysis with other types of omics data also because they were representatives of the two forms of feature engineering/selection tested (pathway-based and mRNA-based respectively) so further comparison would be

[3]Number of discovered clusters (training set) are equal to the number of predicted clusters (validation set)

interesting, because they obtained the highest scores when checked for the conditions to select the best-performing model and finally because the Predictive C-Index and Recovery C-Index improved with respect to the original implementation of the model by Ahmad and Fröhlich (2017).

The convention to interpret the SBC clusters used for the figures presented in this chapter, was to use the labels of "Good","Good moderate", "Bad moderate" and "Bad" according to the mean survival times of the SBC cancer subtypes.

4.1 Block HSIC-Lasso + SBC

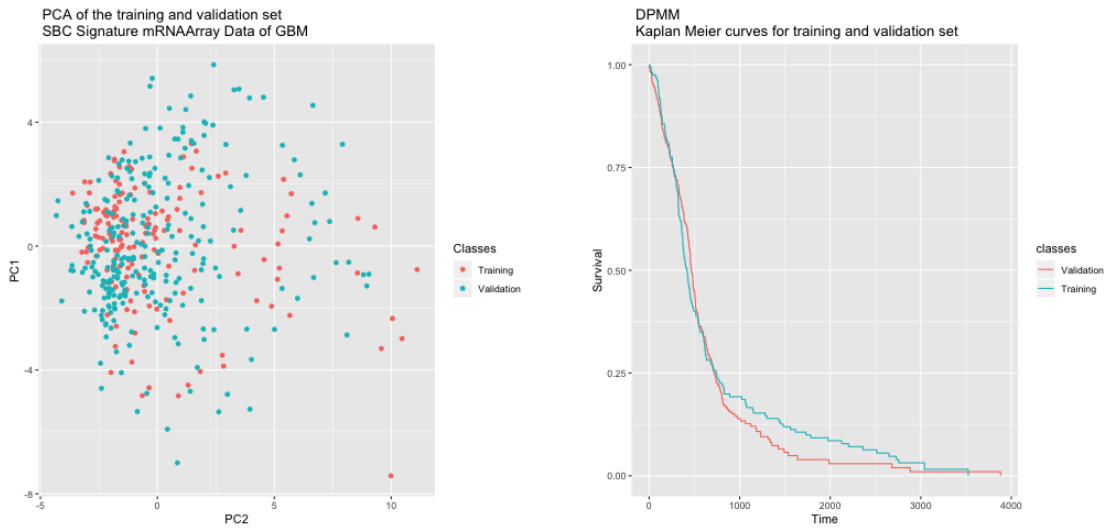
Block HSIC-Lasso (Climente-González et al., 2019) is a state-of-the-art multivariate feature selection algorithm and it was the only mRNA-based feature selection method tested before the implementation of the SBC (this is before SBC signature calculation, training and validation phases, see Figure 3.4). Block HSIC-Lasso selects a group of genes that are dependent on the survival data, taking as input the mRNA microarray expression data and producing as output a reduced mRNA microarray expression matrix with only the pre-filtered subset of selected features. There is no generation of new features in this process.

4.1.1 Model training

By visually inspecting the joint PCA plot and Kaplan Meier curves in Figure 4.1 for batch effects, it is possible to ascertain that the training and validation sample sets are similar molecularly as well as clinically. It is worth mentioning that the survival curves for all the models trained are equal, because each SBC model fit used the same set of patients for training and validation (respectively). However, the PCA of the SBC signature changes between model runs, because the input data to the model is modified (in some cases it was unmodified mRNA gene expression values, in some other the result of an pathway-based feature engineering method that introduced new feature sets).

Figure 4.2 depicts the summarised visualisations for training and validation sets of the Block HSIC-Lasso feature selection model after fitting SBC model on training data and making corresponding predictions on validation data set. The PCA plots are built on the mRNA expression of the calculated SBC signature, which in this case consists of 50 genes selected as input features for model training. Besides, the Kaplan Meier survival curves are based on the overall survival of the patients on each data set, using the recovered and predicted cluster labels on training and validation sets respectively.

Figure 4.2(b) shows four distinct survival curves (log-rank p-value = 1.078e-05) and Figure 4.2(a) shows four relatively separated clusters in the PCA for the training set. For the validation set the results are still significant on the survival curves as it can be seen in Figure 4.2(d) (p-value = 0.047), but the separation is less stark. Similarly, the clusters in the PCA are also less clearly separated for the validation set, as it appears in Figure 4.2(c). In summary, Figure 4.2 shows that the SBC successfully finds clinically relevant cancer subtypes in both the training and validation sets, with distinct survival curves and its corresponding clear molecular profiles based on a 50 gene signature.

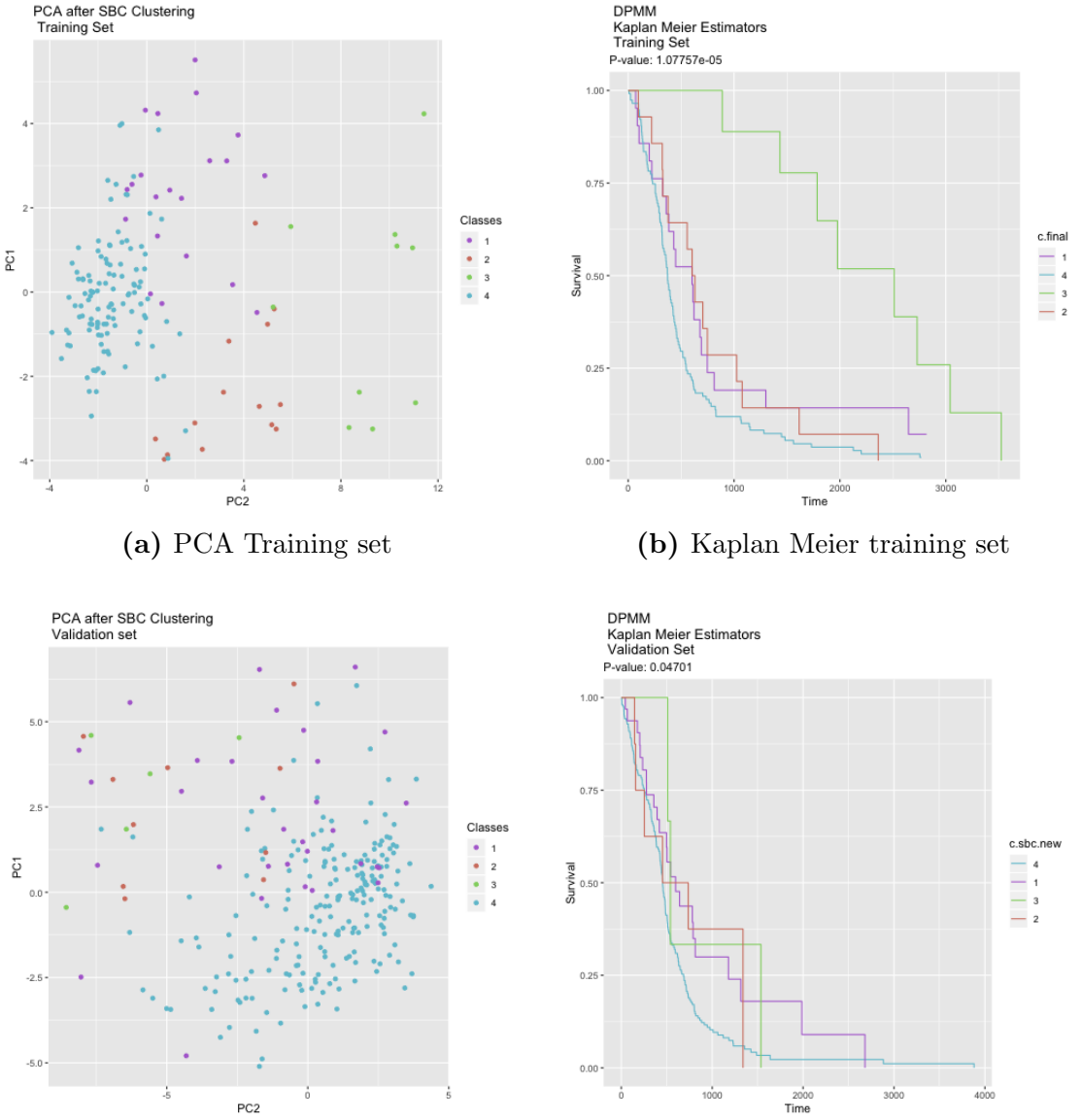


(a) PCA SBC Signature **(b)** Kaplan Meier curves

Figure 4.1: PCA of input features and Kaplan Meier curves after feature selection with Block HSIC-Lasso

From this initial results, it was determined that for the following analyses regarding the HSIC-Lasso feature selection results, cluster 1 will be considered as "Bad Moderate", cluster 2 as "Good Moderate", cluster 3 as "Best" and cluster 4 as "Worst".

Log-likelihood trace plots such as the one in Figure 4.3(a) help assess the convergence of the MCMC sampler which is evident in this case after the 25^{th} iteration. Figure 4.3(b) is the posterior probability plot of selection of covariates in the Lasso penalised cluster specific survival models (AFT), where only the clusters with "Worst" and "Bad moderate" survival times (4 and 1 respectively) have features related to survival.



(a) PCA Training set

(b) Kaplan Meier training set

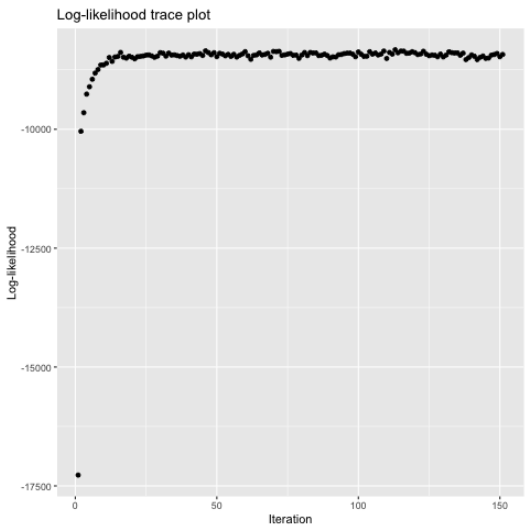
(c) PCA validation set

(d) Kaplan Meier validation set

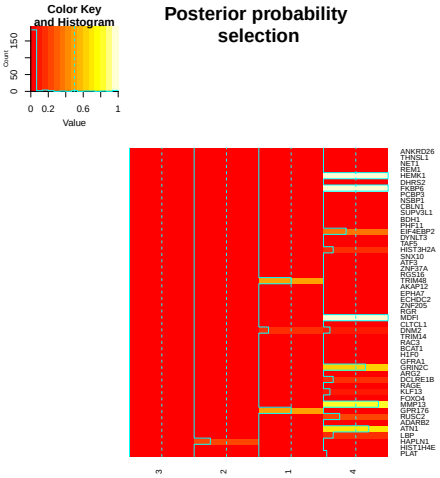
Figure 4.2: Training and validation results of the SBC model after using Block-HSIC Lasso for feature selection. Clinical end-point is overall survival and colours on each plot represent the different predicted classes from the SBC

An uneven distribution of patients among clusters can be seen on Table 4.3. This might be explained by the fact that the SBC can lead to discovery (and later prediction) of clusters with few patients, samples which may not have any covariates associated with them in their corresponding cluster specific penalised AFT survival model (as it is implemented in the SBC). Besides, a cluster of bigger size is more likely to find survival-related features on the data. The following

results point to the fact that SBC model predictions indeed lead to 4 distinct clusters in the validation cohort with its corresponding survival curves. In the coming paragraphs, an effort is made towards trying to understand the model fit of the SBC on the training data by using other omics data sets for characterizing the results.



