Dissecting patient heterogeneity via statistical modeling based on multi-modal omics data

Dissertation zur Erlangung
des Doktorgrades (Dr. rer. nat.) der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Ashar Ahmad
aus Bihar, Indien

Bonn, September, 2018

Angefertigt mit Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität
Bonn

1.Gutachter: Prof. Dr. Holger Fröhlich
2.Gutachter: Prof. Dr. Andreas Weber
Tag der Promotion: 1 Februar, 2019
Erscheinungsjahr: 2019

Abstract

One of the key goals of modern medicine is to treat patients individually, recognizing the heterogeneity that exists within them and thus hoping to provide them with more effective personalized therapies. '-Omics' patient data provides a valuable resource to understand the patient heterogeneity and gain an insight into the biological phenomena at the intracellular level. As it is impossible to dissect patient groups based on single biomarkers or clinical factors, multivariate data mining and statistical modelling approaches (machine learning) play an important role. Moreover, understanding complex disease mechanisms calls for a more comprehensive and integrative approach, hence motivating the use of different kinds of data from the same patient. As individual -omics data sources capture specific kinds of molecular phenomena, there is a pressing need for multi-modal statistical approaches which combine several kinds of -omics data together.

The present thesis addresses the aforementioned issues, viz. the exploration of heterogeneous patient populations based on their multi -omics profiles using statistical and machine learning approaches. More specifically, the main contributions of the thesis include: a) a retrospectively validated prediction model for GBM (Glioblastoma Multiforme) recurrence location and b) development of a new algorithm- Survival based Bayesian Clustering which is a merger of clustering and supervised prediction, this algorithm has been successfully shown to be an important step towards the discovery of clinically relevant patient strata leveraging the potential of multi-omics data integration.

The novel algorithm of Survival based Bayesian Clustering was tested successfully in various scenarios and on different patient populations. The thesis also provides a deep understanding of our proposed technique from a purely statistical point of view. Overall, work in this thesis is a step forward in moving towards the goal of personalized medicine solutions using multi-modal molecular -omics data and statistical modelling.

Contents

Listing of figures

List of Tables

To the loving memory of my father.

Acknowledgments

First and foremost I would like to thank Prof. Dr. Fröhlich as my principal supervisor. I am very glad that I had the opportunity to work in his lab during the course of my PhD. Scientifically, his supervision was incredibly helpful to achieve the goals of this thesis. I am also grateful to him for the inspiring and motivating discussions during difficult times.

I would also like to thank Prof. Dr. Scheffler and Prof. Dr. Glas for the collaboration and cooperation during the IDENTIREST project. This was a very stimulating project to work on and I learnt a great deal through it.

Next, I would like to thank Prof. Dr. Weber for the financial support during the later half of my PhD which allowed me the necessary time and resources to finish this work.

I would also like to thank the group members of our lab, both past and present. They have been wonderful and supportive colleagues to have.

I would like to also thank my mother, father and brother whose support was crucial for me for my PhD studies. My father had always been incredibly supportive of me and his loss was a big setback. I dedicate my success to him and to the endless love that my parents have have shown to me.

Lastly, I would like to thank my wife and her family as well. My wife has always been there for me and she has been pivotal to my success. She has helped me through many difficult times and has been incredibly patient with her love throughout my PhD journey.

"The fundament upon which all our knowledge and learning rests is the inexplicable."

Arthur Schopenhauer

1

Introduction

1.1 Patient Stratification in Cancer

Modern medicine aims at providing a much more personalized treatment of an individual which is tailored to his/her personal characteristics. This approach is fundamentally different from traditional medical practice which is based on the idea of reference treatments. These reference treatments are 'canonical' treatments which have been established based on

large series of patients and are considered universal solutions for treating new patients[GW08] . Diseases such as cancers are known to be highly heterogeneous and may designate, in fact, a myriad of different diseases, each with its own trajectory.

It has been shown that cancer can be characterized by complicated accumulation of genetic and epigenetic alterations thus leading to a lot of heterogeneity within itself[BCH+12]. The idea of patient stratification in this context is to acknowledge that the patient pathology is unique and this uniqueness is driving the choice of treatment. Such a personalized treatment (and hence strata of patients) depends on patient's constitutional genetic background as well as tumor's genetic and epigenetic landscape.

One can easily realize the importance of acknowledging the diversity within patient population in cancer. This diversity and hence stratification is important to:

- Design separate clinical/medical treatment protocols for different strata of patient population
- Better understand the complicated set of molecular alterations within the sub-populations and hence develop stratum- specific drugs.

The second point leads us to defining Biomarkers and its application in Personalized Medicine.

1.2 Personalized Medicine & Biomarker Discovery

Patient stratification is the first step towards individualized treatment also know as Personalized, precision, P4, or stratified medicine. A concrete definition of Personalized Medicine was adopted by EU Health Ministers in their Council as follows: "a medical model using characterization of individuals' phenotypes and genotypes (e.g. molecular profiling, medi-

cal imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention".*

Personalized medicine can be understood as an approach in which patients are stratified based on their predicted disease subtype, disease risk, disease prognosis or treatment response using diagnostic tests[BFM18]. It is a rapidly advancing field of health care which is informed by each person's unique genetic, clinical and genomic information. A key promise of personalized medicine is a much closer molecular understanding of disease to optimize preventive/diagnostic/prognostic health care strategies and drug therapies. As the aforementioned factors are different for every person hence the nature of disease, its onset, its course, and how it might respond to drug or other interventions are as individual as the people who suffer from them[FBB+18]. Briefly, personalized medicine allows the following practical advantages †:

1. Better and informed clinical decision making and disease management
2. Better-targeted therapies that will result in higher desirable outcomes
3. Reduced ill-effects from targeted-therapies.
4. Earlier disease detection and possible disease prevention
5. Reduced health care costs.

*The document can be accessed at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:C:2015:421:FULL&from=EN

†EU commission report on *'The use of '-omics' technologies in the development of personalized medicine'* can be accessed at: https://ec.europa.eu/research/health/pdf/2013-10_personalised_medicine_en.pdf

6. Smarter design of clinical trials due to selection of likely responders at baseline

Personalized or precision medicine stratifies patients into groups based on many factors which include the biological make-up or biomarkers. A formal definition of a biomarker was provided by the National Institutes of Health Biomarkers Definitions Working Group as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention."[Gro01]. The US Food and Drug Administration (FDA) defines the term biomarker as any measurable quantity or score that can be used as a basis to stratify patients, e.g., genomic alterations, molecular markers, disease severity scores, lifestyle characteristics etc.[BFM18].

From a clinical application point of view biomarkers may potentially be used to predict clinical responses to treatments, and in some cases they may represent potential drug targets. Biomarkers in clinical research can be obtained from solid tissues and bio-fluids. Various kinds of biomarkers have been used in clinical practice to detect diseases and predict clinical outcomes. The FDA lists more than 160 pharmocogenomic biomarkers and biomarker signatures which have been used for stratifying patients for drug response[MS17]. Personalized medicine and Biomarker identification are tightly interconnected. It is important to mention that in many cases, it's difficult to identify single biomarkers which can stratify patients. This is due to the complex nature of diseases (especially cancer) which often involves an interplay of many different biological subsystems. Also, drugs for treating diseases are multi-faceted and hence single biomarkers fail to describe their effect. Multivariate biomarker signatures are therefore seen as promising solutions, a highly cited example is MammaPrint[VVDVDV+02] , a 70-gene signature for predicting breast cancer prognosis.

Discovery of such multi-dimensional signatures require advanced statistical models and machine learning methods applied on -omics data sets (genomics, transcriptomics, epigenomics, proteomics, metabolomics). Recently, bio-imaging data (MRT and CT)[MCM+06], electronic medical records (EMRs)[JJB12] have also been used for biomarker discovery.

1.3 Multi Modal Data & Statistical Modelling

The ambitions of personalized medicine and biomarker discovery have been tremendously boosted by the ever growing data availability and the generation of large volumes of high-throughput '-omics' data capturing large scale biology (genome, proteome, transcriptome). A more comprehensive definition of -omics data is provided in the next chapter. We also have (in many cases) large volumes of clinical longitudinal data for patients from Electronic Medical Records (EMR). These various different kinds of data can be referred to as multi-modal (omics, clinical) data as there are different modalities giving rise to these data. This multi-modal data is critical to the goals of personalized medicine and biomarker discovery as it contributes to a much finer understanding of disease at different levels, this in turn might lead to the identification of new biomarkers which might be predictive of the development of a disease, disease prognosis or medicine response or as targets for new treatments. During the course of our present work we focus on the data sources originating from different '-omics' approaches, such as genomic variation and mRNA expression analysis plus the clinical data such as time-to-event.

Statistical Models are tools to analyze this multi-modal data. As the data generated from these multi-modal data is highly multi-dimensional it requires multivariate statistical models. Pre-processing and appropriate normalization methods first prepare the raw multi-

modal data for further application of statistical learning models (also know as machine learning models). The goal of these statistical learning/ machine learning models is to gain insight into the complex structure present within the data and to provide accurate predictions. This essentially involves separating the interesting signals from the noise present in such data. Statistical models capture the statistical dependencies, such as correlation from the data and allow for more comprehensive approaches in Personalized Medicine/Biomarker discovery.

Multi-modal data is widely believed to provide unique opportunities for Personalized Medicine as it allows for capturing and understanding different dimensions of a patient. This aspect could in turn be key for enhancing prediction performance of patient stratification statistical models to a level useful for clinical practice.[BFM18]. There is also the benefit of a deeper understanding of disease mechanisms: recent progress in the investigation of independent 'omic' resources has shown that there is a possibility that the molecular profiles or patterns observed in the potential biomarkers may not be true reflections of primary molecular events which initiate or modulate a disease[BCH+12]. This occurs because disease development is a highly complicated consequence of interplay of different bimolecular pathways. This complicated association of the different large scale biological processes with one another calls for a more holistic and integrative approach. Such an approach not only concentrates at one data source but integrates multi-source data (e.g. multi-omics data sets). The integration of different modalities of clinical and multi-omics data also motivates the present work. In this thesis we refer to multi-modal data in the context of multi-modal -omics data, i.e. data representing different biological modalities and focus on integrating this multi-modal omics data with clinical outcome. We show that such an integrated ap-

proach leads to the development of advanced exploratory/predictive approaches for patient stratification and biomarker discovery.

1.4 Glioblastoma: Case Study in Patient Stratification

Glioblastoma (GBM) is a grade IV astrocytoma, which is the most common primary adult brain tumor. GBM is a fast-growing and most aggressive type of central nervous system tumor[Som17]. The last decade has seen many influential high throughput microarray studies[PKC+06, VHP+10b] on GBM. This line of research has been quite successful to understand GBM more comprehensively. Although the median survival remains low despite advances on many aspects, due to this work there is an increasing understanding of many aspects of this extremely complex disease. In that context, The Cancer Genome Atlas[MFB+08] has been tremendously beneficial for identifying GBM molecular subtypes along with several gene signatures that have been proposed for risk stratification in GBM patients.

One of the key characteristic features of Glioblastoma is that this tumor varies widely in its composition with both inter and intra tumor heterogeneity. Prior to the advent of technologies to decipher the molecular makeup of glioblastoma, histopathology was the only modality available to characterize the tumor and its variants. The World Health Organization in its 2016 classification, characterizes glioblastoma in two major groups prognostic groups based on mutational status of IDH (Isocitrate dehydrogenase) gene-IDH wild type glioblastoma and IDH mutant glioblastoma. It is believed that IDH mutant glioblastoma confers a significantly better prognosis than IDH wild type glioblastoma[Som17]. Other key molecular biomarkers for this disease include: Losses of Chromosomal Arms 1p and 19q, ATRX mutations, TP53 mutations, TERT mutations, EGFR amplifications, PTEN and

MGMT mutations. It has also been established for nearly a decade that Epigenetic silencing of MGMT gene through its promoter methylation results in better response to Temozolomide and it was the most advocated prognostic marker in glioblastoma[Som17]. Glioblastoma was also among the first cancers to be targeted by large scale molecular profiling platforms like comparative genomic hybridization (CGH), single nucleotide polymorphism (SNP) arrays among other omics data types.

In 2006, Phillips et al.[PKC+06] classified Glioblastoma into three subgroups- Proneural, Proliferative and Mesenchymal with the Proneural group showing best prognosis. They used Olig2, DLL3, BCAN(Proneural), PCNA, TOP2A(proliferative), YKL-40, CD44 and VEGF (mesenchymal) as biomarkers to identify their subtypes. In a later work in 2008, Parson et al.[PJZ+08] brought the vital molecule IDH1 (Isocitrate dehydrogenase 1)to the forefront. In later years, Verhaak et al.[VHP+10b] radically altered the molecular classification and identified four subgroups namely: Proneural, Neural, Classical and Mesenchymal. They focused on alterations in PDGFRA, IDH1, EGFR, and NF1 and further highlighted the importance of IDH1 mutation which was seen in the Proneural group predominantly. Though both these groups of scientists (Phillips et al. and Verhaak et al) used distinct methodologies and sample sets, the proneural and mesenchymal groups were robustly concordant in their molecular profiling[DRR+13].

Apart from large heterogeneity, Gioblastoma is also notorious for its inevitable recurrence after maximal safe resection in spite of concomitant radiation and chemotherapy following surgery. The recurrent tumor tends to come back in a more aggressive manner which is more resistant to therapy. Presently, there is no accepted standard therapy for recurrent glioblastoma. Also, there have not been many studies focusing on recurrent

glioblastoma. One major reason for this is that not all recurrent tumors are operable, hence limiting the access to the recurrent tumor tissue. Therefore the scarcity of paired tumor samples precludes any molecular studies on recurrent glioblastoma. Most recurrences are predominantly local (recurring within 2 cm margin of the original tumor) with only a small proportion coming back as distant recurrences (recurring distantly in a different lobe or in contra-lateral hemisphere)[GRS+10]. The genetic makeup of the local recurrences when compared to their its primary counterpart still remains largely unknown.