(a) Log-likelihood trace plot



(b) Posterior probability

Figure 4.3: Log-likelihood trace plot during the burning and posterior probability for the training after using Block-HSIC Lasso for feature selection.

SBC discovered/predicted subtype	Training set	Validation set	Total
"Bad moderate" (Cluster 1)	21 (13%)	32 (12%)	54 (13%)
"Good moderate" (Cluster 2)	14 (9%)	10 (4%)	24 (5%)
"Best" (Cluster 3)	10 (6%)	5 (2%)	15 (4%)
"Worst" (Cluster 4)	115 (72%)	214 (82%)	329 (78%)
Total	160	261	421

Table 4.3: Cluster composition in training and validation sets after Block HSIC-Lasso feature selection. The percentages shown are calculated with the total of each column in the following manner: The "Bad moderate" group has 21 patients in the training set and it is divided by the total number of individuals in the training set ($\frac{21}{160} * 100$)

4.1.2 SBC signature and model interpretation

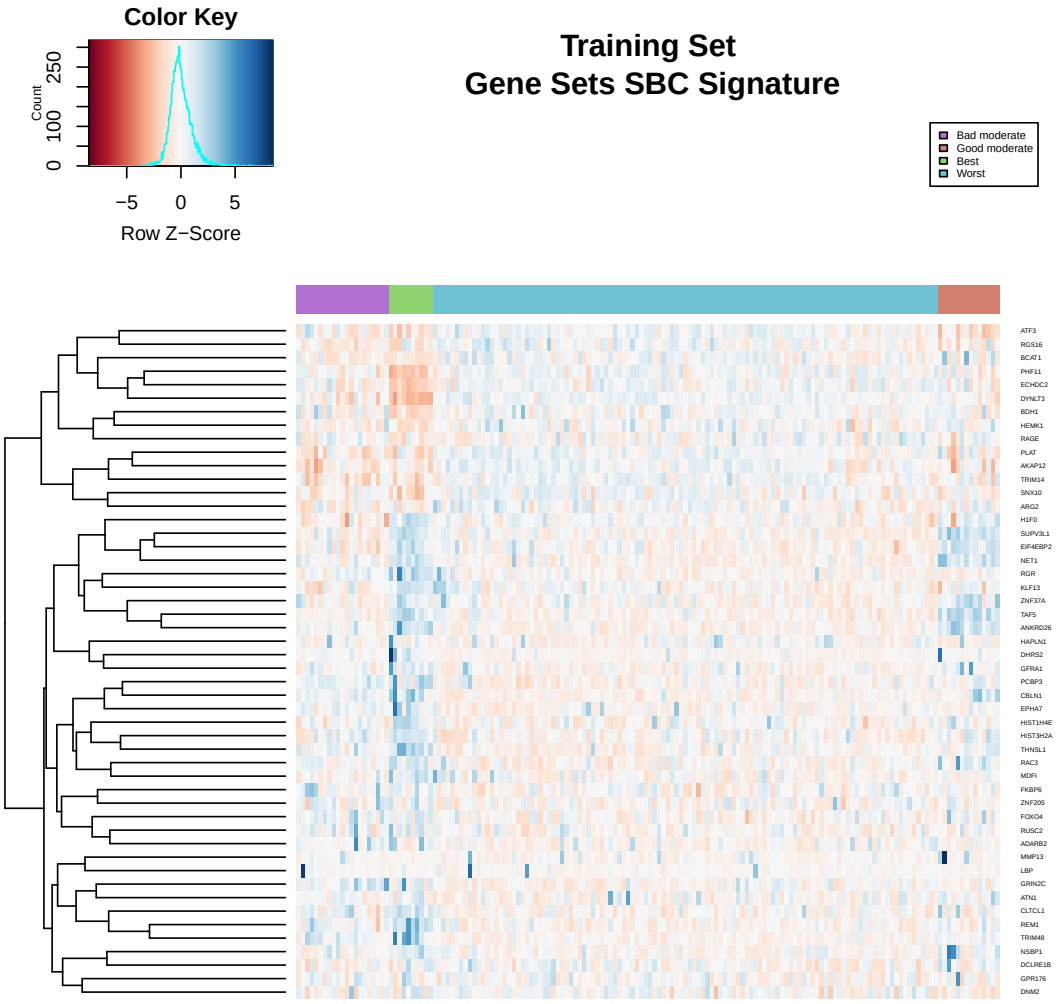


Figure 4.4: Heatmap of the SBC signature for the model trained after Block HSIC-Lasso feature selection.

Figure 4.4 shows the molecular signature profile of the SBC clusters in the training set in a heatmap. The columns which correspond to patients have been ordered from their score using the SBC, and the blocks with high levels of expression are colored in blue, whereas low levels of expression are colored in red on each cluster. This confirms once more that the SBC not only finds subtypes of cancer diagnosed patients according to their survival times, but it also finds distinct molecular profiles that correspond a different survival curve.

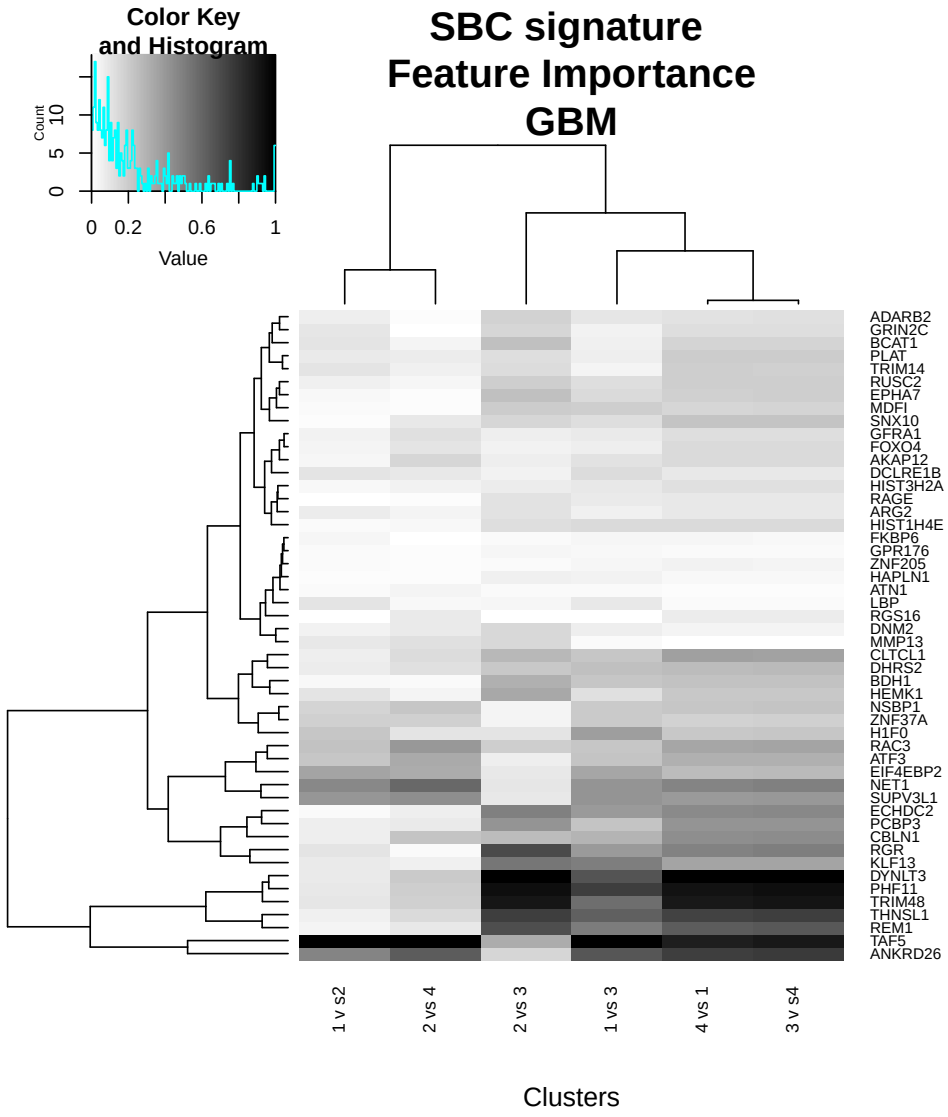


Figure 4.5: Feature importance in discriminating respective clusters of the SBC signature for the model trained after Block HSIC-Lasso feature selection. Darker colours imply stronger effects.

To identify what are the genes that contribute the most to distinguish between different clusters of patients, a feature importance plot (Figure 4.5) is shown, where feature importance is calculated using Equation 2.9. Some genes which have higher contributions across several cluster comparisons like **DYNLT3** have already been reported to have differential expression among long term and short term survival clusters of Glioblastoma Multiforme (Gerber et al., 2014; Y.-W. Kim et al., 2013) patients and have been established as gene markers of GBM prognostic subtypes (Park et al., 2019). On the other hand, **PHF11** which is relevant for

cluster separation in 3 different pairwise comparisons, is also part of the Verhaak et al. (2010) signature. Additionally, **TAF5** which appears as relevant in the SBC model trained after feature selection with HSIC-Lasso had also been used to predict GBM prognosis networks using weighted gene co-expression networks (Xiang et al., 2012).