1.5 IDENTIREST project

In this thesis we develop statistical models for our own -omics data from Glioblastoma patients from the IDENTIREST (Identifizierung neuer Therapieansätze durch Analyse Residualer Tumorzellen) project, funded by the German Federal Ministry of Research and Technology (BMBF) from 2013 till 2016. The goal of the project was to find new candidate drug targets and prognostic biomarker signature for glioblastoma. With our new statistical method (Survival based Bayesian Clustering and its subsequent variations) we identify new patient strata in this cohort with significantly different prognosis.

In this project, two kinds of samples are available for each GBM patient: a)the routinely removed and conventionally studied GBM cells (known also as central samples) from the centre of the tumor resection site and b) Residual cells (known also as peripheral samples) obtained from the periphery of the tumor resection site. The genetic landscape of local recurrences (peripheral samples) was initially thought be similar to the original tumor (central samples). The IDENTIREST project strives to clarify the status of these local recurrence samples, thus gaining molecular insights into the samples from spatially adjacent

areas(residual) of the primary tumor. The goal of the IDENTIREST project was the characterization of these residual cells with the aim to identify new drug targets.

The project builds on the past work of[GRS+10] which concentrated on the isolation and initial characterization of vital residual tumor cells of GBMs. We explore the heterogeneity of the central as well as the residual samples via our statistical models. A further unique contribution of our work is a classification model which allows us to predict with high accuracy the tumour recurrence, hence opening a perspective for preventive treatment of recurrence in GBM patients.

1.6 Thesis Contributions

The aforementioned challenges are addressed in this present thesis. The core contributions of our work are two fold: a) development of classification model in the context of the IDENTIREST project where we successfully predict the spatial recurrence of GBM, and b) formulation of a novel statistical model which uses multi-modal omics data along with clinical outcome data for patient stratification. The development of our statistical model is motivated by the question of patient stratification in Gliobalstoma. From a technical point of view, our Survival based Bayesian Clustering model (or SBC) can be considered as a combination of supervised and unsupervised approaches. The SBC which takes in clinical end-points of patients along with heterogeneous -omics data, performs two tasks in one - a) patient sub-group identification on training data and b) prediction of patient sub-group and survival time on testing data. It's based on the motivation of discovering clusters of patients using their distinct molecular signatures and strong survival curve separability[AFI7]. The plausibility of our SBC approach also lies in the biological interpretability of our re-

sults. The whole thesis is structured as follows:

- Chapter 2 introduces the broad field of -omics data sets and their use in personalized medicine, it talks about the challenges of using single -omics data sets in patient stratification, builds the motivation and application of integrated multi-omics in patient stratification. The chapter also provides a qualitative overview of the various statistical and machine learning methods and approaches that have been devised for the use of multi-omics data sets in patient stratification.

- Chapter 3 talks about some of the practical applications of machine learning methods to personalized medicine in glioblastoma treatment in the context of the IDENTIREST project. It introduces the various facets of the project, including the different omics data sets generated and the potential questions that were answered using different machine learning methods. The last section of this chapter is about the vital contribution in the development of classification model which predicts the spatial recurrence of Glioblastoma in the IDENTIREST cohort of samples. This model is further validated on an independent data set, thus promising to be of importance from a clinical point of view.

- Chapter 4 is the most important part of the thesis from a methodological point of view. It contains the novel statistical model (SBC) along with the motivation and application of the same. Many different questions that can be answered using this model are explored and results on two publicly available cancer data sets are provided. The results and biomarkers are also explored further for their biological relevance. Overall, this chapter introduces a novel machine learning method in patient

stratification and establishes the clinical and biological veracity of the results.

- Chapter 5 gives examples of the application of our SBC approach in Glioblastoma patient stratification. Two different variations of SBC are used to tackle different questions about the glioblastoma data set.

- Chapter 6 elucidates the statistical foundations of SBC and its relationship to other similar methods. The chapter situates SBC as a broad machine learning algorithm which can be seen from different points of views. It explains the similarities and differences with other popular machine learning/statistical methods.

- Conclusion draws the whole thesis to an end summarizing the core message of patient stratification and the use of SBC for this purpose. It also talks about the scientific accomplishment of the present thesis and its limitations along with future directions for research in multi-omics data for personalized medicine.

"Every answer given on principle of experience begets a fresh question."

Immanuel Kant

2

Machine Learning approaches to patient stratification using multi-modal omics data

2.1 Introduction to Omics Data

As discussed in the last chapter, one of the goals of patient stratification is providing personalized treatment. This involves identifying new translational targets in nucleic acid char-

acterizations. In cancer, we know that a series of events occurring at the cellular level disrupt the normal behavior of the cell[BCH+12]. Currently, it is believed that cancers always originate from genomic or epigenomic aberrations. However, consequences of these aberrations can manifest at different biological levels, namely transcriptome, proteome, metabolome etc. Moreover, the interactome between the proteins is probably also affected. Additionally, these changes affect how other kinds of molecules interact with each other, e.g., interactions between transcription factor and DNA. We need biotechnology tools which allow us to better understand tumour progression and improve the classification of tumours. Thus, there is a need to comprehensively quantify the aforementioned changes which occur at different molecular levels. Current -omics biotechnologies enable us to accurately characterize these molecular profiles of each tumor sample: Genomics investigates the DNA alterations (mutation, copy number), miRNomics the microRNA (miRNA) expression, transcriptomics the mRNA expression, proteomics the different proteins, epigenomics the epigenetic modifications like methylation and so on[BCH+12] . A full graphical description of all the different omics technologies can be seen in Fig.2.1.

In the past two decades the advent of omics data sets including genomics, transcriptomics, methylomics and proteomics, has created a huge source of cellular information on the one hand and stimulated parallel developments in statistical methodology and inference, computational tools on the other. Within the context of genomic medicine in cancer research, we focus our attention to different -omics data sources used for personalized medicine along the integration of this high-throughput data from multiple platforms to inform our understanding of the functional consequences of genomic alterations. We now briefly touch upon the technologies used for measuring omics data sets.

The first successful attempt at sequencing DNA was made by[SNC77] and facilitated the full sequencing of both genes and entire genomes. In spite of the fact the method was resource-intensive Sanger sequencing remained the standard method for the coming two decades. The method since its inception has gone through many refinements and technical advancements which increased its efficiency and reliability during the next three decades. Still, Sanger sequencing required large investments. A breakthrough came just before the turn of the century with the emergence of microarrays in the market that lessened run time and could be operated more easily with fewer human resources. Hence, microarray plates started replacing the labor-intensive Sanger method in the mid-1990s. The microarray technology rapidly became the default method to assess the expression of virtually all genes. The microarray technology, since its inception has been adapted in many different forms:

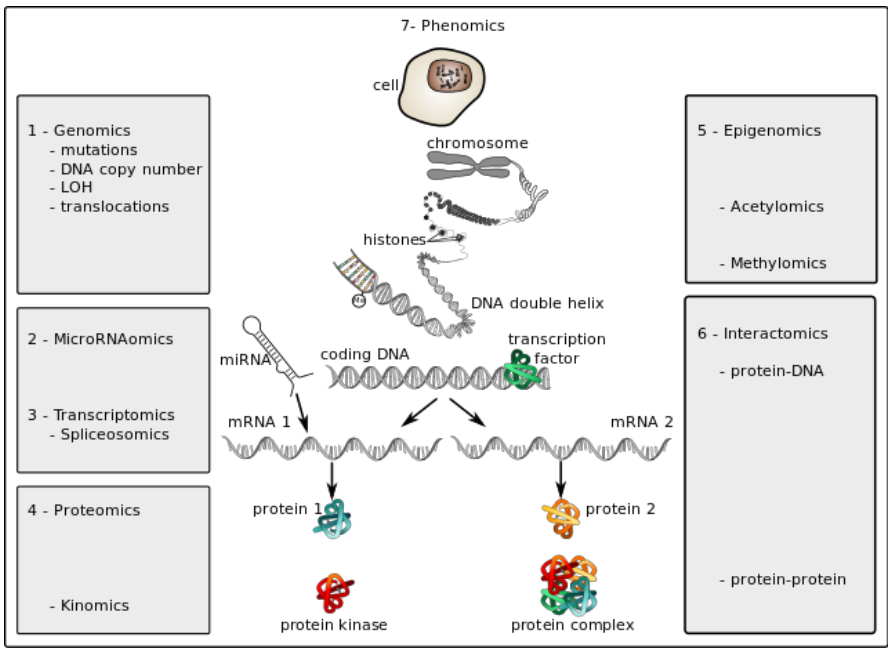

Figure 2.1: Illustrated diagram for different omics data sets. Image reproduced from[BCH+12] under free copy license CC-BY-SA.

genome microarrays like array Comparative Genomic Hybridisation (aCGH) or Single Nucleotide Polymorphism (SNP) arrays which investigate genomic alterations like gain, deletion and point mutations (for SNP arrays); transcriptome arrays for quantifying RNA expression at the the exon level or the transcript level, or microRNA (miRNA) expression; proteome arrays which interrogate protein expression and activities; Chromatin Immunoprecipitation (ChIP) arrays for localizing on the genome protein-DNA interactions or investigating nucleosome modifications[SSD+95].

The microarray technology provides measurements with the ability to quantify the genome-wide expression of thousands of gene on a tiny slide. Generally, for applications in personalized medicine, concentration of the mRNA or SNP array are measured across a range of samples originating from diseased and healthy patients. This provides concentration of a particular gene's transcript at a discrete point in time and differences in expression of the same gene across many samples could provide information to disease specific biomarkers.

Since the mid 2000s, Next-generation sequencing (NGS) has been becoming increasingly popular, NGS allows for simultaneous sequencing of millions of DNA fragments without previous sequence knowledge. NGS is also capable of looking through the entire genome or transcriptome. Since their inception NGS technologies are becoming more and more economically and technically viable, thus increasing in their popularity and often replacing microarrays as high throughput omics data sources for DNA (DNA-Seq), RNA (RNA-Seq) and proteins (ChIP-seq). However, in the context of this thesis only omics data measured via microarray technologies was used (largely due to cost considerations).

2.2 Omics Data in Personalized Medicine

The above mentioned omics biotechnology (producing high-throughput data) has promoted our understanding of the molecular nature of tumors, thus helping us unravel the genetic variations at different molecular levels. Many diseases, like cancer are known to be caused by genetic mutations and hence omics data provide a rich source of evidence to identify these mutations along with the possible mechanisms underlying such mutations. Human cancers are primarily genetic diseases and they can often be characterized by the following molecular changes: DNA sequence changes, copy number aberrations, chromosomal rearrangements and epigenetic modifications such as DNA methylation[BCH+12]. These upstream effects on the DNA can be captured using DNA microarray or DNA methylation arrays. Later or downstream modifications can be understood using mRNA or miRNA and reverse-phase protein microarrays. The overall disease is a complex combination of the aforementioned malignant and trans-formative changes.[MFB+08]. The high-throughput technologies allow for screening of massive amounts of omics-type data. In order to discover clinically relevant molecular markers (also known as biomarkers) one needs to associate such data with a patient's clinical prognosis or with the membership to a clinically relevant disease subtype (positive drug response class vs negative drug response class)[CFI2a].

Traditionally, within each disease (like cancer) patients are usually stratified into subcategories based on clinical information gathered from the patient, such as his/her age and possible previous cases of cancers in his/her family, and about the tumor, such as its size, location or histological type under the microscope. Generally, these informations are collectively referred to as clinicopathological parameters. Such stratification of patients is of

great value in clinical management. Also, for most cancers, guidelines exist to suggest the best therapeutic choices based on these stratification. For example, in Breast Cancer in addition to the histological parameters, the presence of specific markers, such as estrogen (ER), progesterone (PR) and human epidermal growth factor (HER2) receptors, is evaluated by immunohistochemical methods. Aggregated, these clinicopathological parameters currently determine the choice of the therapy proposed to the patient. In spite of the enormous help of such stratification technique for patient management, this approach is limited. Firstly, the consistent and objective assessment of some clinicopathological factors is difficult to make sure. This means that it might not only vary with the particular histological section being studied, but also might depend on the expert analyzing the sample. Secondly, this coarse classification fails to identify many differences between patients that are important for therapeutic treatment and monitoring. It has been known that tumors with similar clinicopathological parameters frequently follow different clinical courses or respond differently to therapies, hinting at the fact that a further level of variability exists within clinicopathological subtypes. Thirdly, clinicopathological parameters do not take into account the molecular differences, which likely have a tremendous impact on disease prognosis and optimal therapy. These limitations of traditional patient stratification calls for a more in-depth and finer classification.

The development of the several aforementioned high-throughput omics technologies has started to revolutionize the way we approach the problem of patient stratification, especially in cancer. Moreover, several omics technologies such as DNA microarrays ensure an unbiased and systematic collection of data, potentially facilitating novel discoveries in hitherto unexplored domains. Gene expression profiling was historically the first omics tech-

nique that was available and has been the most widely employed omics technology used in the area of personalized medicine. The systematic profiling of various cancer types has been among the first applications of microarray-based transcriptomic studies in the early 2000s (e.g[AED+00],[BKH+02],[SPT+01]),the gene expression microarray providing measurements of a set of patients allowing measurement of the biological phenomena and for discovering patterns that potentially allow insights into disease mechanisms. Moreover, microarrays have also been used to identify diagnostic, prognostic and therapeutic biomarkers which are clinically relevant. Many questions related to cancer diversity have potentially been addressed when molecular omics data are collected on different tissues and patients. For example:

- Is there observable diversity at the molecular level corresponding to that which we are already familiar with at the macroscopic level or under the microscope?
- Is it possible to define new, robust classification schemes based on molecular biomarkers ?
- What biological insight (mechanisms, pathways of action) can we get from comparing the molecular portraits of diverse samples?
- Is it possible to obtain better disease prognosis models and better predictive biomarkers for therapy response?

One of the hypothesis that researchers have tried to ascertain is the fact whether some of the clinicopathological parameters such as the dosage of protein markers are directly related to measures that we can perform at the molecular level, such as the expression level of the corresponding or related genes. It has indeed been shown by the likes of[DHKW+08] that the ER and HER2 status usually measured by pathologists in the clinics can be recovered, with

good accuracy, from the expression level of a few genes (see Fig.2.2). This allows in principle the automatic classification into the classical subtypes based on the expression profiles (see Fig.2.3). A landmark work in the area of using omics data set for Breast Cancer stratification identified a 70-gene signatures for metastasis prognosis [VVDVDV+02], [VDVHV+02]. This 70-gene signature has been validated prospectively and led to an FDA approved diagnostic test for clinical practice, MammaPrint®.