4.1.3 Characterisation of SBC clusters in other omics data sets

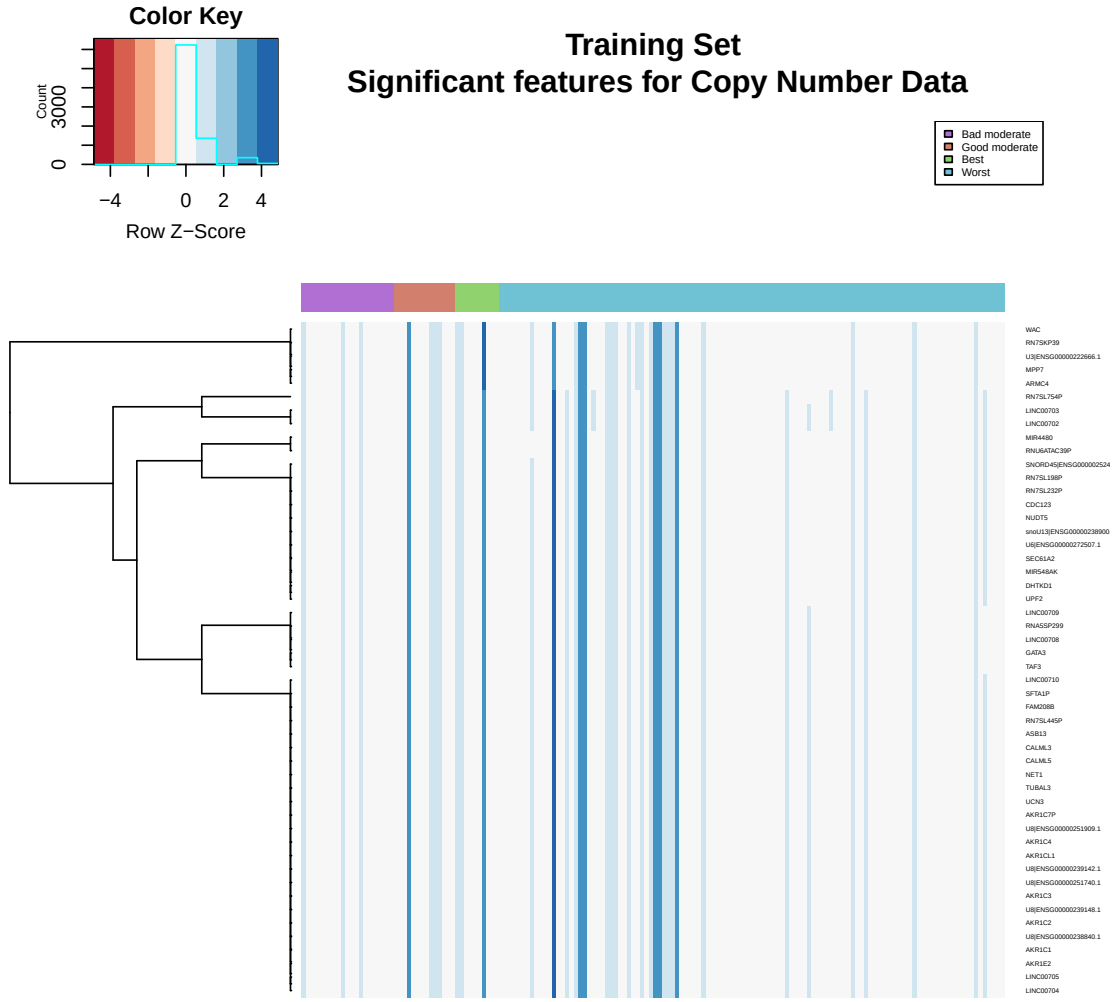
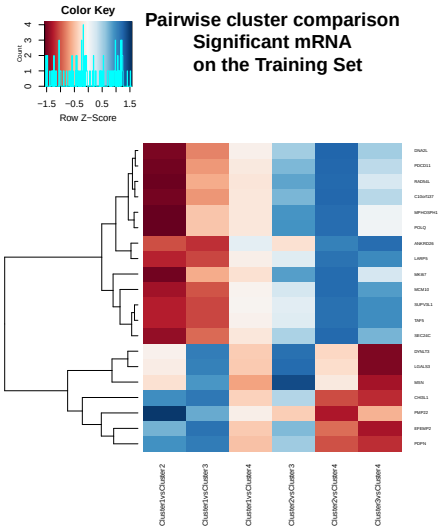
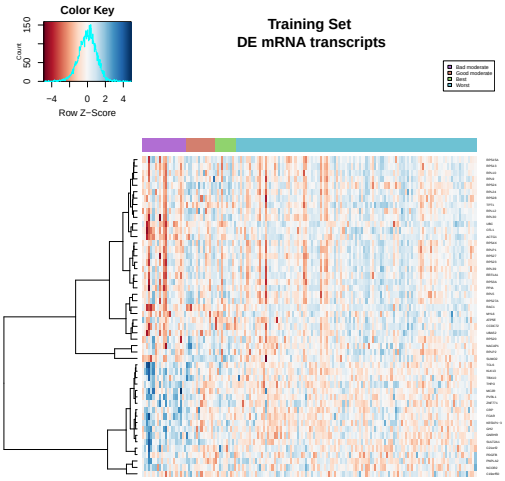


Figure 4.6: Copy Number Variation heatmap after Block-HSIC Lasso feature selection.

The use of Copy Number Variation data as external data set, i.e., data that is not used as input for the SBC model, was a way to evaluate if the discovered SBC clusters had distinct expression patterns visible in other data types apart from the mRNA microarray data used to train the model.



(a) Heatmap of DE mRNA in cluster pairwise comparison



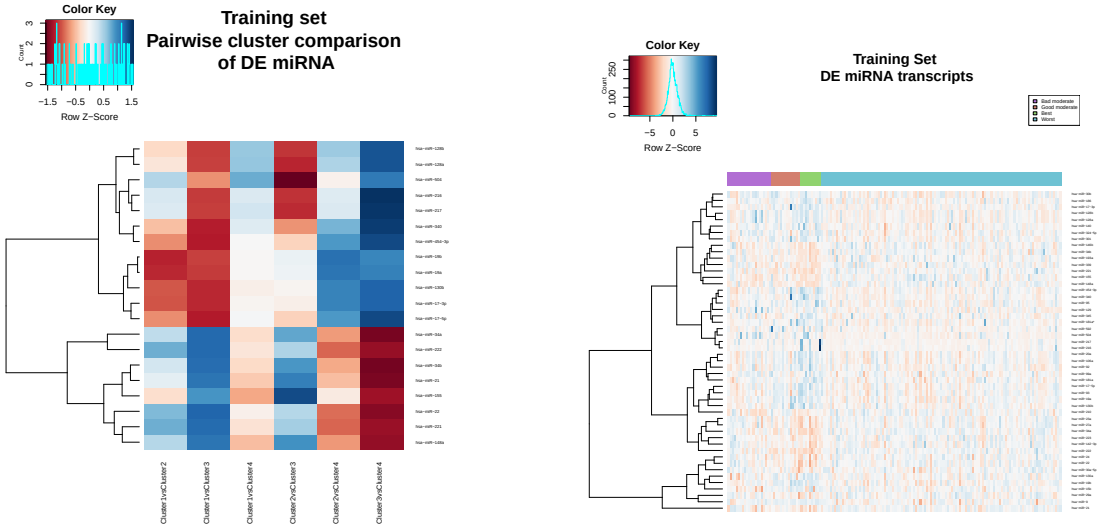
(b) Heatmap of DE mRNA per cluster

Figure 4.7: DE mRNA transcripts in pairwise cluster comparison and per cluster comparison using the predicted SBC labels. After the differential expression analysis with *(limma)* on the mRNA data set, two plots were obtained: (a) shows the log expression of the most significant mRNA transcripts per pairwise comparison between the SBC clusters, (b) shows the top 50 most significantly DE genes (FDR < 0.05). The SBC labels on the plots were obtained with thecorrespond to the model trained after HSIC-Lasso feature selection

When exploring Copy Number Variation data, the idea was to account for a possible association between the expression pattern of CNV changes and the predicted SBC labels. In order to facilitate visualisation, the top 50 significant genes were used to plot Figure 4.6, based on a Chi-squared test to rank the most significantly associated CNVs to the SBC clusters. There seems to be a pattern of high values for the Z-score for the cluster with "Worst" survival, meaning that the cluster with the shorter survival times had events of insertions of 50 bp or longer in their genome. Besides, the most significant Copy Number Variation events are related to mutations in the form of insertions (blue colours), and not so much to deletions (red colours). Additionally, **NET1** overlapped between the SBC signature and the genes with copy number changes, a neuroepithelial cell transforming gene that has proven to be significantly up-regulated among different TCGA-GBM

survival-related subtypes after CNV analysis (Y.-W. Kim et al., 2013).

mRNA and miRNA microarray expression data are used now in a post-hoc manner to interpret the discovered SBC clusters. The idea was to obtain the deferentially expressed feature sets which that are distinguishable pairwise amongst clusters and overall between the four clusters, using both the microarray mRNA gene expression data and the miRNA microarray expression data available for the samples selected from the TCGA-GBM cohort.



(a) Heatmap of DE miRNA in cluster pairwise comparison

(b) Heatmap of DE miRNA per cluster

Figure 4.8: DE miRNA transcripts in pairwise cluster comparison and per cluster comparison using the discovered SBC labels. After the differential expression analysis with *(limma)* on the miRNA data set, two plots were obtained: (a) shows the log expression of the most significant miRNA transcripts per pairwise comparison between the SBC clusters, (b) shows the top 50 most significantly DE miRNA (FDR < 0.05). The SBC labels on the plots correspond to the model trained after HSIC-Lasso feature selection

After performing a Differential Expression Analysis on the mRNA TCGA-GBM data, once more the expression patterns of the predicted SBC clusters seem to be distinct across SBC subtypes (Figure 4.8 (b)). **DYNLT3**, **ANKRD26**, **TAF5**, **SUPV3L1**, **NET1** and **TRIM48** appear both among the Deferentially Expressed genes and in the SBC signature. **TAF5** stands out because it is also part of the Verhaak signature. This gene codifies for a component of the transcription factor IID complex, essential for mediation regulation of the RNA polymerase transcription.

For the miRNA differential expression analysis, the "Best" and "Good moderate" cancer subtypes seem to have similar miRNA expression patterns. Interestingly **miR-222** and **miR-221**, which appeared both as DE miRNA (see Figure 4.8 (b)) and as distinctive features for the pairwise cluster comparison of SBC subtypes (see Figure 4.8 a), have been reported as regulators of glioma tumorigenesis through the control of PTPμ protein expression (Quintavalle et al., 2012). Additionally, **miR-34a** which is also DE among the predicted SBC clusters, was reported as downregulated in GBM patients (Guessous et al., 2010) and was found as a relevant for the analysis of MicroRNA regulatory networks in Glioblastoma Multiforme (Sun et al., 2012). This miRNA regulates proteins involved in cell cycle, apoptosis, cellular differentiation and cellular development (Chen and Hu, 2012). Similarly, **miR-9** appears in the literature as a suppressor of mesenchymal differentiation in GBM by downregulating expression of JAK kinases and inhibiting activation of STAT3 (T.-M. Kim et al., 2011) and is expressed distinctively among the SBC subtypes predicted after Block HSIC-Lasso feature selection.

4.1.4 Characterisation of SBC clusters using other clinical data

The use of other clinical features apart from survival data for cluster characterisation, was intended to increase cluster interpretability and help in the detection of possible confounding factors, lurking variables that might have led to the discovery of trivial clusters. For this reason, several Chi-squared test were performed between the features and the SBC labels. Finally, after a Pearson's Chi-squared test clinical features available for the samples except for the survival data, the **date of initial pathological diagnosis** (FDR = 3.8e-02) and **age** (FDR = 1.9e-03) correlate with the discovered SBC labels on the training set after performing feature selection with Block HSIC-Lasso on the mRNA level. Furthermore, the discovered subtypes correlate only slightly (p-value =0.072) with the Verhaak labels reported in the literature (Mesenchymal, Proneural, Neural and Classical). This might be due to the patterns in gene expression which both SBC and Verhaak models try to capture but using different modelling assumptions.

4.2 PAFT pathway-based feature engineering + SBC

A Penalized Accelerated Failure Time model was used here as a pathway-based engineering approach previous to the calculation of the SBC signature, by using mRNA pathway data from Oncogenic gene set collection of mSigDB, a collection

that contained 189 gene sets. One model was fit for the mRNA expression matrix of the training data for each gene set, that is to say, the gene expression matrix containing only the genes belonging to each gene set. The vector of linear predictors for each model fit were concatenated and used as input features for the training phase. The model fit on the training data was later used to obtain a matrix of linear predictors on the validation set, which would then be used on the testing phase.