Spurred by the early success of using -omics data sets for clinically relevant patient stratification, National Institute of Health launched The Cancer Genome Atlas (TCGA) -omics data base back in 2006 [MFB+08]. This project has generated comprehensive, multi-dimensional maps of the important molecular changes in 33 types of cancer. The TCGA dataset has also been made publicly available.Some years later, another massive world-wide collaboration project, the International Cancer Genome Consortium (ICGC) [HAA+10], was

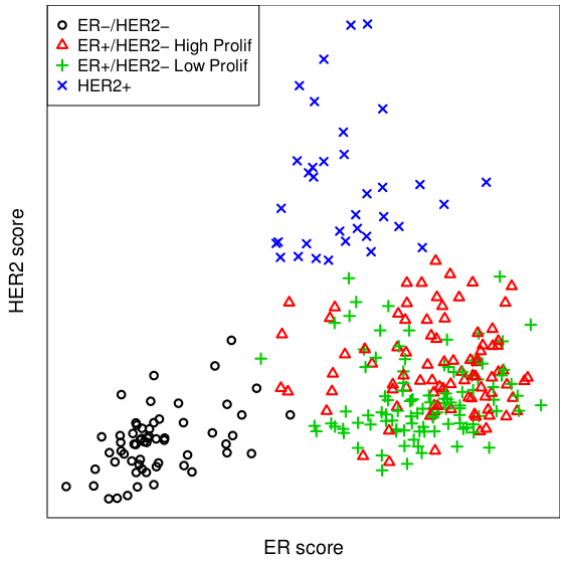

Figure 2.2: Breast cancer diversity in 2 dimensions. Global view of the 286 tumors in the Wang dataset, organized in terms of ER and HER2 status. Image reproduced from [BCH+12] under free copy license CC-BY-SA.

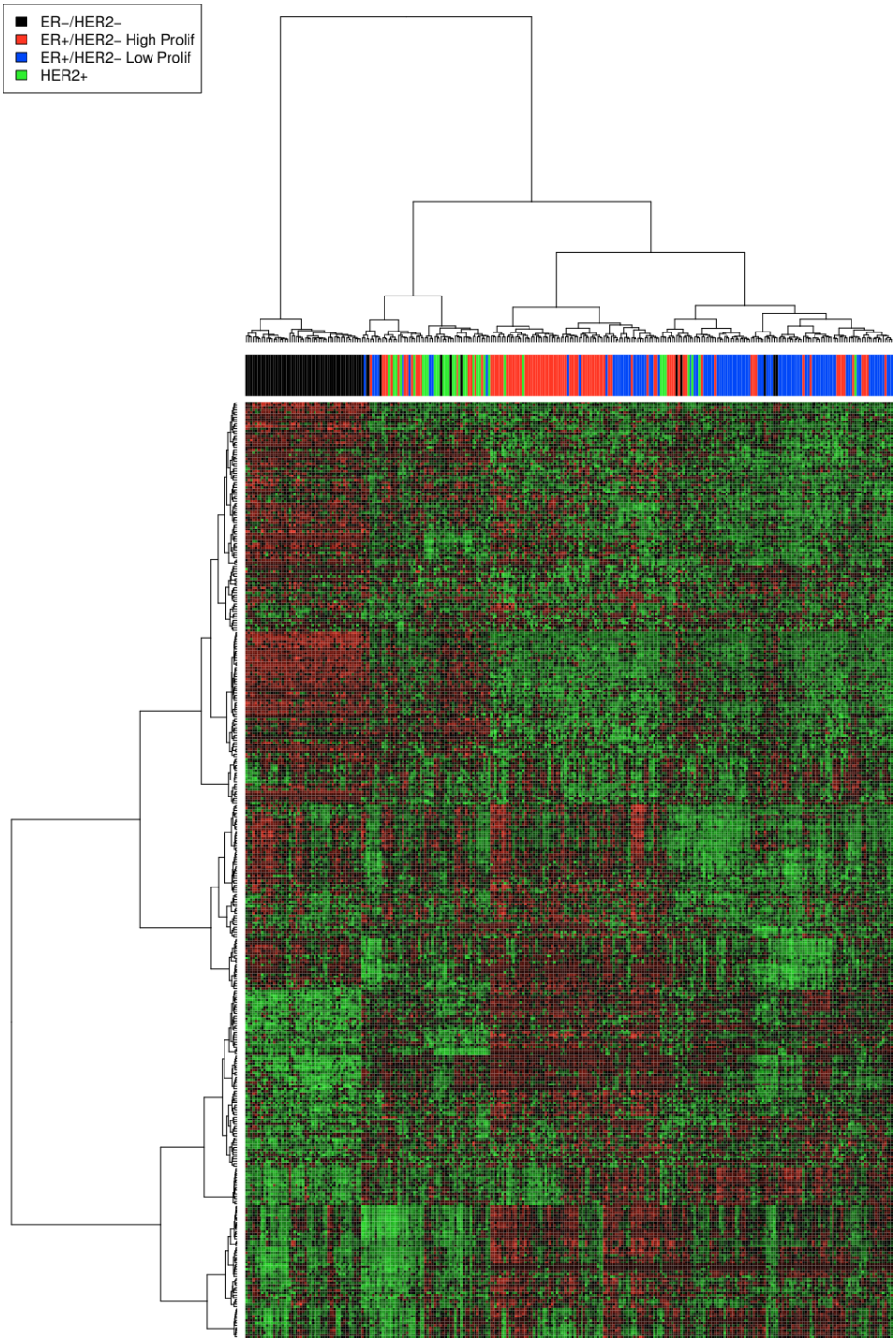

Figure 2.3: Molecular subtypes in Breast Cancer from mRNA expression profiles. The heatmap is based on 286 samples in the Wang breast cancer dataset. Image reproduced from BCH+ 12 under free copy license CC-BY-SA.

started with the goal of characterizing the molecular profiles of more than 50 cancer types with larger tumor samples. The samples in both these massive databases are accompanied with relevant clinical features along with corresponding molecular profiles which contain genomic, transcriptomic and epigenomics profiles. Both these repositories are open source and have resulted in large volumes of progress in patient stratification.

2.3 Statistical Methods for single Omics Data

We now turn our attention to the statistical techniques used for multi-variate analysis for single -omics data sets. Broadly speaking, there are two different paradigms of the statistical learning models when applied to patient stratification and biomarker discovery using -omics data: the first approach is an exploratory one which is also known as 'unsupervised' learning. The goal of this approach is to discover meaningful patterns within the patient population without being guided by any pre-defined patient classes (and thus the name unsupervised). One is interested in different clusters of patient population which are closer to one another than with those patients which are outside their cluster. This approach is particularly interesting when we have no prior knowledge about the underlying patient (or sample) population structure. Learning unsupervised models implies the learning of the cluster-specific parameters. Unsupervised learning approaches model a set of inputs, based on similarity without any reference to the class labels of those inputs. The model learning phase discovers the class labels. Popular unsupervised learning approaches are: Hierarchical clustering[Joh67], partition based methods like k-means and k-medoids[RK90] and matrix factorization methods like NMF (Non-negative matrix factorization)[LS99]. Unsupervised approaches for patient stratification strive on detecting subgroups from molecular charac-

terization of tumors with a hope of paving way for new, robust and unbiased taxonomies of cancers. The goal of such unsupervised, exploratory analysis is to provide parallel stratification schemes or to refine the classical stratification schemes based on clinicopathological factors. It could also reveal new molecular factors underlying the stratification, such as the activation of particular signaling pathways or the alterations of particular genomic regions, thus consolidating our current understanding of the molecular underpinnings of cancers.

The second approach is known as supervised learning' as it uses pre-defined patient stratas (e.g. high risk vs low risk) to obtain its model parameters. This use of already defined patient strata makes such models powerful from a predictive point of view. This means that once the parameters of such a supervised model are learnt using a set of patient data (known as training data), the model can then predict patient stratum for every new sample (known as testing data). This predictive approach of supervised learning becomes especially relevant when we have prior knowledge about the patient strata and we would like to stratify future patients based on our current data and model. Supervised learning generates a function that maps inputs to desired outputs (patient stratum also called class labels which are pre-defined by human experts). This learnt function is then used to predict class labels for other patients. Supervised learning uses patients clinical end-points (metastasis or not, time to relapse, overall survival or disease evolution etc.) This, then can be combined to define prognostic biomarkers which can be measured for future patients. These future patients (patients not used for building the statistical model) can then then be classified for example into high or low risk. This helps the physician to prescribe appropriate treatment. The rationale for this decision making is as follows: most cancer patients after their initial diagnosis and treatment are given adjuvant therapy in the form of cytotoxic drugs

which have strong deleterious side effects. However, based on patient stratification, this adjuvant therapy can be only given to high risk groups and thus sparing the low risk group of the harmful side effects of Chemotherapy. Therefore such predicted stratification can not only spare the morbidity of a treatment but can also justify a more aggressive adjuvant treatment to patients belonging to the bad prognosis or high risk class. Class-labels could be discrete (classification problem) or could be continuous indicators (like survival time). Famous examples of supervised models applied in Patient stratification are: support vector machines (SVM)[CV95], linear discriminant analysis (LDA)[Wel05], multinomial regression, Random forests[Bre01], Cox regression, Boosting[FSA99] etc.

Apart from stratifying patients into clinically relevant groups, statistical methods also focus on identifying relevant biomarkers and signatures. Discovery of biomarker signatures is an important aspect of the above mentioned statistical models. Owing to the oftentimes high dimensionality of the generated data, statistical methods (machine learning algorithms) employ some kind of variable selection. Variable selection is employed due to the fact that high-dimensional data contains a large amount of noisy, useless information that needs to be filtered out. Hence, almost every statistical modelling technique used in this domain also identifies few relevant features which could be of value[BVDG11]. These few relevant features can be interpreted as biomarker signatures. Variable selection not only allows for the detection of a small set of features which can be later validated, it also allows us to bypass important statistical challenges in high-dimensional statistics, viz. overfitting. Overfitting refers to the phenomenon in which the statistical model describes the training data perfectly, however performs poorly on unseen test data. When the number of samples (which is typically of the order of some tens or hundreds) is much smaller that the number

of features (of the order of some thousands), as is the case in most microarray technologies, building robust and stable statistical models calls for efficient feature selection. In order to tackle the problem of high-dimensionality, many advances have been made in the field of high dimensional statistics which caters to the statistical modelling for high-throughput omics data sets (see[CF12b] for a comprehensive review) . High-dimensional statistics refers to statistical inference when the number of unknown parameters p is of much larger order than sample size n, that is: $p \gg n$. A successful signature is a relatively small collection of q features, i.e. $q \ll p$ which can validated on other external data sets. The problem of defining a signature within the context of statistical modelling is associated with the more general problem of feature selection in machine learning. Feature Selection is an integral aspect of high-dimensional statistics, both in supervised and unsupervised approaches. It allows to circumvent many problems including like high-dimensionality and high-correlation within the omics data.

2.4 Motivation for Multi Omics Data Integration

The strength of data-driven statistical methods normally increases when more data are analyzed jointly. Therefore, during the last years there has been an increasing interest to analyze multiple, heterogeneous omics data in an integrated manner in order to gain a more comprehensive picture on complex biological systems[HHR11]. For example, large scale initiatives such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC)[HAA+10] now provide transcriptomics, methylomics, proteomics and genomics data of hundreds of patients for several cancer entities, allowing novel insights into cancer biology[MFB+08]. In the following section we review a few computational strategies

which have been proposed for integrative analysis over multi-modal omics data sets and the associated challenges. The text that follows is closely adapted from our review paper[AF16].

While most authors agree on the chances of omics data integration, the associated challenges have changed considerably over the last decades: While in the middle of the last decade data availability was seen as one of the big issues, later works mention statistical challenges, such as the risk of overfitting[CP12], and the difficulties associated with different technical platforms, for example differing normalization protocols and batch effects[KLR+14].

Altogether the challenges for integrating heterogeneous omics data may be summarized as follows: omics data of different modality (e.g. transcriptomics vs. proteomics) are measured with different techniques. Hence, these data have differing numerical types (e.g. discrete counts vs. continuous signals) and scales, coupled with large differences in the number of measured features (several hundreds of thousands of SNPs vs. few hundreds of miRNAs). Furthermore, each technical platform has another noise level and sensitivity. Consequently, naive merging of heterogeneous omics data increases the dimensionality of the data and thus increases the chance to produce false positive hypothesis testing results. In a machine learning setting the chance increases to overfit the data. In order to circumvent these problems the key question is therefore, how to identify and combine relevant features from each data modality in a way that respects known biological dependencies.

The goal of the subsequent sections is to give an overview about recent statistical inference and learning techniques that have been devised to address issues related to multi-omics data integration for patient stratification. Here[AF16], instead of emphasizing specific mathematical details of selected methods, we try to characterize the overall methods landscape. More specifically, we pose the question : in which way integrated omics data could be used

for personalized patient treatment in a supervised as well as unsupervised learning setting. We also focus on ideas adopted in the past to overcome the above mentioned challenges vis-a-vis multi -omics data integration.

2.5 Statistical Methods for Multi-Omics Data Integration.

As it has been mentioned before, one of the primary goals of personalized medicine is to stratify patients into clinically relevant sub-populations based on suitable biomarker signatures[CF13]. In this section we ask, in which way integrated omics data could be used for better personalized patient treatment in a supervised as well as unsupervised learning setting. Different classes of algorithms are discussed for both application tasks. Existing and future challenges for data integration methods are pointed out. An overview about the associated statistical learning techniques discussed in this section can be found in Table 2.1 and 2.2.

2.5.1 Clinical Outcome Prediction

During the last decade computational research in the personalized medicine area has mainly focused on learning predictive models based on one data modality (e.g. gene expression), possibly also in combination with biological background knowledge (see[CF12b] for a review). The advent of multiple, heterogeneous omics data modalities from the same patient (e.g. somatic mutations plus gene expression data) now raises the question, whether predictive models which utilize several combined data sources could improve prediction performance. Hence, the primary objective for omics data integration in personalized medicine is to enhance model learning and prediction performance.

In the machine learning community traditionally three general strategies for data inte-

gration are distinguished[PWCG01,MGKP08]: Early integration methods focus on extraction of common features from several data modalities, resulting into one integrated data matrix. In a second step conventional machine learning methods can then be applied. Late integration algorithms first learn separate models for each data modality and then combine predictions made by these models, for example with the help of a meta-model trained on the outputs of data source specific sub-models. The latter strategy is called stacking[Wol92]. Intermediate integration algorithms are the youngest branch of data fusion approaches. The idea is to join data sources while building the predictive model. An example of this strategy is Support Vector Machine (SVM) learning with linear combinations of multiple kernel functions[LCB+04].

All three data integration strategies have been applied in the area of personalized medicine: Pittman et al.[PHD+04] integrated clinical and gene expression data into a Bayesian decision tree classifier to predict breast cancer prognosis. Following an early integration approach the authors first summarized gene expression data into meta-genes[HIP+03], which were then joined with clinical variables. Selection of relevant variables was carried out via forward selection.

Boulesteix et al. first used partial least squares (PLS) regression to extract features from gene expression data[BPD08]. These features were then combined with clinical variables to train a Random Forest classifier for predicting breast and colorectal cancer outcome. In a similar vein Cao et al.[LCMM10] proposed a mixture of experts model to jointly model the effect of gene expression and patient clinical data to predict patient outcomes, they concluded that using gene expression data can provide valuable insights to understanding survival mechanisms by identifying prognostic biomarkers.

Gevaert et al. [GST+06] employed a Bayesian Network to combine clinical and gene expression based on the 70 gene breast cancer signature by van't Veer et al. [vDv+02]. The authors compared an early integration strategy based on simple pasting of data matrices with an intermediate and a late strategy. In the intermediate integration the authors first learned separate BN structures for each data sources and then join these networks based on the node representing the clinical outcome they had in common. In the late strategy only predictions by the two separate BN models were weighted and aggregated. The authors found the intermediate strategy to be most promising.

Daemen and co-workers suggested the use of a multiple kernel learning (MKL) framework to predict disease outcome of rectal cancer based on gene and protein expression data, and of prostate cancer based on transcriptome and CNV data [DGO+09]. Within the MKL framework separate kernel functions were defined for each omics data modality. A linear combination of these kernels was then employed to train a least squares SVM (LS-SVM). The authors reported a better prediction performance of this intermediate data integration strategy compared to model stacking.

Following again the idea of MKL, Thomas et al. suggested a weighted LS-SVM classifier to combine gene expression and clinical data [TBSM14]. Compared to models built on each individual data modality as well as compared to an early integration strategy using generalized singular value decomposition, the authors found a significant improvement of their approach for predicting breast cancer outcome.

Wang et al. [WBM+13] developed an integration scheme based on probabilistic graphical models and Bayesian inference. Their iBAG algorithm (integrated Bayesian Analysis) combines miRNA, DNA methylation and mRNA data to predict patient survival of Glioblas-

toma Multiforme (GBM) patients. Their approach explicitly takes into account the biological relationship between different data modalities. The authors identified separate gene sets related to disease outcome and demonstrated better prediction power to detect disease related genes than non-integrative methods.

Gade et al. first constructed a correlation weighted bipartite miRNA-target gene graph [GPF+11]. This graph was then used to guide feature selection with a component-wise likelihood boosting algorithm for predicting prostate cancer outcome [BS09]. Going one step further other authors also considered protein-protein interaction information [CFI3]. Their method first smoothes marginal t-statistics of genes and miRNAs over the structure of the integrated PPI and miRNA-target gene network via random walk kernels. Most relevant features are then determined via a permutation test. Subsequently a conventional SVM classifier is trained. The authors demonstrated the benefit of this approach compared to stacking for predicting disease prognosis in several cancers.

Arguable one of the most advanced but also computationally costly approaches for intermediate data integration in the field of personalized medicine has recently been suggested by Zitnik and Zupan [ZZ14]. The authors combined gene expression and histological data from animals and human with protein-protein interactions and GO annotation to predict liver injury induced by chemicals. This was done based on a constrained matrix trifactorization algorithm suggested by the same authors [ZZ15].

Vliet et al. made a comparison of several integration strategies (pasting of feature matrices, linear combination of distance measures or kernel functions, stacking) and classifiers to predict breast cancer outcome [vHv+12]. The authors reported most success via an intermediate strategy using a nearest mean classifier our via a late strategy using a logical OR function.

2.5.2 Unsupervised Patient Subgroup Detection

Apart from supervised patient stratification using defined clinical endpoints (e.g. survival times), a lot of effort has been made to detect patient sub-populations in a completely unsupervised manner based on molecular data. An example of this approach is the detection of four different molecular subtypes of Glioblastoma Multiforme (GBM) patients based on gene expression data by Verhaak et al.[MFB+08]. As more molecular data modalities from the same patient become available now, many authors explored the possibility of fusing these data for discovering stronger patterns (see[CKB+14] for a review)

Akin to the case of supervised learning for patient stratification, unsupervised data fusion approaches can be broadly classified into three groups, which involve early, late and intermediate integration schemes. Early integration methods work with a joint feature matrix and modify traditional clustering algorithms, such as k-means, to calculate a weight for each data source[CXHY13]. Late integration combines patient similarity matrices obtained from independent clusterings of distinct data types. Intermediate integration methods typically aim for extracting common features from different data modalities combined with clustering of patients.

An example of an intermediate integration strategy is non-negative matrix factorization (NMF)[LS01]. The idea behind NMF is to factorize a data matrix into a product of two matrices, one indicating discriminative feature combinations between clusters and one indicating cluster assignments of patients. While originally NMF was designed to work with one data modality only, later work has extended the approach to simultaneous clustering of several data types. For example, Zhang et al. used an extended NMF framework to cluster 385 ovarian cancer patients based on joint gene expression, DNA methylation and miRNA

profiles[ZLL+12].

Another popular intermediate integration approach is the iCluster method by Shen et al.[SOL09,SMS+12]. This technique combines ideas from sparse matrix decomposition and latent factor models and has also remarkable similarities to probabilistic PCA[TB99] and k-means[DH04]. Furthermore, the iCluster method can be seen as a special case of Bayesian canonical correlation analysis with a sparsity prior for the coefficient matrix[KVK13], facilitating model identifiability and interpretability. In[SMS+12] the authors used iCluster to integrate gene expression, DNA methylation as well as CNV data of Glioblastoma Multiforme (GBM) patients. The iCluster method treats information from all patients with the same confidence, which may lead to erroneous results, if there are patients with disconcordant information from different omics data modalities. The latter issue was taken up by Yuan et al.[YSM11], who developed a Patient Specific Data Fusion (PSDF) model, which gives different patients separate weights within a non-parametric Bayesian framework. A unique aspect of PSDF is that it allows for the separation of concordant and dis-concordant signals from patients and unlike iCluster does not force patients to cluster together. The obtained disease subtypes via PSDF were reported to be prognostically relevant by the authors. A limitation of the PSDF method is in the required data discretization, which may lead to considerable loss of information. Similar to the PSDF method Kormaksson et al.[KBF+12] proposed a mixture-model for integrative clustering of gene expression and DNA methylation data. Unlike PSDF, the method does not require data discretization. However, a limitation is the assumption of statistical independence of molecular features.

Another recent mixture model approach is the MDI (Multiple Data Integration) approach by Kirk et al.[KGS+12] and Savage et al.[SGG+13]. Following a Bayesian non-parametric

clustering approach MDI assumes a Dirichlet Process Prior over cluster assignments. Moreover, and in contrast to PSDF, MDI learns exact dependencies between the different data sources as a directed acyclic graph. This implicitly results in a preference to put patients into the same cluster, if they tend to group together in each of the different data sources. However, at the same time each data source still retains its own clustering, reflecting the fact that different molecular data may express partially non-concordant patient groupings. Savage et al.[SGG+13] used the MDI model to integrate genomic, epigenomic and transcriptomic information of GBM patients and reported clinically relevant disease sub-types. MDI is flexible in modelling continuous (e.g. gene expression) as well as discrete (e.g. CNVs) data. A limitation is the assumption of statistical independence of molecular features.

Generative modelling approach, such as MDI and PSDF, require to express explicitly the joint statistical distribution over different data modalities. This complication is avoided in late integration techniques. Examples are Similarity Network Fusion (SNF)[WMD+14] and Multiview Genomic Data Integration (MVDA)[SFF+15]. These techniques use independent clustering algorithms for each data modality and aggregate results of patient similarity matrices from each data source. Thus, late integration potentially allows for incorporating thousands of features for each data modality. Furthermore, late integration techniques are typically more robust to small sample sizes. A limitation is the difficulty to explicitly model dependencies between data modalities. The SNF method models patient similarities as networks with nodes representing patients. Each data modality generates its own network, and these networks are then fused into a consensus network using a message-passing algorithm. The authors in this way integrated gene expression, DNA methylation as well as miRNA profiles over five cancer datasets and performed graph clustering on the consensus network

to identify disease subtypes. The MVDA approach[SFF+15] concatenates patient-patient similarity matrices obtained from different data sources and then uses matrix factorization of the concatenated matrix to come up with a consensus clustering.

Biclustering is yet another popular statistical technique for simultaneous clustering of the rows and columns of a data matrix and has recently also been employed for data fusion. The original method along with its modifications has since many years found several applications in biological data analysis (see[MO04] for a comprehensive review). Recently, Bunte et al.[BLSK16] developed a novel bi-clustering algorithm to cluster cancer cell lines treated with different drugs while including CNV, DNA methylation, mRNA, protein abundance and exome sequencing information. The model is based on the previous work of the same group of authors on the Group Factor Analysis Model[KVLK15]. Another technique based on biclustering has been proposed by Sun et al[SBK13,SBK14]. Their method is based on sparse singular value decomposition (SSVD) and was applied to combine SNP information with clinical data for disease subtyping and identification of subtype-specific genotype variations.

Table 2.1: Selected Statistical Learning Techniques for Personalized Medicine using Multiple Data Sources

Method	Objective	modelling Approach	Input	Output	Assumptions	Advantages	Limitations
Daemen et al. [DGO+09]	supervised clinical outcome prediction	Multiple Kernel Learning	mRNA, CNV, clinical data	clinical outcome	linear kernel combination can enhance prediction performance	flexible and extendable framework	computationally costly
iBAG, Wang et al. [WBM+13]	supervised clinical outcome prediction	graphical model	miRNA, mRNA, methylation	patient survival	model consistent with biological data and at least partially identifiable	fully probabilistic approach	framework not easy to extend; computationally costly
Gade et al. [GPF+11]	supervised clinical outcome prediction	correlation, statistical meta-analysis, boosting	miRNA, mRNA	patient survival	miRNA-target predictions largely consistent with biological reality	conceptually simple	framework not easy to extend; computationally costly
Zitnik et al. [ZZ15]	supervised clinical outcome prediction	matrix factorization	miRNA, PPI, GO annotation, histological data	chemical induced liver injury	biologically relevant information can be extracted from linear subspace of the data	flexible and extendable framework, can integrate various types of information	relies on relations between biological entities (e.g. GO terms and genes), computationally costly

Table 2.2: Selected Statistical Learning Techniques for Personalized Medicine using Multiple Data Sources

Method	Objective	modelling Approach	Input	Output	Assumptions	Advantages	Limitations
Zhang et al. [ZLL+12]	unsupervised disease subgroup identification	matrix factorization	mRNA, miRNA, methylation	disease subtypes	biologically relevant information can be extracted from linear subspace of the data	flexible and extendable framework	same influence of each data source
Shen et al. [SMS+12]	unsupervised disease subgroup identification	matrix factorization	mRNA, miRNA, methylation	disease subtypes	biologically relevant information can be extracted from linear subspace of the data	flexible and extendable framework	same influence of each data source
PSDF, Yuan et al. [YSM11]	unsupervised disease subgroup identification	Bayesian non-parametric (Dirichlet Process Mixture Model)	mRNA, CNV	disease subtypes	model consistent with biological data and at least partially identifiable	fully probabilistic, flexible and extendable framework	data discretization, computationally costly
MDI, Kirk et al. [KGS+12]	unsupervised disease subgroup identification	Bayesian non-parametric (Dirichlet Process Mixture Model)	mRNA, DNA methylation, CNV	disease subtypes	model consistent with biological data and at least partially identifiable	fully probabilistic, flexible and extendable framework	assumes statistical feature independence; computationally costly
SNF [WMD+14]	unsupervised disease subgroup identification	patient similarity, message passing	mRNA, miRNA, DNA methylation	disease subtypes	thresholding patient-patient similarities defines subtypes	flexible and extendable framework	neglects biological dependencies between data modalities

If I have seen further it is by standing on the shoulders of Giants.

Isaac Newton

3

Machine Learning Approaches to Personalized Medicine in Glioblastoma

3.1 IDENTIREST: Identifying new therapeutic targets in Glioblastoma

As described earlier, Glioblastoma (GBM) is a brain tumor with an incidence rate of 3-4 cases per 100,000 people. GBM is the most malignant brain tumor in adults and is also

one of the most aggressive human tumors. Tumor resection along with the use of radiation and chemotherapy have a positive influence on the survival of the patients. Still, the prognosis remains poor with a median survival of only around 14 months. Apart from the classic characteristics of tumors, GBM has additionally a number of peculiarities that are currently insufficiently taken into account for the diagnosis and treatment planning. Although, various histological and molecular subgroups exist[PKC+06, VHP+10b], they are often grouped together as one entity leading to a very heterogeneous course of the disease. In addition, there are also a variety of cellular and functional phenotypes within tumor tissue that have not been adequately classified to date. Residual tumor cells that remain beyond the margins of every glioblastoma (GBM) resection are believed to play an important role in relapse of the disease[GRS+10]. These residual cells are also know to be resilient to post-surgical therapy. These residual tumor cells have not been studied in the past and the goal of the IDENTIREST project was the characterization of these cells with a goal to identify new drug targets.

The project builds on the past work of[GRS+10] which concentrated on the isolation and initial characterization of vital residual tumor cells of GBMs. It was shown that the residual cells have different properties than the routinely removed and conventionally studied GBM cells (known also as central cells). They are e.g. migratory and proliferative active, but have a lower content of stem cells. Moreover, the authors observed a different expression of relevant therapeutic targets along with different response to in vitro therapy (see Fig. 3.1).