4.2.1 Model training

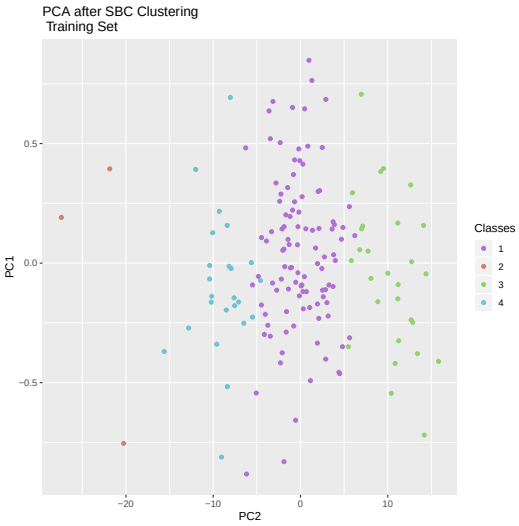
(a) PCA SBC Signature

(b) Kaplan Meier curves

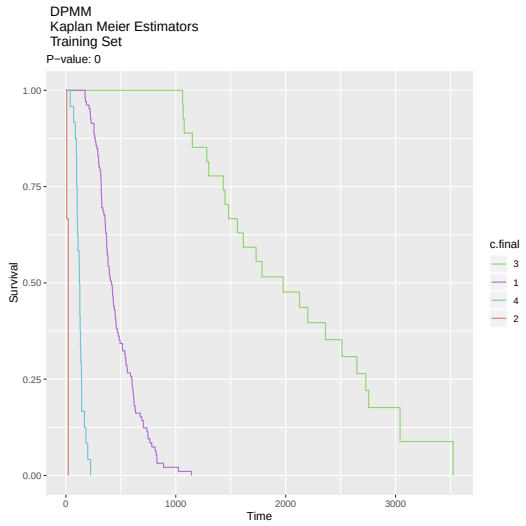
Figure 4.9: PCA of input features and Kaplan Meier curves after feature selection with PAFT model

Figure 4.9(a) shows the PCA of the input features after performing the feature selection with the linear predictors of the PAFT model. After a visual inspection at the survival curves and the PCA plots, and in comparison with Figure 4.1(a), it is possible to confirm that the training and validation sets are more or less similar clinically as well as molecularly. As mentioned before, the input features of every model were always independently scaled and centered, because this yielded the best results. Nonetheless, the PCA for this model showed a certain bias between test and training sets, despite the Karnofsky Index correction and the independent Z-score scaling. Additionally, it is worth noting that Figure 4.9(a) and Figure 4.1(b) are exactly the same, because the training and validation splitting of the TCGA-GBM data (as it was explained in the subsection 3.1.2) was the same for

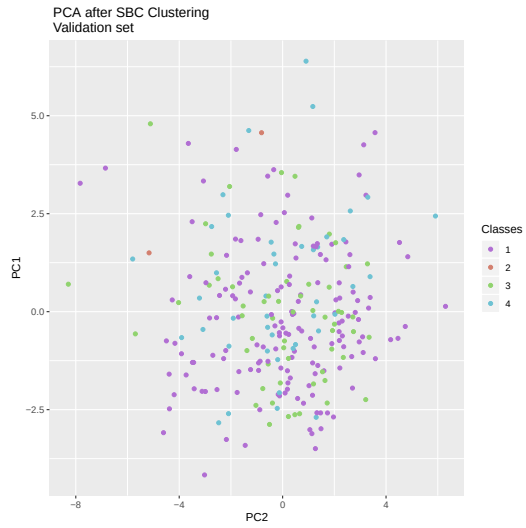
every model trained, therefore the Kaplan Meier curves for both scenarios remain the same.



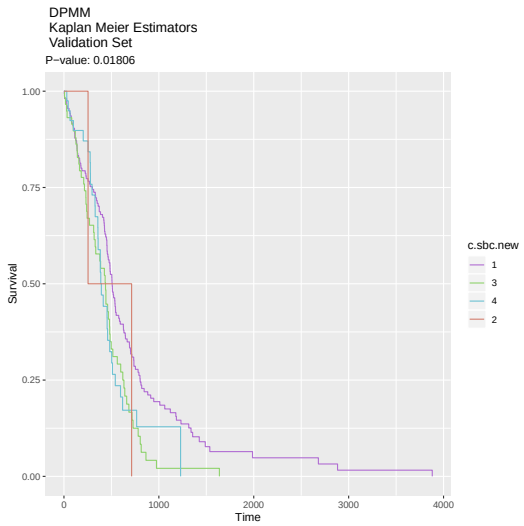
(a) PCA training set



(b) Kaplan Meier training set



(c) PCA validation set



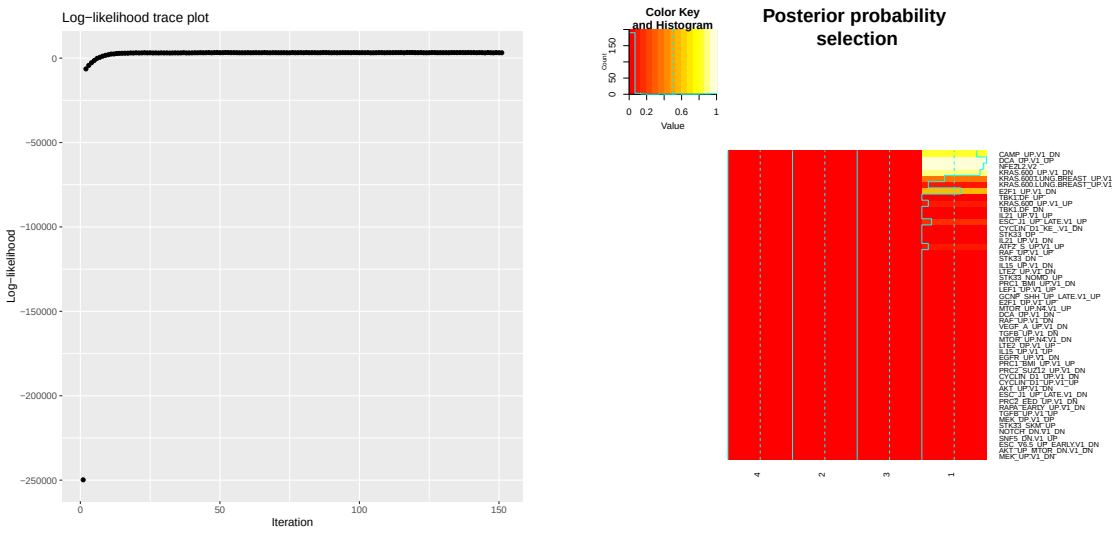
(d) Kaplan Meier validation set

Figure 4.10: Training and validation results of the SBC model after using PAFT for feature selection. Clinical end-point is overall survival and colours on each plot represent the different discovered (training) and predicted (validation) classes from the SBC

Figure 4.10 shows the summarised visualisations for training and validation sets of the PAFT feature selection model. The PCA plots are built on the PAFT linear predictors of the calculated SBC signature, which in this case consists of 50 oncogenic genes sets used as input features for model training. Besides, the

Kaplan Meier survival curves are built with the overall survival of the patients on each data set, using the discovered SBC labels. Figure 4.10(b) shows four distinct survival curves and Figure 4.10(a) shows four relatively separated clusters in the PCA for the training set. For the validation set the results are still significant on the survival curves as it can be seen in Figure 4.10(d) (log-rank p-value = 0.018), but the separation is again less stark. Similarly, the clusters in the PCA are also less clearly separated for the validation set, as it appears in Figure 4.10(c). When compared to 4.2, there is a bigger difference in the survival curves from the training data set after the PAFT than the HSIC-Lasso feature selection. In summary, Figure 4.10 shows that the SBC successfully finds clinically relevant cancer subtypes in both the training and validation sets, with distinct survival curves and its corresponding molecular profiles based on the aggregated score of the 50 oncogenic gene sets best correlated with survival.

Based on this initial results, it was determined that for the following analyses regarding the PAFT feature results section, cluster 1 will be considered as "Good Moderate", cluster 2 as "Worst", cluster 3 as "Best" and cluster 4 as "Bad moderate".



(a) Log-likelihood trace plot **(b)** Posterior probability

Figure 4.11: Log-likelihood trace plot during the burning and Posterior probability plot for the training after using PAFT for feature selection.

Log-likelihood trace plots such as the one in Figure 4.11(a) help assess the convergence of the MCMC sampler, which is evident in this case after roughly the 12^{th} iteration. When compared to Figure 4.3(a), the model converges faster after PAFT than after HSIC-Lasso feature selection. On the other hand, Figure

4.11(b) shows the posterior probability plot, where only the "Good Moderate" subgroup (cluster 1) has features related to survival. Once more, the small sample size of other clusters explains this result (see Table 4.4). Interestingly, the most populated cluster after PAFT feature selection is the one with "Good moderate" survival, whereas after the Block HSIC-Lasso feature selection the most populated cluster was the one with the "Worst" survival times.

SBC discovered/predicted class	**Training set**	**Validation set**	**Total**
"Good moderate" (Cluster 1)	105 (66%)	157 (60%)	262 (62%)
"Worst" (Cluster 2)	3 (2%)	2 (1%)	5 (1%)
"Best" (Cluster 3)	28 (17%)	59 (23%)	87 (21%)
"Bad moderate" (Cluster 4)	24 (15%)	43 (16%)	67 (16%)
Total	160	261	421

Table 4.4: Cluster composition in training and validation sets after PAFT feature selection. The percentages shown are calculated with the total of each column in the following manner: The "Good moderate" group has 105 patients in the training set and it is divided by the total number of individuals in the training set ($\frac{105}{160} * 100$)

4.2.2 SBC signature and model interpretation

In order to identify the genes that contribute the most to distinguish between different clusters of patients in the SBC, a feature importance plot using Equation 2.6 was built. Figure 4.12 shows molecular signature profile of the SBC clusters in the training set in a heatmap, where blocks with high values are coloured in blue and blocks with low values are coloured in red. Again, the SBC not only finds distinct subtypes of cancer diagnosed patients according to their survival times, but it also finds distinct molecular profiles (in the form of aggregated features) that have a corresponding survival curve. Interestingly the "Worst" and "Bad moderate" groups have low expression of the PAFT aggregated scores on oncogenic gene sets, whereas the "Best" subtype has a high expression. Furthermore, the molecular profile on based on the SBC signature seems more contrasting after the PAFT pathway-based feature engineering than after the Block HSIC-Lasso feature selection (see Figure 4.4).