As a Pilot project, 12 paired cell samples were used to generate genome-wide transcription profiles (using Affymetrix array). The data analysis of the transcription profiles revealed that the molecular signature of residual tumor cells differs significantly from the

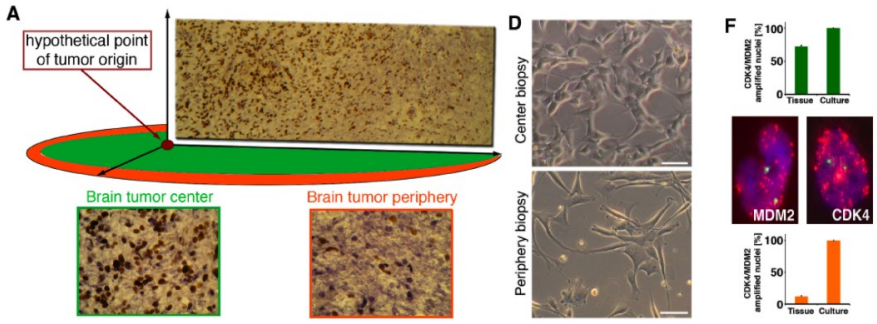

Figure 3.1: Figure taken from Glas et al., 2010 [GRS+10]: (A) GBM tissue taken from routinely-picked and conventionally analyzed (green). In the resection margin (orange), however, tumor cells (residual tumor cells) always remain in the patient (D). Center biopsy morphology of GBM cells. Resection margin (Periphery biopsy), the location of the residual tumor cells. (F) GBM cells (identified and quantified by patient-specific amplifications, here the MDM2 and CDK4 genes) can be isolated from the tissue of the tumor center as well as the resection margin and accumulated to a similar extent. This allows the comparative in vitro analysis of both cell populations.

signature of central GBM cells. 109 significantly differentialy expressed genes were found which on applying a stricter filter criterion led to 14 candidate genes (see Fig.3.2)

A pathway enrichment analysis using the Kyoto Encyclopedia of Genes and Genomes (KEGG, www.genome.jp/kegg/) resulted in 33 signaling cascades which are significantly differentially active in both cell populations. Altogether, 109 significantly differently expressed genes could be specified (Figure 3.2A). By applying filter criteria, this list could be narrowed down to 14 candidates for further analysis (Figure 3.2B). Expression of these genes was confirmed by qPCR on a collective of five paired samples (Figure 3.2C). The distribution profile of the candidate genes was shown to be stable even with prolonged in vitro expansion of the cells (Figure 3.2D).

The functional relevance of these regionally expressed genes, which are potential targets has also been explorative and demonstrated (Figure 3.2E and F): (i) Targets of central GBM cells: The proliferative activity of GBM cells of the tumor center was checked via inhibition of Fibroblast growth receptor FGFR1 . Residual GBM cells that express less of this

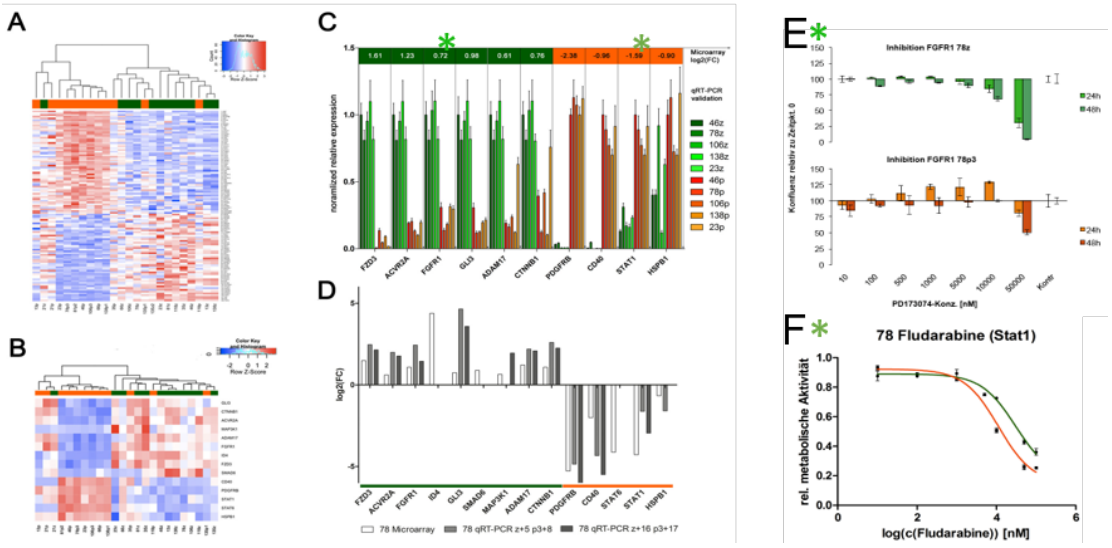

Figure 3.2: Figure taken from unpublished manuscript (Glass, Laurel, Cheerful, Riemenschneider, Scheffler in preparation.)(A) Unsupervised cluster analysis of the 109 distinctly differently expressed 109 genes in central (z) and residual (p) GBM cells. (B) By applying filter criteria to (A), a shortlist of 14 candidate genes could be generated. (C) qPCR validation of candidate genes on five pairs of samples (D) Stability of the profiles (cell passage was analyzed 3 vs. 17 in vitro, underlying microarray data from passage 5). (E) Pilot experiment for FGFR1 (cell confluence determination by Cellavista®) and (F) STAT1 (measurement of metabolic activity by alamarBlue® assay) show that central and residual GBM cells can be differentially inhibited due to the different expression of the target structures (color coding of the data : green, central GBM cells, orange, residual GBM cells)

receptor, on the other hand, were much less inhibited. (ii) Targets of Residual GBM Cells: Conversely, fludarabine15-mediated inhibition of the transcription factor STAT1 in residual GBM cells produced significantly greater inhibition of metabolic activity than in the corresponding tumor central cell samples (Figure 3.2C and F). STAT1 is expressed more strongly in residual cells.

The goal of the IDENTIREST project was to study and analyze these residual tumor cells in large population cohort. The aim was to then validate the results of the pilot experiments, which could potentially lead to the development of novel targeted therapies. These novel therapies would also be patient-specific and would thus lead to the goal of new

tailor-made remedies for GBM. For this purpose, a cohort of around 270 samples coming from around 60 patients were used. As the Primary GBM and Residual GBM samples had widely differing properties, a new and innovative biotechnology technique was employed (using stem cell biological technologies) to grow and expand the cells such that they could reflect and map patient-specific properties in vitro for a longer time. It was made sure that there were at least two tissue samples are available per patient: one from the tumor center and one from the resection wall after completion of routine intervention. The latter sample was then used for purification of the Residual tumor cells.

Two kinds of molecular data was available: Whole genome transcriptomics data (using Affymetrix GeneChip™ Human Transcriptome Array 2.0) for around 220 samples and SNP data for around 190 samples (using Infinium CoreExome-24 and Infinium Psych DNA microarray). Apart from molecular data, patient level clinical data was also available like age, sex, pre and post-surgery Karnofsky Index. The clinical data has been summarized in Fig.3.3

In the following sections we try to answer the important questions raised by the IDENTIREST project. Each section describes a certain problem which was necessary to first gain more clarity about the nature of the Peripheral Samples and secondly would lead to patient-specific targets (based on certain biomarkers). Firstly, a set of differentially expressed genes were identified between the central and peripheral samples (details of which follow in the next section). Following that a significant amount of work was done by others to prioritize these potential target genes leading to a validated set of 32 genes (the details of which have been left out here as the work falls outside the scope of this thesis).

The present chapter tackles questions related to the nature of the samples derived from

patients, providing algorithms for sample stratification (not patient stratification). This in turn can be understood as providing answers to smaller parts of a much bigger puzzle (i.e. of finding novel personalized treatments for GBM patients). The key accomplishment of this chapter is the development of the statistical classifier which can successfully predict the relative location of tumour recurrence from peripheral gene expression profiles. This classifier has very important implications in the clinical management of the GBM patients. Using the predictions of this classifier, regions of tumor recurrence could be identified in future patients leading to more targeted adjuvant therapy.

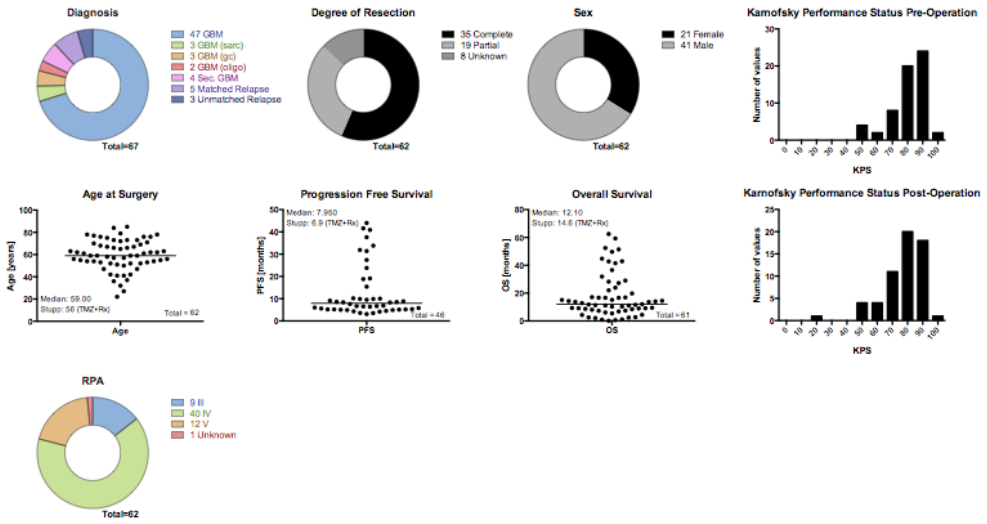

Figure 3.3: Figure describing the clinical data from the patients in the IDENTIREST project

3.2 Sample heterogeneity in Transcriptomics data

One of the key premises of IDENTIREST project is the existence of two kinds of cells, viz. the central tumor cells and the residual tumour cells which are located on the immediate periphery of the central cells (see Fig.3.4). After tumor resection, the surgeon collects sam-

ples of both kinds from the patient. Our first task was to explore the differences between residual tumor cells (known as periphery or P cells) and the tumour cells (the central cells or C cells) using transcriptomics data.

For this purpose we first used the Transcriptomics data set (gene expression array). We performed the following normalization steps: We used '`arrayQualityMetrics`' R-package[KGH08] to perform initial data quality check. Then using the R-package '`affy`'[GCBI04] probe level summarization was performed via a median polish and background correction carried out via RMA[IBC+03 Iri03]. Mapping of probe-sets to genes was based on the Affymetrix annotation file, which can be downloaded from the Affymetrix web page.

After the aforementioned normalization of the raw data we looked into the expression profiles of the 220 samples. These cells were not only P or C, but other kinds of cell types were also included for reference. As the transcriptomic data contains around 70000 features, PCA plots are a way to understand the heterogeneity within the sample population on a low dimension. This is presented in Fig.3.5. As one can see from the Fig.3.5, there exist a large heterogeneity within the cell population of Peripheral and Central Cells. The reference samples (like GBM Cell lines, NPCs, Astrocytes and Neurons) are well separated from P and C cells, however there are two clusters which contain both the Peripheral and Central Cells.

3.3 Sample Identification using Genomics Data

Apart from the mRNA expression data, we also had access to the SNP array data for 187 samples genotyped on two Illumina arrays - Infinium CoreExome-24 and Infinium PsychArray. Our focus in using the genomics data was on identifying sample characteristics

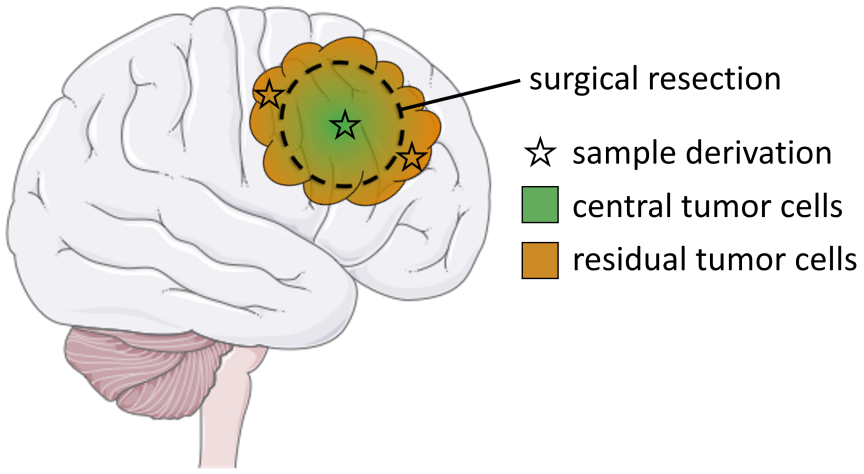

Figure 3.4: Figure describing the origin of the peripheral and central cells

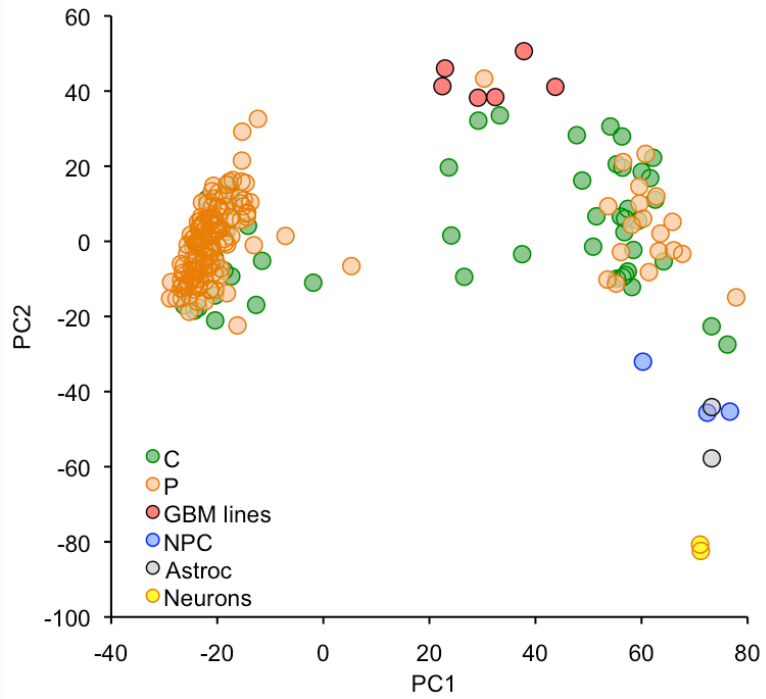

Figure 3.5: Figure describing the first two principal components of the Transcriptomics data

based on the differences in genomic profiles for Central (C or Z) and Peripheral (P) samples. We hoped to find key differences in these two cell population. In a way, the Genomics data served to validate our hypothesis of differences between the two kinds of cell types.

The input for our analysis was Genotype, LogR and B-allele Frequency values for 588,454 SNP locations (for 187 samples). The corresponding annotation of the SNPs on the chromosome was also provided. We used open source library PennCNV-1.0.3 to identify larger structural changes like Copy Number variations (CNVs). Approximately 25,000 CNVs were detected using the software. Further filtering of CNVs by specifying minimum CNV length of 50kb and minimum number of contained SNPs to 10 leaves around 16,000 CNVs. In the next step only those CNVs which are locally associated to gene regions (within 5 kb pairs) are annotated to Genes. The 16,000 CNVs map to around 20,000 genes, this is because one CNV could map to multiple neighbor genes. We only consider those CNVs which show changes in at least 5 samples, also we only use those genes who have non-conflicting CNV changes (meaning that the CNV regions that map to the same gene should all have the same Copy Number change). This leaves us with 1070 genes with their corresponding copy number changes.

Given this list of genes with their copy number variations in each sample, we explore two important questions: a) Is the CNV data associated to the Peripheral/Central annotation of the samples (from the surgeons annotation) and b) Is the CNV data associated to the gene expression pattern if we account for the class-specific (i.e. P or C) variations. In order to answer these questions we fit separate gene-specific generalized linear model for both questions. In a) class annotation is used as outcome variable (P or C) accounting for Patient IDs and gender as additional covariates apart from CNV data; in b) gene expression value is

used as outcome variable while accounting for Patient IDs, gender and sample class (P or C) as additional covariates apart from CNV data. We next use log-likelihood test to check the significance of the association of the outcome variable in each case to the CNV data.

Out of the 1070 genes, 473 genes have a significant (FDR 0.01) association with Peripheral/Central cell types. This association can also be seen in the CNV heatmap of top 202 CNV containing genes for the 180 samples which contained CNVs in Fig.3.6

The frequency of CNV calls are also significantly different in P samples and C samples. (p-value of $7e-04$). This difference can be seen in the Frequency of CNV in P vs C samples as depicted in Fig.3.7. This hints again to the fundamentally different nature of the P samples and the need for further exploration.

As mentioned before significant amount of work was done by other to prioritize potential target genes (differentially expressed between P and C sample types) leading to a validated set of 32 genes. We found that out of these 32 genes, 7 of them have a significant (FDR 0.01) association (using the linear model approach described above) of the Copy Number to their corresponding gene expression scores. We have visualized this in Fig.3.8. The results shown in this section point to the fact that CNV data provides a valuable information in two ways:

- Providing a molecular data modality which can be used to distinguish Peripheral and Central samples annotation provided by the surgeon.

- Corroborating the evidence of the changes in the gene expression pattern between the P and C cells with corresponding CNV changes which are significantly associated (see Fig.3.8)

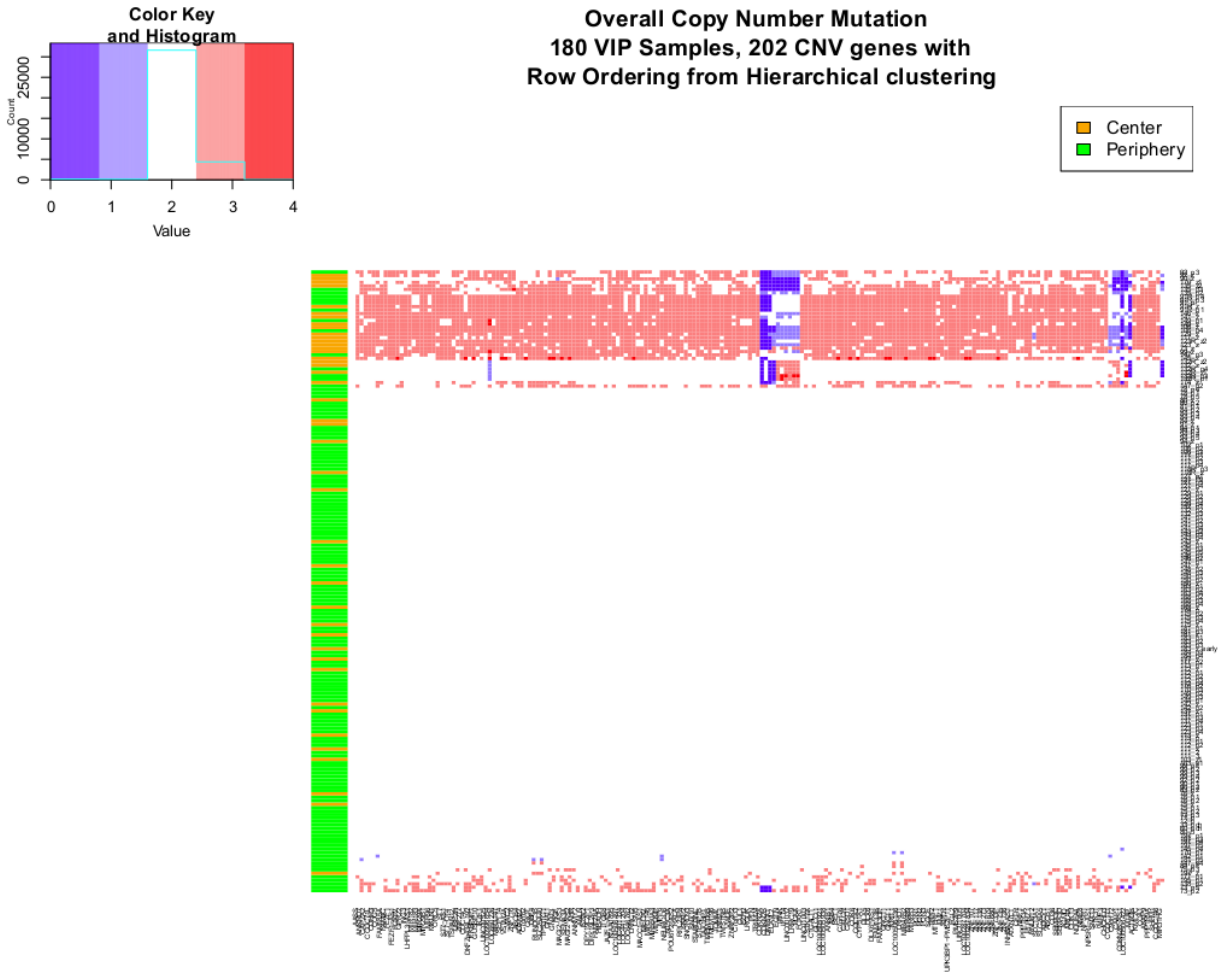

Figure 3.6: Heatmap showing CNVs of selected genes for the samples which are annotated as either Peripheral or central cell types.

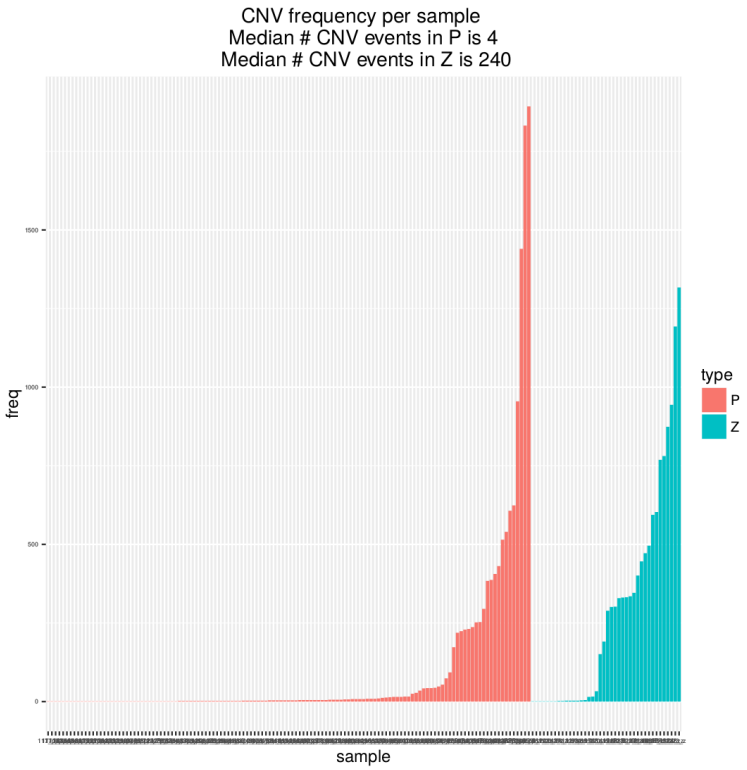

Figure 3.7: CNV call frequency for each of 180 CNV samples. There is a significant difference between Peripheral and Central cell types

3.4 Characterizing Samples based on Verhaak Classification

Another way to explore the differences between the Central and Peripheral cell types is to use a known classification scheme in Glioblastoma and to classify the two cell types. The classifier predicts a defined class for each of the samples in our project. We can then analyse the differences of the classification results on the two different kinds of samples (P and C samples). For this purpose, we use gene expression data described in Section 2.2. For the established classification scheme we choose the one of Verhaak et al. 2010 which is a landmark work in the area of stratification of samples for Glioblastoma.

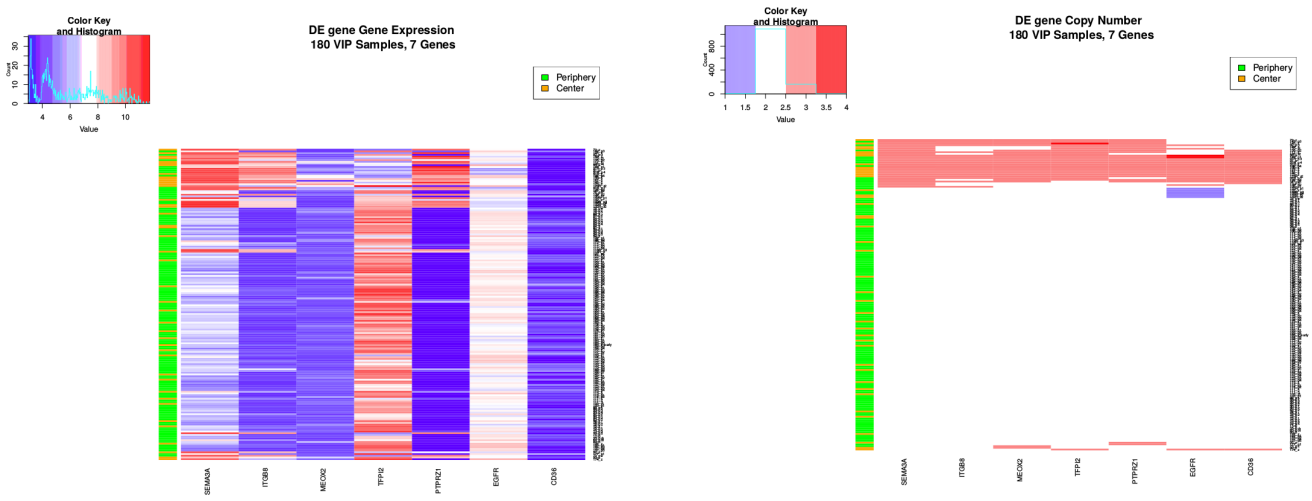

Figure 3.8: List of seven target genes which showed simultaneous association of the CNV changes to expression profiles (Left) as well as class annotation (Peripheral or Central cell types) (Right)

Verhaak et al. 2010 [VHP+10b] established a molecular classification of Glioblastoma based on Gene Expression profiles of 202 patient samples. The authors used the expression pattern of 1,740 genes in a consensus average linkage hierarchical clustering (Monti et al. 2003 [MTMG03]) and obtained four classes based on maximum clustering stability. Further, they defined a 840 gene signature (210 genes per class) using ClaNC, a nearest centroid-based classifier which balances the number of genes per class (Dabney, 2006 [Dab05]). This 840 gene set showed the minimum cross-validation error. The four classes were named as Classical, Mesenchymal, Proneural and Neural. The dysregulation of each of the genes- EGFR, NF1, and PDGFRA/IDH1 were highlighted as "defining characteristics" of the Mesenchymal, Proneural and Neural Classes respectively. This gene signature was further used to classify samples based on a separate validation set. The reproduction of a similar gene expression pattern in the validation set was taken to be strong evidence on the reproducibility

of their 840 gene signature. A further functional annotation of the subtypes was done by integrating the Genomic data to identify statistically significant subtype-specific copy number variations. The 840 gene signature and the expression profiles of 202 patients was downloaded from the TCGA's webpage: https://tcga-data.nci.nih.gov/docs/publications/gbm_exp/. This data was sub-setted for only those genes which were part of the genes present in our IDENTIREST study, thus leaving 751 genes.