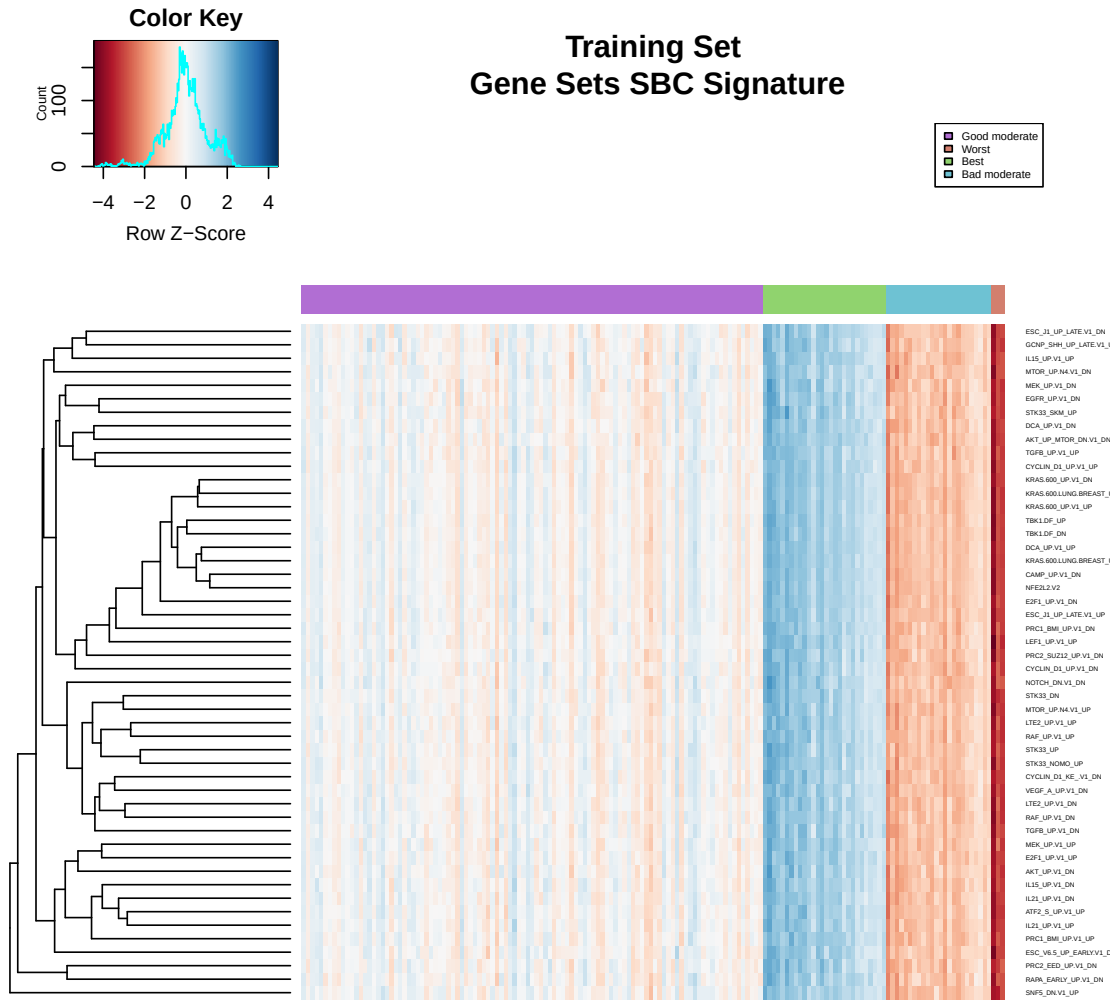


Figure 4.12: Heatmap of the SBC signature for the model trained after PAFT pathway-based feature engineering on the Oncogenic gene sets expression.

Once more, the gene sets that contribute the most to distinguish between different clusters of patients, are calculated with the Feature Importance measure (Equation 2.9) incorporated in the SBC, and produce the Feature Importance plot seen in Figure 4.13. Noticeably the oncogenic gene set **CAMP_UP.V1_DN** is one of the two most relevant features for the differentiation in all pairwise cluster comparisons. This set is that of down-regulated genes in primary thyrocyte cultures in response to cAMP signaling pathway activation by thyrotropin (TSH) (Staveren et al., 2006). It is reported in the literature that patients with GBM diagnosis are prone to metabolic changes such as hypothyroidism (Faghih-Jouybari

et al., 2018), moreover, glioblastoma cells are thyroid hormone dependent and it is known that mild hypothyroidism may improve survival of GBM patients (F. B. Davis et al., 2006).

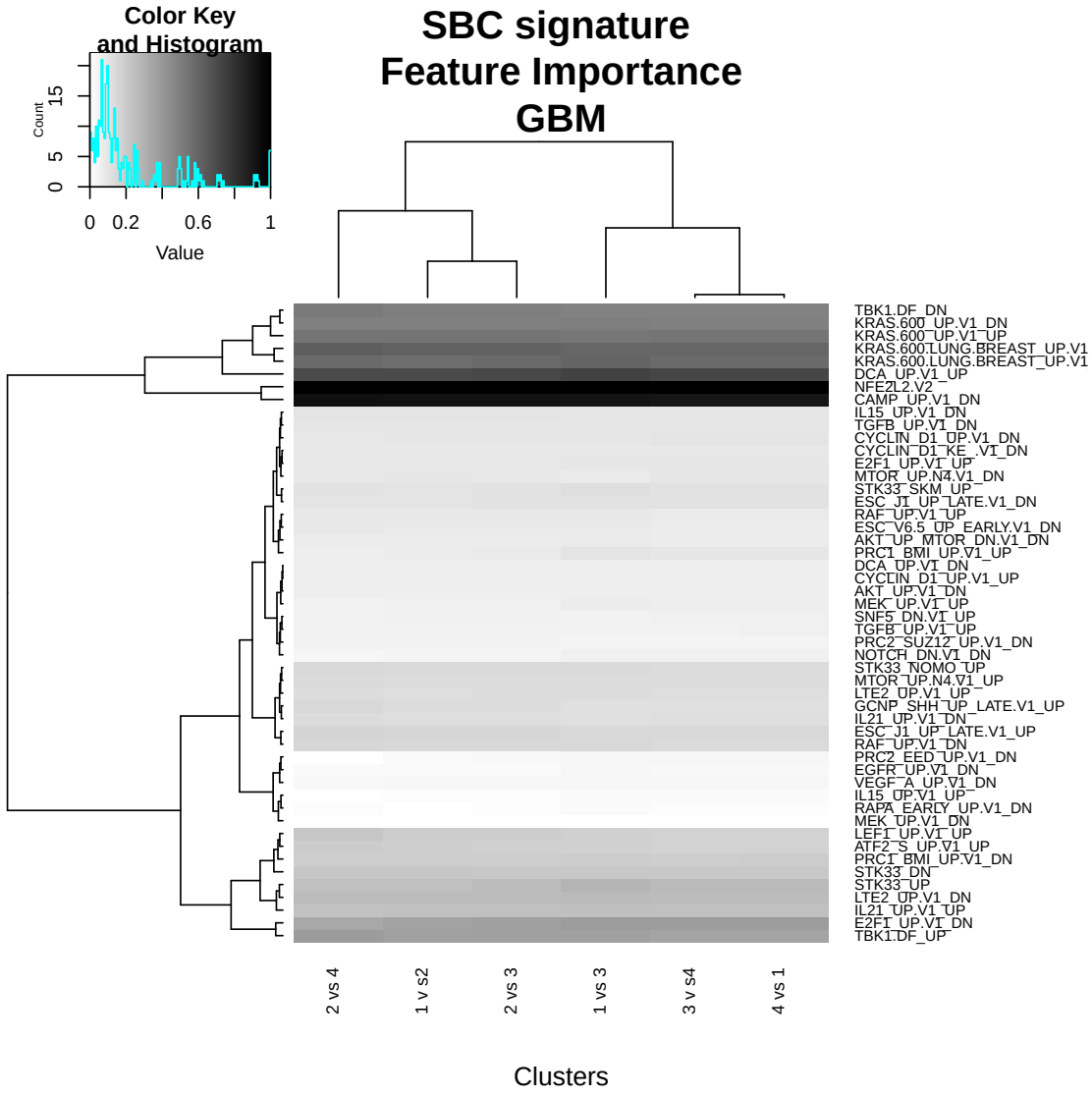


Figure 4.13: Feature importance in discriminating respective clusters of the SBC signature for the model trained after PAFT pathway-based feature engineering. Darker colours imply stronger effects.

The second most relevant gene set seen in Figure 4.13 is **NFE2L2.V2**. This set of genes are down-regulated after the knockout of the protein coding gene **NFE2L2**, which codes for the transcription factor **NRF2**. The latter is a transcriptor activator in charge of inducing a cytoprotective response to oxidative stresses and whose mutations provide constitutive activation in cancer (J. W. Kim

et al., 2016). This transcription factor plays a pivotal role in cancer survival and tumour growth (Fan et al., 2017; Ji et al., 2014; Seidel et al., 2010), and its up-regulation confers therapeutic resistance (Singer et al., 2015).

A third important oncogenic gene set that can bee seen in Figure 4.13, is **DCA_UP.V1_UP**, a group of genes up-regulated in Glioblastoma Multiforme cells treated with Dichloroacetate (DCA), an inhibitor of the mitochondrial pyruvate dehydrogenase kinase. DCA shifts metabolism from glycolysis to glucose oxidation and decreases mitochondrial membrane potential in cancer cells (Bonnet et al., 2007). This compound is a metabolic modulator that has been suggested as a viable therapeutic approach in the treatment of Glioblastoma Multiforme (Michelakis et al., 2008).

4.2.3 Characterisation of SBC clusters in other omics data sets

Just as it was done with the results after Block HSIC-Lasso mRNA-based feature selection, this subsection is intended to describe the possible repeated expression patterns in other types of molecular data which were not used as input for the SBC, in order to look for some sort of validation in other omics data types.

One possible association between the expression pattern of Copy Number Variation changes and the discovered SBC labels after PAFT feature selection can be visualised with the heatmap from Figure 4.14. **CAMK2G**, a gene widely associated with lifetime and progression-free survival (Boom et al., 2003; Network et al., 2008; Suzuki et al., 2004; Verhaak et al., 2010), also has significant correlation (FDR = 1.525e-11) between the discovered SBC subtypes in this model run and the CNV data. Interestingly, most of the CNV events correspond to insertions in the genome (blue colours) but they do not really form a clear block pattern with respect to the SBC labels.

The differential expression analysis on microarray mRNA and microarray miRNA data was done as described in section 3.3, in order to look for significant up-regulated and down-regulated features in two types of comparisons: cluster pairwise and among all discovered SBC subtypes. After doing the differential expression analysis on the mRNA data between discovered SBC cancer clusters with *(limma)*, **NCOR2** stands out for being deferentially expressed among all SBC labels (corrected p-value = 1.38e-248) and also being part of the Verhaak Gene Signature. This gene codifies for a transcriptional corepressor that promotes chromatin condensation and prevents transcription. Besides, it is targeted and silenced by the miRNA **miR-100**, which has been reported to be down-regulated in GBM patients (Alrfaei et al., 2013).