The gene expression data with 180 samples and 751 genes is used for obtaining class membership probabilities for the four Verhaak classes. An Empirical Bayes method (Johnson et al. 2007[JLR07]) is first used for cross-platform normalization of the IDENTIREST dataset with the Verhaak data set. We use the implementation of Empirical Bayes in the CONOR package in R (Rudy et al. 2011[RV11]). After an additional gene wise z-score normalization, a four class linear SVM (Schölkopf 2002[SS02]) classifier is built on the basis of the cross-platform normalized training data by Verhaak et al. The cost or regularization parameter is chosen by minimizing the Cross Validation error across a series of cost values. The final classifier (with probabilistic predictions) is then applied to our IDENTIREST samples, and probabilistic predictions are made separately for C samples (see Fig.3.9) and P samples(see Fig.3.10). The Central cells can be seen to belong to all the four Verhaak classes thus further evidencing the fact that they are regular cancer cells with cellular heterogeneity. The behavior of the P cells on the other hand is atypical of normal cancer GBM cells as they mostly belong to the Mesenchymal subtype of the Verhaak classification (small minority belonging to Neural and Classical subtypes). This behavior further points to the special nature of the Periphery cells. It has been shown that the Mesenchymal phenotype in glioblastoma (GBM) and other cancers drives tumor aggressiveness along with resistance to treatment,

hence leading to therapeutic failure and often recurrence of disease.[OLX+17]

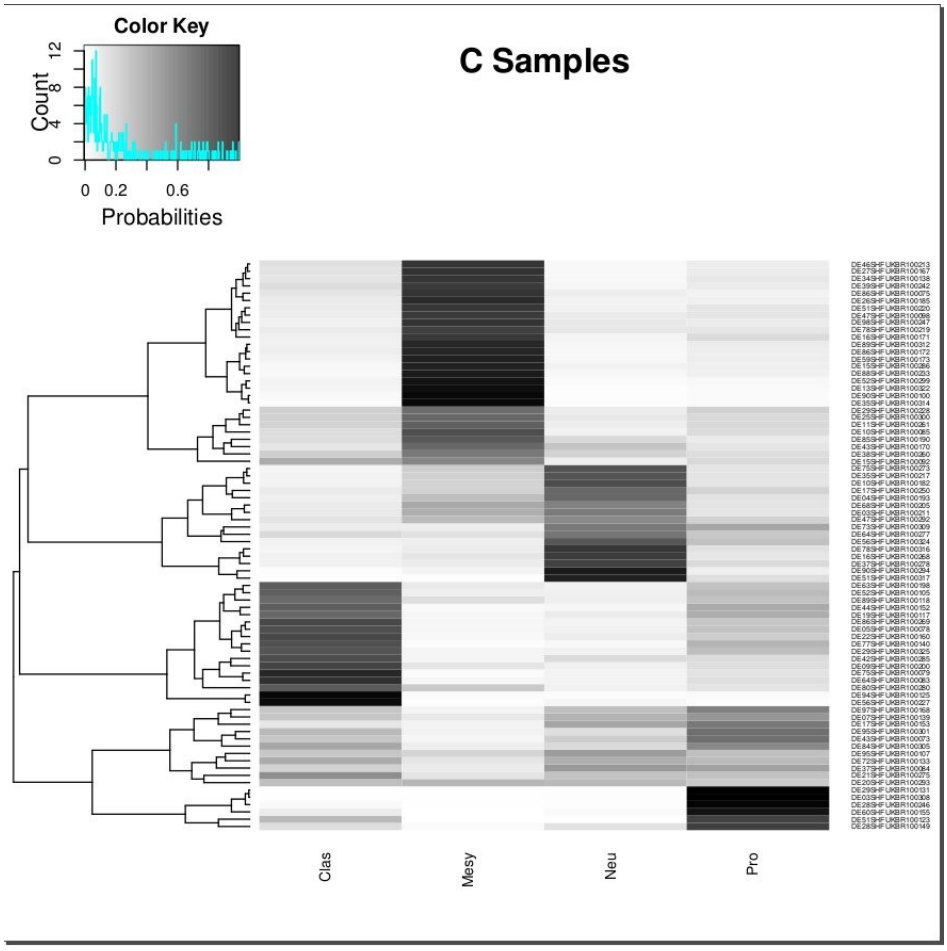

Figure 3.9: Results of the Verhaak Classification scheme on the Central Samples with each row denoting the probability for belonging to that particular subtype.

3.5 Statistical Modelling for predicting spatial recurrence in Glioblastoma

We have established the distinct nature of Peripheral cells as compared to that of Central cells which are cancer cells and have been shown to possess GBM cell-like properties. We have also explored the fact that there is heterogeneity within the C-cells (see Verhaak classifi-

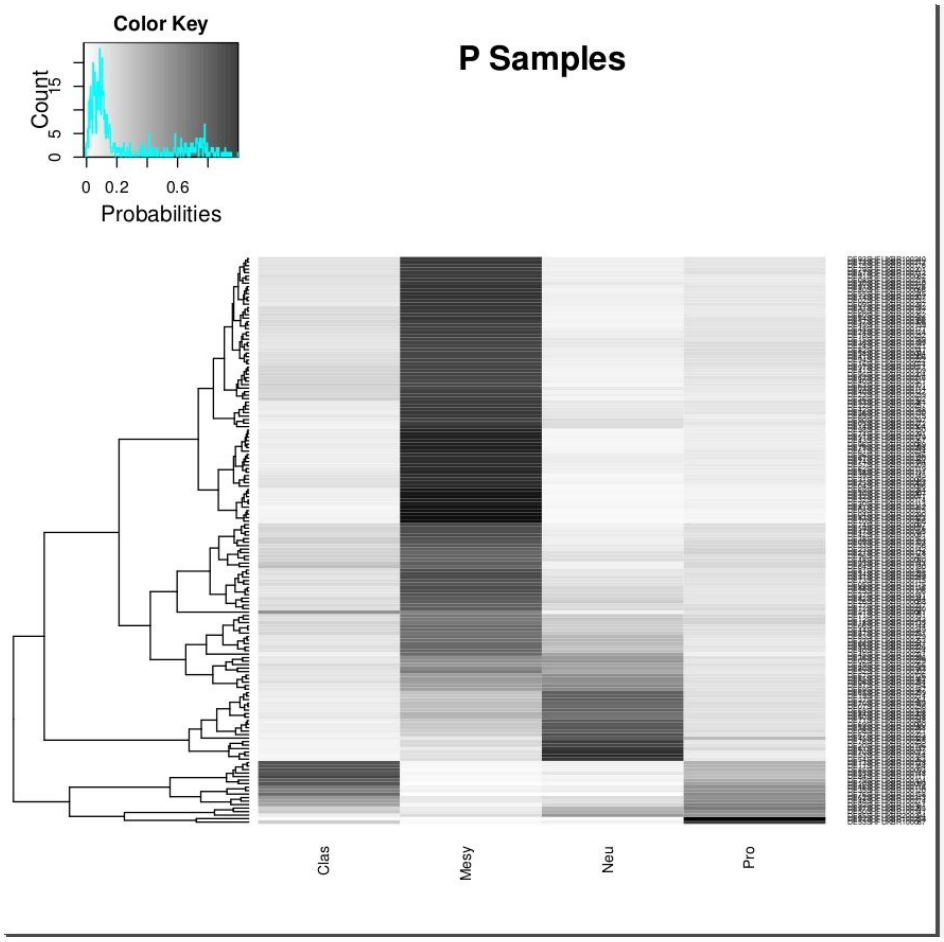

Figure 3.10: Results of the Verhaak Classification scheme on the Peripheral Samples with each row denoting the probability for belonging to that particular subype.

cation results) as well as within the P-cells (see PCA plot in Fig.3.5). In this section we probe further the heterogeneity within the P-cell population. This is also, clinically, a very relevant issue: In our IDENTIREST project, many peripheral samples were collected for the same patient. These different peripheral samples originated from different brain regions of the patient. As we know that in most cases there is a recurrence of the GBM after a certain time. The question we tried to answer here was whether we could predict the brain region where the recurrence began. Fig.3.11 depicts the sample collection from multiple regions surrounding the initial tumor resection. This means that out of the many P sample biopsies that we get, some will be associated with disease relapse (designated as RI samples) while others not (designated as RU samples).

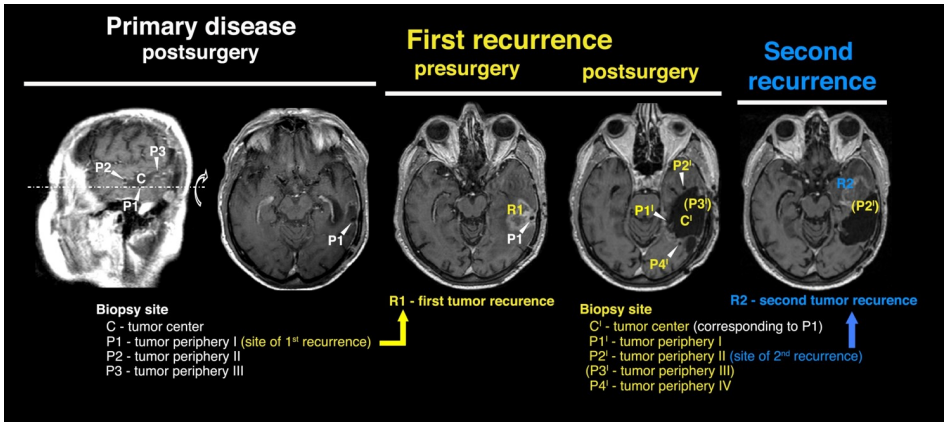

Figure 3.11: Figure depicting the collection of multiple biopsies from tumor center C vs. periphery P. The exact localization of each biopsy is marked during initial phase of disease for longitudinal follow up. After which they are labeled as 'RI' (involved site) or 'RU' (uninvolved site)

We developed an elastic net based machine learning classier (Zou, 2005 [ZH05]) to discriminate between RI and RU samples. The classifier takes as input the gene expression profile in brain tissue surrounding the tumor resection area. Moreover, the information whether a sample stems from those particular patients, from which also other tissue material are

available, is also made use of (details are explained later). The model forecasts whether the recurrent tumor will appear in an involved (RI) or uninvolved region(RU). Figure 3.12 provides an overview about the approach that we have taken to develop the classifier and to validate it. Briefly, we started from an initial training cohort of 73 expression profiles from 26 patients (28 RI, 45 RU). Based on these data we evaluated and compared two different methods (Figure 3.12A):

1. An elastic net classifier using microarray features.

2. An elastic net classifier using biological pathway activity scores based on Single Sample Gene Set Enrichment Analysis (Barbie, 2009[BTB+09]). Single-sample GSEA (ssGSEA) calculates separate enrichment scores for each sample in each gene set. Each ssGSEA enrichment score represents the degree to which the genes in a particular gene set are coordinately up- or down-regulated within a sample.

3.5.1 Details of Classifier Development

Both the above mentioned approaches included an initial filtering step to reduce the dimensionality of the data. Consequently, two different classifiers were developed. We evaluated and compared both approaches within a 10 times repeated 10-fold nested cross-validation scheme. That means we randomly split our data into 10 folds. While sequentially leaving out 1 out of the 10 folds for testing our models the remaining 90% of the data were used for building the two different machine learning models, as described above. Importantly, all feature filtering was done within the cross-validation procedure.

Based on the cross-validation based evaluation we selected the model yielding a higher area under ROC curve (AUC), which is a measure for prediction performance. 50% AUC

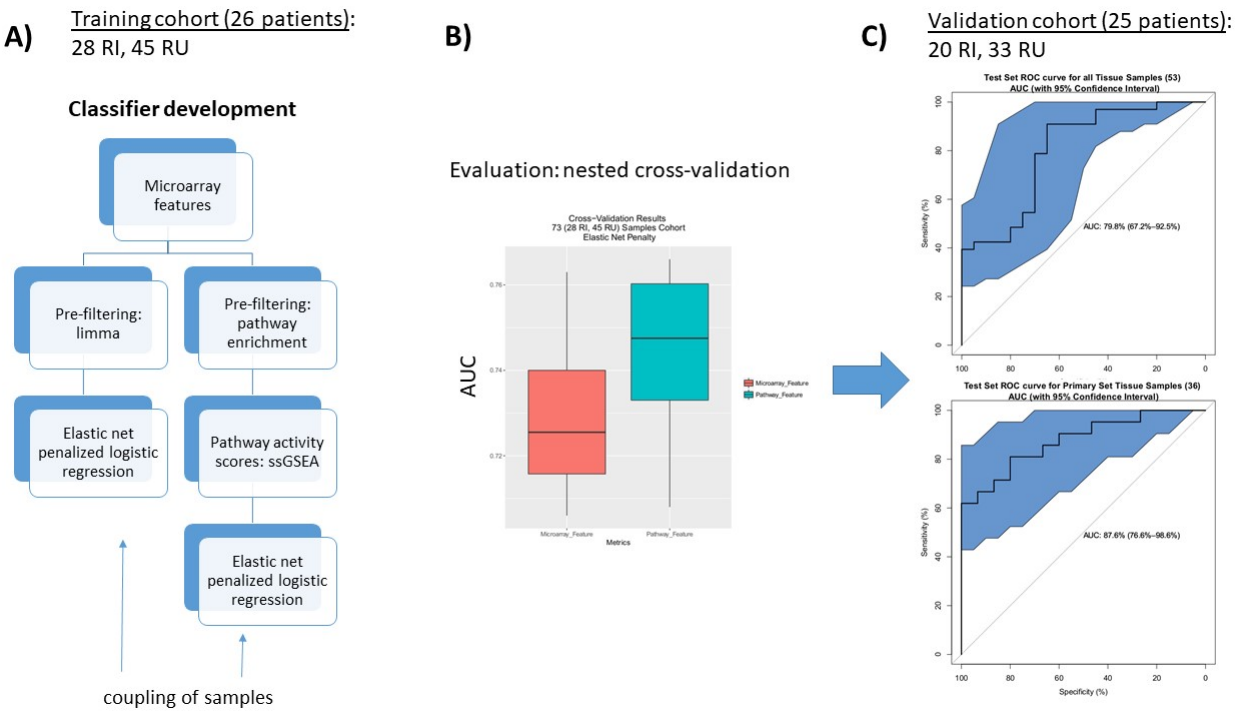

Figure 3.12: Overview about the approach to develop and validate a machine learning classifier to predict the relative location of tumor recurrence (i.e. RI or RU).

indicates chance level and 100% a perfect prediction. The microarray feature based classifier achieved an AUC of 72%, the pathway feature based one of 75% (Figure 3.12B), which yields a preference for the pathway feature based approach. Additional evaluation of a pathway features based classifier with a strategy where sequentially all samples from one of the 26 patients were left out resulted into a higher prediction performance of 87% AUC.