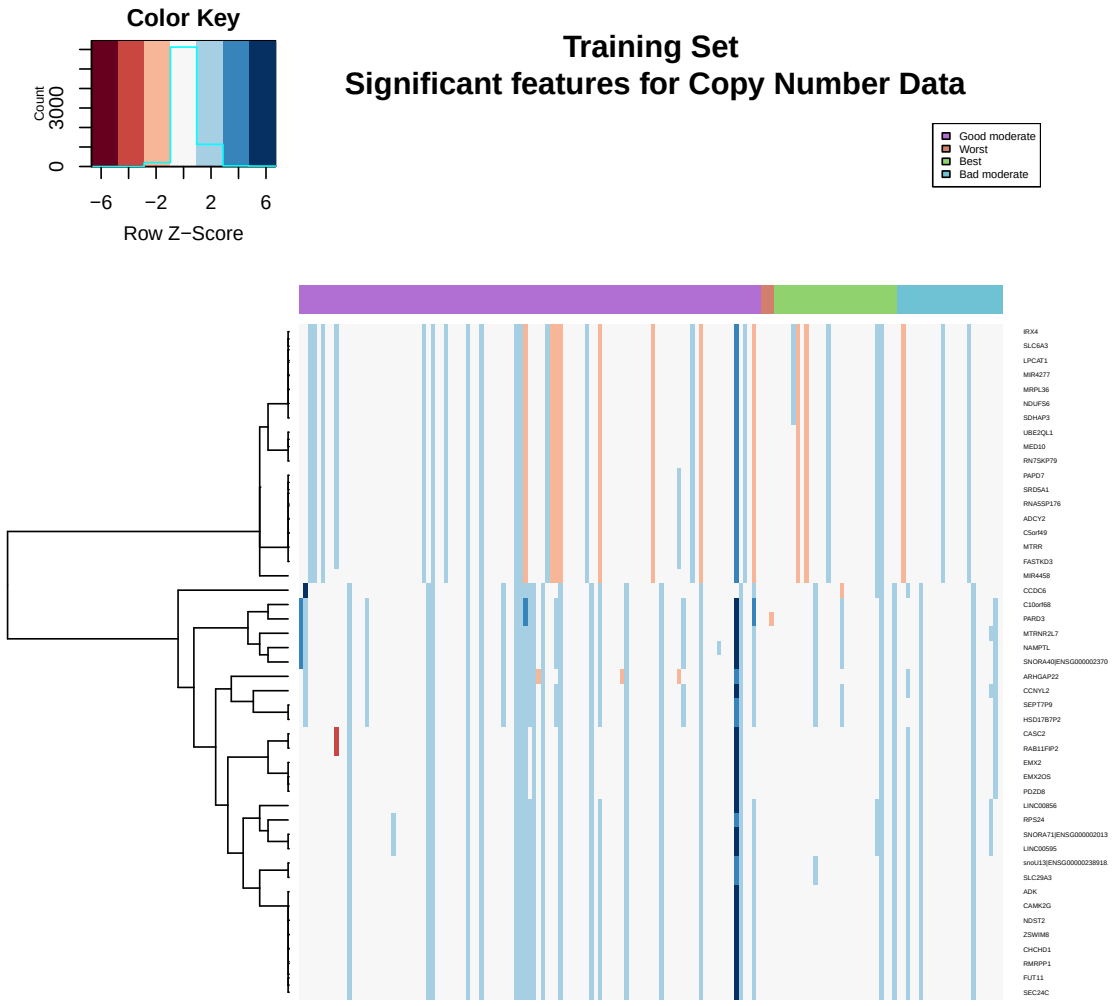
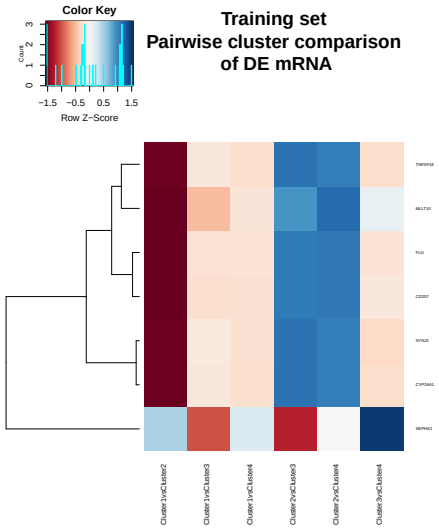


Figure 4.14: Copy Number Variation heatmap built with the SBC discovered labels after PAFT pathway-based feature engineering.

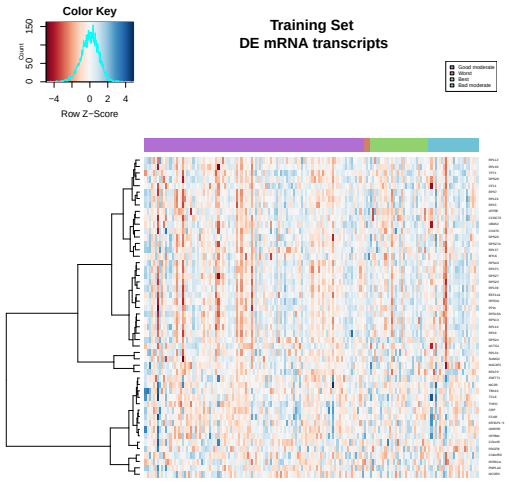
Figure 4.15 shows a distinct expression pattern for the mRNA when making pairwise comparisons in between discovered SBC clusters (Figure 4.15 (a)), as well as in the heatmap of the most differentially expressed genes between all clusters (Figure 4.15 (b)).

After analysing miRNA data, only **miR-31** and **miR-222** had a significant difference in expression in both pairwise cluster comparison and among all clusters. The former is reported as down-regulated in Glioblastoma Multiforme compared to normal brain tissues, and it affects cell migration and motility (Hua et al., 2012). The latter has already been reported as relevant in the miRNA GBM

landscape (Quintavalle et al., 2012) and appeared as differentially expressed with the Block HSIC-Lasso feature selection SBC model labels as well. There are no plots for those two significant miRNA transcripts since it did not make sense to do a heatmap of two features only.



(a) Heatmap of DE mRNA in cluster pairwise comparison



(b) Heatmap of DE mRNA per cluster

Figure 4.15: DE mRNA transcripts in pairwise cluster comparison and per cluster comparison using the discovered SBC labels. After performing a differential expression analysis with *(limma)* on the mRNA data set, two plots were obtained: (a) shows the log expression of the most significant mRNA transcripts per pairwise comparison between the SBC clusters, (b) shows the top 50 most significantly DE genes (FDR < 0.05). The SBC labels on the plots correspond to the model trained after PAFT feature selection

4.2.4 Characterisation of SBC clusters using other clinical data

Finally, when analysing other clinical features, only **age** (FDR= 4.0e-02) and **radiation therapy** (FDR = 2.5e-11) correlate with the SBC labels after pathway-based PAFT feature selection. Once more, the age of the patients whose data samples were part of the TCGA-GBM cohort, correlates with the survival-related clusters discovered by the model. This result is not surprising as survival time naturally correlates with how old a patient is.

Furthermore, the discovered labels correlate only slightly (p-value =0.072) with the Verhaak labels reported in the literature (Mesenchymal, Proneural, Neural and Classical).

5 Discussion

5.1 Summary

The Survival Based Bayesian Clustering algorithm has been successfully tested by using survival data along with molecular data to perform two main tasks: patient subtype identification on the training set and subtype prediction on the validation set. Both were evaluated in a simulated, a Breast cancer (Van De Vijver et al., 2002) and a Glioblastoma Multiforme (Verhaak et al., 2010) cohort of patients. After the work done in this thesis, it is possible to conclude that the SBC has proven to successfully replicate the results of the training set on an external data set, the TCGA-GBM, using the older Verhaak samples as training set and other samples in the TCGA-GBM as validation set.

In order to perform the SBC external validation, which was one of the two main objectives of this project, several challenges appeared: sample selection bias correction, selection of a method for SBC model initialisation and the involved process of feature engineering and selection. The bulk of work in this thesis was therefore devoted to coming up with a more informative feature set (SBC signature) which would provide satisfactory results on the training and also validation data sets.

The feature engineering and selection methods tested, further categorized in this thesis into pathway-based and mRNA-based, showed improved predictive and recovery C-index when compared with the results obtained in the model with mRNA and traditional SBC feature selection (which uses univariate Cox models). After careful evaluation of several alternatives to the feature selection problem, the two most successfully methods showed improvements in the Recovery C-Index (Block HSIC-Lasso feature selection = +8.7%, PAFT feature selection = +5.0%) and Predictive C-Index (Block HSIC-Lasso feature selection = +1.5%, PAFT feature selection = +27.6%)(see Table 5.1). Furthermore, both methods helped the SBC to find of distinct molecular clusters alongside their corresponding statistically significant different survival curves. As it is for any clustering and prediction algorithm, models are sensitive to the input features in which they are trained, and the rigorous feature engineering process assessed in this master thesis, was crucial to obtain results that validated the SBC in an external data set.

The process to implement the pathway-based feature engineering methods tested

Variable	Block HSIC-Lasso (mRNA-based feature selection)	PAFT (Pathway-based feature engineering)
P-value on the training set	1.078e-05	<5e-10
P-value on the validation set	4.701e-02	1.806e-02
Predictive C-Index	0.565 (+0.087)	0.528 (+0.05)
Recovery C-Index	0.698 (+ 0.015)	0.959 (+0.276)
Cluster composition	-2% "Worst". -15% "Bad moderate". -66% "Good moderate". -17% "Best".	-73% "Worst". -13% "Bad moderate". -9% "Good moderate". -6% "Best".
Significant SBC signature features	DYNLT3,PHF11,TAF5	Down-regulated genes after cAMP activation with TSH. Down-regulated genes after knockout of NFE2L2. Up-regulated genes after DCA treatment
Relevant CNV features	NET1	CAMK2G
Relevant mRNA transcripts	DYNLT3, ANKRD26, TAF5, SUPV3L1, NET1, TRIM48	NCOR2
Relevant miRNA	miR-222, miR-221, miR-34a, miR-9	miR-31, miR-222
Relevant clinical features	-Date of initial pathological diagnosis. -Age.	-Radiation therapy. -Age.

Table 5.1: Comparison of main results obtained with the two selected feature selection methods: Pathway-based PAFT and mRNA-based Block HSIC-Lasso. Values inside parenthesis correspond to the difference between the value obtained with the feature selection method from each column and the initial SBC model run without any feature selection (initial Predictive C-Index = 0.478, Initial Recovery C-Index = 0.683)

in this thesis was more convoluted than the mRNA-based feature selection, and it required the additional steps of selecting a relevant gene set collection as well as a feature aggregation method. These reasons made this path invariably longer when compared to the mRNA-based feature selection tested, the Block HSIC-Lasso, whose code is swiftly and easily evaluated in Python. Nonetheless, the creation of aggregated and biologically meaningful features yielded more distinct survival curves in both validation and training sets, a larger improvement in the Recovery C-Index when compared to the initial SBC implementation, and it allowed for a more extensive biological interpretation of the SBC clusters.

Another interesting result comes with the pathway-based feature selection methods using KEGG and Canonical pathways, that did not produce significantly distinct survival curves for the discovered subtypes and produced different cluster number for the training and the validation sets (see Table 4.1, Model 3 and Model 4 respectively). It is possible that by introducing an aggregated score based on those gene sets the survival signal is lost, i.e., the expression of those gene sets is unrelated to the clinical end-point data of the patients in the sample, and therefore the corresponding SBC signature in these cases failed to make good predictions.

Along the same line, a much lower p-value on the Kaplan Meier curves for the PAFT based feature selection (p-value < 5e-10) compared to the Block HSIC-Lasso based feature selection (p-value = 1.078e-05) is most certainly due to the fact that the former constructs those molecular features that correlate with survival than the latter, moreover the oncogenic gene sets already contained a previously validated cancer-related group of features with an underlying relation with survival and hence may have stronger feature sets with respect to survival than KEGG or the Canonical Pathways. This reasoning also explains the decision of not trying out the PAFT based feature aggregation method or the Penalised Cox feature aggregation method on a gene set collection other than the oncogenic. These gene sets (and their aggregated scores) might not necessarily be correlated with survival and the SBC would receive noise as input features, and consequently its performance would decrease.

5.2 Limitations and Future work

The problem of feature aggregation scores using pathway information has been addressed in different ways already reported in the literature, and four of the most prominent and cited methods were evaluated in this thesis (PLAGE, Z-Score, ssGSEA and GSVA), along with the implementation of two the most widely used survival analysis models (Penalized Accelerated Failure Time and Penalised Cox Proportional Hazard model) for feature engineering. As part of some exploratory results that are not presented in this thesis, a set of simple autoencoders were trained to obtain a latent representation of both the mRNA molecular data and the clinical endpoint data, nonetheless feature aggregation method using denoising autoencoders, requires a much rigorous hyper-parameter optimisation and building a network architecture that was outside of the scope of this master thesis due to time limitations. The latent representation of the high-dimensional data while preserving the survival signal to input to the SBC would be an interesting approach for future works.

The notorious unbalance in cluster composition both after Block HSIC-Lasso feature selection and PAFT feature selection might be due to a common problem present in the field of Machine Learning in Healthcare, namely that of small data sets. Despite the huge effort made by organisations such as the Broad Institute and the NCI, there are big difficulties in collecting enough data points in order to optimally train algorithms that are both able to detect the local patterns and make consistent predictions across patient cohorts, specially if there is a sample selection bias in the training set (Panch et al., 2019). Nonetheless, overcoming this problem

requires an ample amount of effort that can not only be solved algorithmically. Furthermore, it would be interesting to look into possible modifications to the SBC model itself, in a way that it forces the discovery of more homogeneous cluster sizes, taking into account that this form of cancer subtype discovery is meant to be clinically relevant.

Another interesting path to take in the future in order to improve the clinical relevance of the SBC clusters, is to include other clinical features such as the ones examined in this project in order to avoid discovering trivial clusters due to possible confounding factors. That is to say that it is ideal that the model in fact predicts cancer subtypes based on the molecular and survival data, and not on other features such as age of the patient or whether the patient received treatment or not.

Lastly, it is reassuring to corroborate how the cancer subtypes found by the SBC correlate with the reported literature on differential expression analyses done on GBM mRNA data, and that the expression patterns of other types of data sets such as miRNA and CNV vary among the SBC labels. This means the discovered subtypes are not artificial divisions of a sample into random subgroups, but they are in fact real subtypes with a clinical and biological underlying mechanism behind them, that correspond to distinct patterns in the omics data. To conclude, despite the aforementioned limitations, the SBC remains a viable method worth of further exploration when the objective is to retrieve clinically relevant subgroups of patients from a sample with survival and molecular data.

Bibliography

Ahmad, A. (2019). "Dissecting patient heterogeneity via statistical modeling based on multi-modal omics data". PhD thesis. Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn (cit. on pp. 4, 14, 17).

Ahmad, A. and H. Fröhlich (2017). "Towards clinically more relevant dissection of patient heterogeneity via survival-based Bayesian clustering". In: *Bioinformatics* 33.22, pp. 3558–3566 (cit. on pp. vii, 3, 6, 12, 13, 15, 16, 23, 24, 30, 37, 40).

Alrfaei, B. M., R. Vemuganti, and J. S. Kuo (2013). "microRNA-100 targets SMRT/NCOR2, reduces proliferation, and improves survival in glioblastoma animal models". In: *PLoS One* 8.11, e80865 (cit. on p. 56).

Bonnet, S., S. L. Archer, J. Allalunis-Turner, A. Haromy, C. Beaulieu, R. Thompson, C. T. Lee, G. D. Lopaschuk, L. Puttagunta, S. Bonnet, et al. (2007). "A mitochondria-K+ channel axis is suppressed in cancer and its normalization promotes apoptosis and inhibits cancer growth". In: *Cancer cell* 11.1, pp. 37–51 (cit. on p. 56).

Boom, J. van den, M. Wolter, R. Kuick, D. E. Misek, A. S. Youkilis, D. S. Wechsler, C. Sommer, G. Reifenberger, and S. M. Hanash (2003). "Characterization of gene expression profiles associated with glioma progression using oligonucleotide-based microarray analysis and real-time reverse transcription-polymerase chain reaction". In: *The American journal of pathology* 163.3, pp. 1033–1043 (cit. on p. 56).

Brennan, C. W., R. G. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S. R. Salama, S. Zheng, D. Chakravarty, J. Z. Sanborn, S. H. Berman, et al. (2013). "The somatic genomic landscape of glioblastoma". In: *Cell* 155.2, pp. 462–477 (cit. on p. 3).

Broad Institute (August 7 of 2019[a]). *C2 sub-collection CP: Canonical Pathways*. URL: `http://software.broadinstitute.org/gsea/msigdb/collection_details.jsp#C2` (cit. on p. 28).

— (August 7 of 2019[b]). *C2 sub-collection CP: Canonical Pathways*. URL: `http://software.broadinstitute.org/gsea/msigdb/collection_details.jsp#C6` (cit. on p. 28).

Chen, F. and S.-J. Hu (2012). "Effect of microRNA-34a in cell cycle, differentiation, and apoptosis: A review". In: *Journal of biochemical and molecular toxicology* 26.2, pp. 79–86 (cit. on p. 49).

Climente-González, H., C. A. Azencott, S. Kaski, M. Yamada, et al. (2019). "Block HSIC Lasso". In: (cit. on pp. 26, 27, 40).

Colaprico, A., T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, et al. (2015). "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data". In: *Nucleic acids research* 44.8, e71–e71 (cit. on pp. 9, 11).

Cox, D. R. (1972). "Regression models and life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2, pp. 187–202 (cit. on p. 30).

Davis, F. B., H.-Y. Tang, A. Shih, T. Keating, L. Lansing, A. Hercbergs, R. A. Fenstermaker, A. Mousa, S. A. Mousa, P. J. Davis, et al. (2006). "Acting via a cell surface receptor, thyroid hormone is a growth factor for glioma cells". In: *Cancer research* 66.14, pp. 7270–7275 (cit. on p. 55).

Ding, C. and H. Peng (2005). "Minimum redundancy feature selection from microarray gene expression data". In: *Journal of bioinformatics and computational biology* 3.02, pp. 185–205 (cit. on p. 26).

Faghih-Jouybari, M., S. Naderi, S. Mashayekhi, T. Padeganeh, and S. Abdollahzade (2018). "Hypothyroidism among patients with glioblastoma multiforme". In: *Iranian journal of neurology* 17.3, p. 149 (cit. on p. 54).

Fan, Z., A. Wirth, D. Chen, C. Wruck, M. Rauh, M. Buchfelder, and N. Savaskan (2017). "Nrf2-Keap1 pathway promotes cell proliferation and diminishes ferroptosis". In: *Oncogenesis* 6.8, e371 (cit. on p. 56).

Freire, P., M. Vilela, H. Deus, Y.-W. Kim, D. Koul, H. Colman, K. D. Aldape, O. Bogler, W. A. Yung, K. Coombes, et al. (2008). "Exploratory analysis of the copy number alterations in glioblastoma multiforme". In: *PloS one* 3.12, e4076 (cit. on p. 12).

Friedman, J., T. Hastie, and R. Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of statistical software* 33.1, p. 1 (cit. on p. 30).

Fröhlich, H., R. Balling, N. Beerenwinkel, O. Kohlbacher, S. Kumar, T. Lengauer, M. H. Maathuis, Y. Moreau, S. A. Murphy, T. M. Przytycka, et al. (2018). "From hype to reality: data science enabling personalized medicine". In: *BMC medicine* 16.1, p. 150 (cit. on p. 4).

Gerber, N. K., A. Goenka, S. Turcan, M. Reyngold, V. Makarov, K. Kannan, K. Beal, A. Omuro, Y. Yamada, P. Gutin, et al. (2014). "Transcriptional diversity of long-term glioblastoma survivors". In: *Neuro-oncology* 16.9, pp. 1186–1195 (cit. on p. 45).

Gershon, D. (2002). "Microarray technology: an array of opportunities". In: *Nature* 416.6883, pp. 885–892 (cit. on p. 11).

Görür, D. and C. E. Rasmussen (2010). "Dirichlet process gaussian mixture models: Choice of the base distribution". In: *Journal of Computer Science and Technology* 25.4, pp. 653–664 (cit. on p. 14).

Group, F.-N. B. W. et al. (2016). "BEST (Biomarkers, EndpointS, and other Tools) resource". In: (cit. on p. 4).

Guessous, F., Y. Zhang, A. Kofman, A. Catania, Y. Li, D. Schiff, B. Purow, and R. Abounader (2010). "microRNA-34a is tumor suppressive in brain tumors and glioma stem cells". In: *Cell cycle* 9.6, pp. 1031–1036 (cit. on p. 49).

Gui, J. and H. Li (2005). "Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data". In: *Bioinformatics* 21.13, pp. 3001–3008 (cit. on p. 30).

Guo, Y., Q. Sheng, J. Li, F. Ye, D. C. Samuels, and Y. Shyr (2013). "Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data". In: *PloS one* 8.8, e71462 (cit. on p. 11).

Hanif, F., K. Muzaffar, K. Perveen, S. M. Malhi, and S. U. Simjee (2017). "Glioblastoma multiforme: a review of its epidemiology and pathogenesis through clinical presentation and treatment". In: *Asian Pacific journal of cancer prevention: APJCP* 18.1, p. 3 (cit. on p. 5).

Hänzelmann, S., R. Castelo, and J. Guinney (2013). "GSVA: gene set variation analysis for microarray and RNA-seq data". In: *BMC bioinformatics* 14.1, p. 7 (cit. on pp. 28–30).

Hua, D., D. Ding, X. Han, W. Zhang, N. Zhao, G. Foltz, Q. Lan, Q. Huang, and B. Lin (2012). "Human miR-31 targets radixin and inhibits migration and invasion of glioma cells". In: *Oncology reports* 27.3, pp. 700–706 (cit. on p. 57).

Huang, J., A. Gretton, K. Borgwardt, B. Schölkopf, and A. J. Smola (2007). "Correcting sample selection bias by unlabeled data". In: *Advances in neural information processing systems*, pp. 601–608 (cit. on p. 21).

Jensen, M. A., V. Ferretti, R. L. Grossman, and L. M. Staudt (2017). "The NCI Genomic Data Commons as an engine for precision medicine". In: *Blood* 130.4, pp. 453–459 (cit. on p. 10).

Ji, X., H. Wang, J. Zhu, L. Zhu, H. Pan, W. Li, Y. Zhou, Z. Cong, F. Yan, and S. Chen (2014). "Knockdown of Nrf2 suppresses glioblastoma angiogenesis by inhibiting hypoxia-induced activation of HIF-1α". In: *International journal of cancer* 135.3, pp. 574–584 (cit. on p. 56).

Johnson, W. E., C. Li, and A. Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1, pp. 118–127 (cit. on p. 21).

Kim, J. W., O. B. Botvinnik, O. Abudayyeh, C. Birger, J. Rosenbluh, Y. Shrestha, M. E. Abazeed, P. S. Hammerman, D. DiCara, D. J. Konieczkowski, et al. (2016). "Characterizing genomic alterations in cancer by complementary functional associations". In: *Nature biotechnology* 34.5, p. 539 (cit. on p. 55).

Kim, T.-M., W. Huang, R. Park, P. J. Park, and M. D. Johnson (2011). "A developmental taxonomy of glioblastoma defined and maintained by MicroRNAs". In: *Cancer research* 71.9, pp. 3387–3399 (cit. on p. 49).

Kim, Y.-W., D. Koul, S. H. Kim, A. K. Lucio-Eterovic, P. R. Freire, J. Yao, J. Wang, J. S. Almeida, K. Aldape, and W. A. Yung (2013). "Identification of prognostic gene signatures of glioblastoma: a study based on TCGA data analysis". In: *Neuro-oncology* 15.7, pp. 829–839 (cit. on pp. 45, 48).

Kitano, H. (2002). "Computational systems biology". In: *Nature* 420.6912, p. 206 (cit. on p. 4).

Kuhn, M. and K. Johnson (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press (cit. on p. 24).

Law, C. W., Y. Chen, W. Shi, and G. K. Smyth (2014). "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". In: *Genome biology* 15.2, R29 (cit. on p. 33).

Le Tourneau, C., E. Borcoman, and M. Kamal (2019). "Molecular profiling in precision medicine oncology". In: *Nature medicine* 25.5, p. 711 (cit. on p. 4).

Lee, E., H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee (2008). "Inferring pathway activity toward precise disease classification". In: *PLoS computational biology* 4.11, e1000217 (cit. on p. 29).

Liberzon, A., A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov (2011). "Molecular signatures database (MSigDB) 3.0". In: *Bioinformatics* 27.12, pp. 1739–1740 (cit. on p. 27).

Louis, D. N., A. Perry, G. Reifenberger, A. Von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison (2016). "The 2016 World Health Organization classification of tumors of the central nervous system: a summary". In: *Acta neuropathologica* 131.6, pp. 803–820 (cit. on p. 5).

MacKay, D. J. and D. J. Mac Kay (2003). *Information theory, inference and learning algorithms*. Cambridge university press (cit. on p. 15).

Mathur, S. and J. Sutton (2017). "Personalized medicine could transform healthcare". In: *Biomedical reports* 7.1, pp. 3–5 (cit. on p. 4).

Mermel, C. H., S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz (2011). "GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers". In: *Genome biology* 12.4, R41 (cit. on p. 12).

Michelakis, E., L. Webster, and J. Mackey (2008). "Dichloroacetate (DCA) as a potential metabolic-targeting therapy for cancer". In: *British journal of cancer* 99.7, p. 989 (cit. on p. 56).

Mounir, M., M. Lucchetta, T. C. Silva, C. Olsen, G. Bontempi, X. Chen, H. Noushmehr, A. Colaprico, and E. Papaleo (2019). "New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx". In: *PLoS computational biology* 15.3, e1006701 (cit. on p. 9).

NCI (2019). *NCI Dictionary of Cancer Terms*. URL: `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/precision-medicine` (visited on 08/05/2019) (cit. on p. 4).

NCI, N. C. I. (March 6 of 2019). *TCGA Computational Tools*. URL: `https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/tools` (cit. on p. 10).

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models". In: *Journal of computational and graphical statistics* 9.2, pp. 249–265 (cit. on pp. 14, 15).

Network, C. G. A. R. et al. (2008). "Comprehensive genomic characterization defines human glioblastoma genes and core pathways". In: *Nature* 455.7216, p. 1061 (cit. on p. 56).

Omuro, A. and L. M. DeAngelis (2013). "Glioblastoma and other malignant gliomas: a clinical review". In: *Jama* 310.17, pp. 1842–1850 (cit. on p. 3).

Panch, T., H. Mattie, and L. A. Celi (2019). "The "inconvenient truth" about AI in healthcare". In: *npj Digital Medicine* 2.1. URL: `https://doi.org/10.1038/s41746-019-0155-4` (cit. on p. 61).

Park, J., J.-K. Shim, S.-J. Yoon, S. H. Kim, J. H. Chang, and S.-G. Kang (2019). "Transcriptome profiling-based identification of prognostic subtypes and multi-omics signatures of glioblastoma". In: *Scientific reports* 9.1, p. 10555 (cit. on p. 45).

Prados, M. D., S. A. Byron, N. L. Tran, J. J. Phillips, A. M. Molinaro, K. L. Ligon, P. Y. Wen, J. G. Kuhn, I. K. Mellinghoff, J. F. De Groot, et al. (2015). "Toward precision medicine in glioblastoma: the promise and the challenges". In: *Neuro-oncology* 17.8, pp. 1051–1063 (cit. on pp. 4, 5).

Purves, D., G. J. Augustine, D. Fitzpatrick, W. C. Hall, A.-S. LaMantia, J. O. McNamara, and L. White (2001). "Neuroscience. ed". In: *Glutamate. Sunderland (MA): Sinauer Associates* (cit. on p. 5).

Quintavalle, C., M. Garofalo, C. Zanca, G. Romano, M. Iaboni, M. d. B. De Caro, J. Martinez-Montero, M. Incoronato, G. Nuovo, C. Croce, et al. (2012). "miR-221/222 overexpession in human glioblastoma increases invasiveness by targeting the protein phosphate PTPμ". In: *Oncogene* 31.7, p. 858 (cit. on pp. 49, 58).

Ravikumar, P., J. Lafferty, H. Liu, and L. Wasserman (2009). "Sparse additive models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.5, pp. 1009–1030 (cit. on p. 26).

Samur, M. K. (2014). "RTCGAToolbox: a new tool for exporting TCGA Firehose data". In: *PloS one* 9.9, e106397 (cit. on pp. 10, 19).

Santarius, T., J. Shipley, D. Brewer, M. R. Stratton, and C. S. Cooper (2010). "A census of amplified and overexpressed human cancer genes". In: *Nature Reviews Cancer* 10.1, p. 59 (cit. on p. 12).

Seidel, S., B. K. Garvalov, V. Wirta, L. von Stechow, A. Schänzer, K. Meletis, M. Wolter, D. Sommerlad, A.-T. Henze, M. Nister, et al. (2010). "A hypoxic niche regulates glioblastoma stem cells through hypoxia inducible factor 2α". In: *Brain* 133.4, pp. 983–995 (cit. on p. 56).

Singer, E., J. Judkins, N. Salomonis, L. Matlaf, P. Soteropoulos, S. McAllister, and L. Soroceanu (2015). "Reactive oxygen species-mediated therapeutic response and resistance in glioblastoma". In: *Cell death & disease* 6.1, e1601 (cit. on p. 56).

Srinivasan, S., I. R. P. Patric, and K. Somasundaram (2011). "A ten-microRNA expression signature predicts survival in glioblastoma". In: *PloS one* 6.3, e17438 (cit. on p. 12).

Staveren, W. C. van, D. W. Solıs, L. Delys, D. Venet, M. Cappello, G. Andry, J. E. Dumont, F. Libert, V. Detours, and C. Maenhaut (2006). "Gene expression in human thyrocytes and autonomous adenomas reveals suppression of negative feedbacks in tumorigenesis". In: *Proceedings of the National Academy of Sciences* 103.2, pp. 413–418 (cit. on p. 54).

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550 (cit. on p. 29).

Sun, J., X. Gong, B. Purow, and Z. Zhao (2012). "Uncovering microRNA and transcription factor mediated regulatory networks in glioblastoma". In: *PLoS computational biology* 8.7, e1002488 (cit. on p. 49).

Suzuki, T., M. Maruno, K. Wada, N. Kagawa, Y. Fujimoto, N. Hashimoto, S. Izumoto, and T. Yoshimine (2004). "Genetic analysis of human glioblastomas using a genomic microarray system". In: *Brain tumor pathology* 21.1, pp. 27–34 (cit. on p. 56).

Tarca, A. L., G. Bhatti, and R. Romero (2013). "A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity". In: *PloS one* 8.11, e79217 (cit. on p. 29).

Tomczak, K., P. Czerwińska, and M. Wiznerowicz (2015). "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge". In: *Contemporary oncology* 19.1A, A68 (cit. on p. 9).

Tomfohr, J., J. Lu, and T. B. Kepler (2005). "Pathway level analysis of gene expression using singular value decomposition". In: *BMC bioinformatics* 6.1, p. 225 (cit. on p. 29).

Van De Vijver, M. J., Y. D. He, L. J. Van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, et al. (2002). "A gene-expression signature as a predictor of survival in breast cancer". In: *New England Journal of Medicine* 347.25, pp. 1999–2009 (cit. on pp. vii, 3, 5, 6, 59).

Verhaak, R. G., K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, et al. (2010). "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1". In: *Cancer cell* 17.1, pp. 98–110 (cit. on pp. vii, 3, 5–7, 46, 56, 59).

Wesseling, P. and D. Capper (2018). "WHO 2016 classification of gliomas". In: *Neuropathology and applied neurobiology* 44.2, pp. 139–150 (cit. on p. 5).

Xiang, Y., C.-Q. Zhang, and K. Huang (2012). "Predicting glioblastoma prognosis networks using weighted gene co-expression network analysis on TCGA data". In: *BMC bioinformatics*. Vol. 13. 2. BioMed Central, S12 (cit. on p. 46).

Yamada, M., W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama (2014). "High-dimensional feature selection by feature-wise kernelized lasso". In: *Neural computation* 26.1, pp. 185–207 (cit. on p. 26).

Yildirim, I. (2012). "Bayesian inference: Gibbs sampling". In: *Technical Note, University of Rochester* (cit. on p. 16).

Zarrei, M., J. R. MacDonald, D. Merico, and S. W. Scherer (2015). "A copy number variation map of the human genome". In: *Nature reviews genetics* 16.3, p. 172 (cit. on p. 12).

Glossary

2016 CNS WHO 2016 World Health Organization Classification of Tumor in the Central Nervous System. 5

AFT Accelerated Time Failure. 15, 16, 30, 41, 42

CNV Copy Number Variation. vii, xi, xii, 7, 12, 19, 21, 33, 46–48, 56, 57, 62

DCA Dichloroacetate. 56

DE Deferentially Expressed. xi, xii, 47–49, 58

DP Dirichlet Process. 14

DPMM Dirichlet Process Mixture Models. ix, 13, 14, 17

EHR Electronic Healthcare Records. 4

FDR False Discovery Rate. 33, 47–49, 56, 58

GBM Glioblastoma Multiforme. vii, ix, 3, 5–7, 11, 12, 45, 46, 49, 54–57, 59, 62

GDC Genomics Data Commons. 9–12, 19, 34

GISTIC Genomic Identification of Significant Targets in Cancer. 12

GSE Gene Set Enrichment. xi, 26, 28, 31, 32

GSEA Gene Set Enrichment Analysis. 29

GSVA Gene Set Variation Analysis. 29, 31, 61

HSIC Hilbert-Schmidt Independence Criterion. 26

MCMC Markov Chain Monte Carlo. 15–17, 32, 33, 41, 52

miRNA MicroRNA. vii, xi, 7, 9, 12, 19, 21, 33, 48, 49, 56–58, 62

mSigDB Molecular Signatures Database. 27, 28, 31, 49

NCI National Cancer Institute. 9–11, 61

NHGRI National Human Genome Research Institute. 9

PAFT Penalized Accelerated Failure Time. vii, x, xii, 13, 30, 31, 37–39, 49–59, 61

PCA Principal Component Analysis. xi, 20–23, 40–42, 50–52

PLAGE Pathway level Analysis of Gene Expression. 29, 31, 61

SBC Survival Based Bayesian Clustering. vii, ix–xii, 3–7, 12–17, 19, 21, 23–35, 37, 39–49, 51–62

ssGSEA Single Sample Gene Set Enrichment Analysis. 29, 31, 39, 61

SVD Singular Value Decomposition. 29

TCGA The Cancer Genome Atlas. ix, xi, 3, 9–12, 19–21, 29

TCGA-GBM The Cancer Genome Atlas Glioblastoma Multiforme. vii, xi, 3, 6, 19–21, 23–25, 28, 34, 35, 47, 48, 50, 58, 59

TSS Tissue Source Site. 20

WHO World Health Organisation. 5