After pre-filtering of features, an elastic-net penalized logistic regression[ZH05] was used to build the classification model using the R package '`glmnet`'[FHT10] . The Elastic-net penalty provides a mix between traditional L1 an L2 penalties with parameters, controlling their respective contributions. These parameters were tuned via a leave-one-out cross-validation. The elastic net penalized logistic regression effectively shrinks coefficients towards zero, thus achieving a sparse model fit, i.e. selecting a subset of features. An important aspect of our data is the fact that several samples exist from the same patient, which leads to statistical dependencies between data samples. In order to account for this aspect, we applied a data augmentation scheme where we added a extra binary matrix to the original data features. This data matrix contained one column for each unique patient ID. The samples are given the value 1 if they come from that patient, else 0. Thus, for every column, all those samples which stem from that patient have 1s, while the other samples have 0s. Regression coefficients for the auxiliary features that had been added in order to augment the data were not penalized in our model, i.e. were not shrunken. The resulting small set of relevant features is henceforth referred to as "signature". When learning the elastic-net classifier based on microarray features we obtained a set of 14 genes while pathway features led to 4 selected pathways. The list of these selected 4 pathways is shown in Table 3.1.

We also evaluated the pathway based classifier on a separate collection of 53 expression

Pathway	Stability	Coefficient
hsa03008: Ribosome biogenesis in eukaryotes	95	0.2746625
hsa03320: PPAR signaling pathway	98	-0.5208185
hsa04512: ECM-receptor interaction	82	0.2106235
hsa00230: Purine metabolism	84	0.0856415

Table 3.1: Pathway signature discriminating RI and RU samples. Column "Stability" refers to the frequency by which the corresponding pathway was selected during a 10 times repeated 10-fold cross-validation. The frequency can range from 1 – 100, where 100 means perfect consistency. Column "Coefficient" reflects the relative contribution of each pathway. A larger magnitude implies more impact on model predictions (more positive = more impact on RU, more negative = more impact on RI).

profiles stemming from 25 patients. (20 RI, 33 RU). 16 patients had gene expression profiles in the validation as well as in the training cohort (the expression profiles are referred to as primary tissue samples), however these profiles were not identical. Evaluations were separately performed on:

- The entire sample collection
- Only primary tissues.

As is typical with transcriptomics data, initially a strong batch effect could be observed between training (containing 73 gene expression profiles) and validation data (containing 53 gene expression profiles). Correspondingly, we corrected for this effect by adding the difference of means between feature values in both datasets. Effectively, this translates the validation set to the mean of the training set and can be understood as moment matching. A graphical depiction is shown in Fig. 3.13.

We obtained an AUC of 80% (95% CI: 67 - 93%) for the entire sample collection and of 88% (95% CI: 77 – 99%) for primary tissues in the validation set. Thus, we have been able to successfully validate the multivariate pathway signature to predict the site of local recurrence of Glioblastoma patients in this retrospective study.

3.5.2 Interpretation and visualization of signature

All four pathways that we obtain as signature (see Table3.1) for our RI vs RU classifier have been linked to GBM and cancer in general in the literature: The most stable feature, PPAR signaling pathway contains the nuclear receptor transcription factor PPARγ which has been found to be expressed in high grade gliomas[EK14]. Changes in ribosome biogenesis have been linked to induce cancer by down-regulating the tumor suppression potential in cells[MTD12]. Furthermore, dysregulation of purine metabolism has been connected to the development of tumor initiating cells in glioma[WYX+17]. Finally, the ECM receptor interaction has been shown to play a key role in the proliferation of glioma cells[UJPK09]. In order to better understand our four Pathway signature in terms of genes which are annotated to the respective pathways, we visualize the fold changes of those genes in the Fig.3.14 where a total of 397 genes are shown with their fold changes (between RI and RU) along with edges connecting them to the respective pathways. We have used Cytoscape[SMO+03] software to visualize

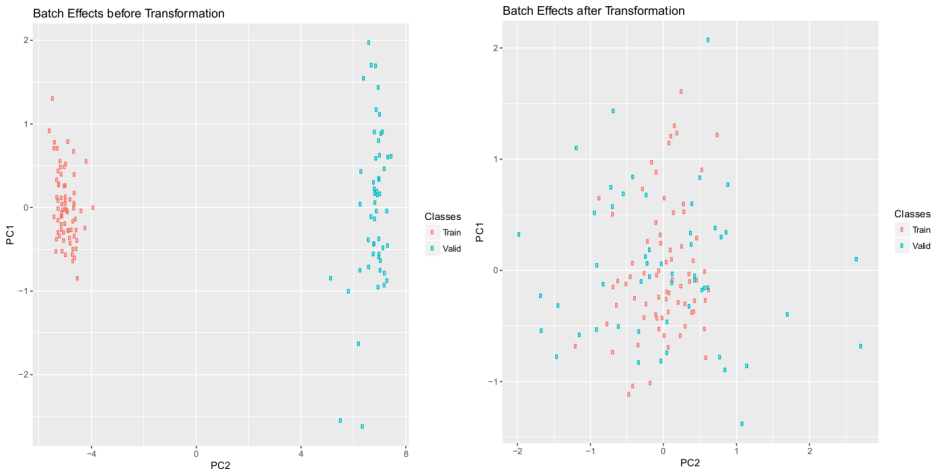

Figure 3.13: The batch effect is clearly observed on the left between the training and validation data sets. On the right is the PCA plot after batch correction

the fold changes of the genes in these four pathways. Cytoscape is an open source software enabling the visualization of complex biological and molecular networks.

Next, we explored the importance of the 4 Pathway signature in terms of its correlation with Progression Free Survival Times (PFS) and Overall survival (OS) in GBM. For this purpose, firstly we first selected patients from the IDENTIREST project which have unique Central Cell biopsies. Overall, 34 patients fulfilled this criterion and had PFS and OS information available. We use a Cox Proportional Hazard's model (CoxPH) as implemented in the `'coxph'` function in the R-package `'survival'` [TL15]. We found that the 4 pathway activity scores were significantly related to both to PFS (p-value 0.003) and OS (p-value 0.005) with fitted C-Index of 0.68 (for PFS) and 0.71 (for OS). Correcting for Age and Gender as additional covariates, we find that the 4 pathway signature significantly improves the base Cox-model (fitted with Age and Gender) in case of both PFS (p-value 0.002) and OS (p-value 0.0004). These results on the IDENTIREST samples suggest that the pathway signature which predicts the spatial recurrence of the tumor is also related to the PFS and OS.

Going one step further, we next used a different cohort of Verhaak et al. [VHP+10a] GBM samples to check if we can ascertain the correlation of the pathway scores (as calculated by the `'ssgsea'` [BTB+09] method) to the PFS and OS. 77 Verhaak patients had PFS information and 342 had OS information. We found that in both cases the pathway signature was significantly correlated with PFS (p-value 0.009) and OS (p-value 0.005). Moreover, after accounting for age and gender, the 4 pathway signature resulted in significant improvements in the CoxPH model fit for PFS (p-value 0.01) and OS (p-value 0.02). Hence, we can say that molecular signatures predicting spatial recurrence also has prognostic value as

measured using PFS or OS.

3.5.3 Clinical potential for the signature

To summarize, we have established the 4 pathway signature to classify RI and RU samples from 73 expression profiles stemming from 26 patients. Additionally, we have retrospectively validated the pathway signature based classifier in a cohort of 25 patients providing 53 gene expression profiles. The natural next step would be a prospective validation study.

If we are successful in prospectively validating our pathway signature based classifier, this would open up its potential use in clinical practice as a prognostic tool. Essentially, such a prognostic tool would be able to predict for future patients the site of recurrence based on their spatial biopsies. Based on the predictions produced by such a prognostic tool, patients would then be provided with targeted radiotherapy, which (hopefully) should be useful to delay (or even stop) the progression of GBM. A clinical study would have to assess the efficacy of such targeted radiotherapy, this would involve comparing the patients who have received targeted radiotherapy to those who receive standard of care treatment with respect to their respective disease progressions. Such disease progressions could be measured by, e.g., comparing the Kaplan Meier curves of the two groups based on their OS of PFS (time to next recurrence). Such a clinical study, if successful, would establish the pathway classifier as a viable prognostic tool in clinical practice and this would be the future direction of research.

In order for our 4 pathway signature based classifier to be used widely as a prognostic tool, specific tailor made assays need to be designed. The most successful example of such customized arrays are MammaPrint® and BluePrint® which are marketed as 'Breast Cancer

Recurrence and Molecular subtyping kit' by the diagnostics company Agendia. A product similar on those lines can also be envisioned for our pathway based classifier which would be useful in clinical management of GBM. Although the original 70-gene signature[VVDVDV+02] for breast cancer was established using microarrays, the successful clinical application lead to the development of the Laboratory Developed Test (LDT) (essentially a customized array) which are then also subject to FDA (or other regulatory agency's) approval. Such approval is based on randomized prospective clinical trials validating clinical utility of these LDTs. The development of customized arrays (and hence LDTs) raise non-trivial challenges like mapping of the gene microarray based classifier to the new customized array classifier (e.g., RT^2 Profiler PCR Arrays from QIAGEN which provides measurement of some hundreds of pathway-focused genes). Furthermore, clinical trials and regulatory approval is a very time and cost demanding process that needs to be assessed, planned in terms of and in relation to the differential clinical benefit of such LDTs to the GBM patients in clinical practice.

Thus, a road-map has been laid out for the future direction of the work for bringing the RI vs RU pathway based classifier from the field of research to a commercially viable and clinically effective prognostic subtyping toolkit.

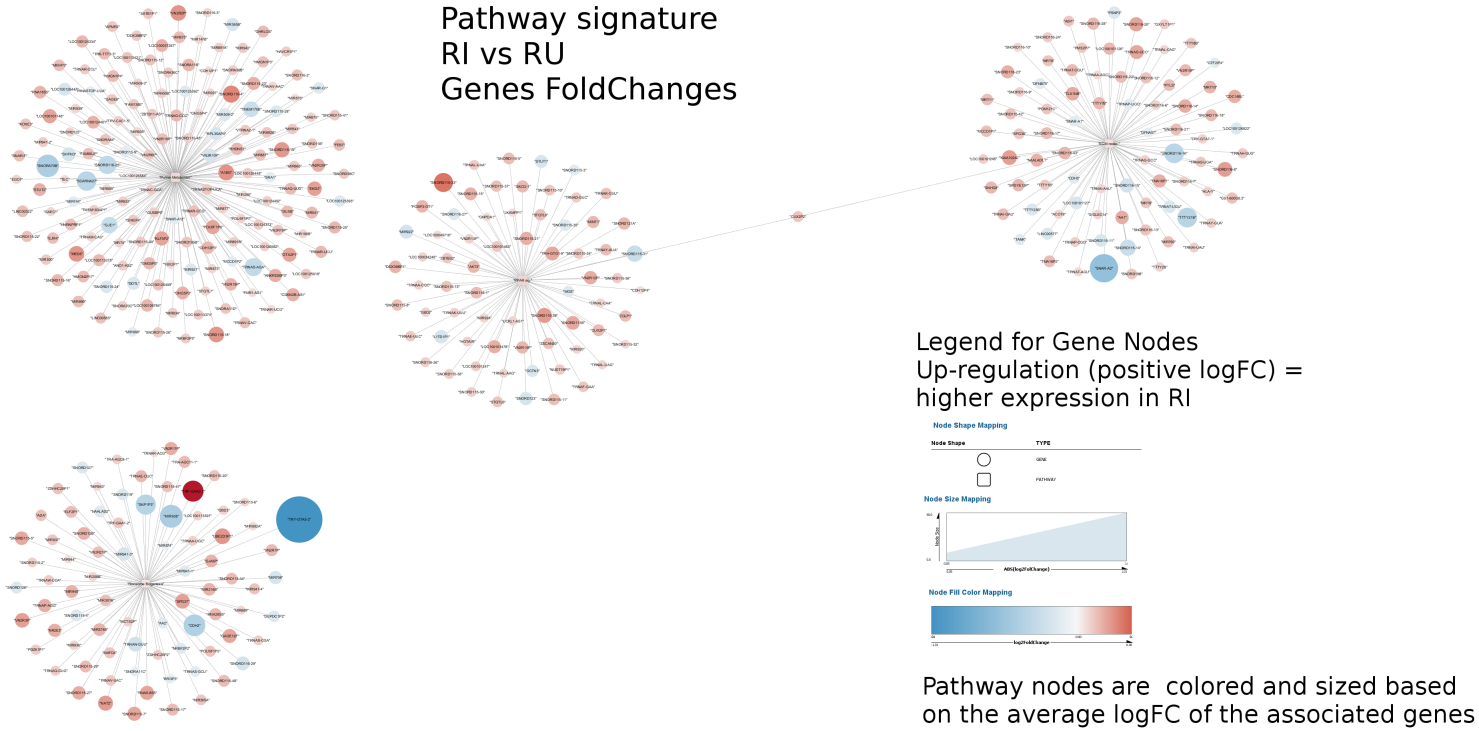

Figure 3.14: Graphical depiction of the 4 Pathway signature along with the constitutive gene sets. All the genes annotated to the 4 Pathways,which also have corresponding AffyIDs on the microarray, have been depicted with their respective fold changes

"Doubt is the origin of wisdom"

Rene Descartes

4

SBC – A novel technique for patient stratification

4.1 Motivation for SBC

As explained earlier, the goal of personalized medicine is an individually optimized patient treatment. This idea typically implies a stratification of patients into sufficiently homo-

geneous sub-populations. In that context, characterization of disease sub-types is of high relevance. Some early land-mark studies of patient stratification based on gene expression data set ([AED+00], [BKH+02], [VVDVDV+02], [LLH+04]) have generated a lot of interest in this direction. Disease subtype identification with an emphasis on patient survival prediction, which is highly relevant to the promise of individual therapies, can be approached using either the molecular -omics data alone (fully unsupervised) or based entirely on patient survival data (fully supervised) by dichotomizing the patients into predefined groups like ”low-risk” and ”high-risk” and then using standard discriminative analysis tools like support vector machines (SVM), linear discriminant analysis (LDA) or multinomial regression [SRT+02] to make predictions. The success of the supervised approach thus critically depends on the a priori definition of patient sub-groups and the correspondence of these groups to molecular data.

Unsupervised clustering (such as hierarchical clustering - [YFM+01]) on the other hand focuses on discovery of molecular separable disease sub-types without any clinically motivated a priori definition of patient sub-populations. Once the disease sub-types are established a post-hoc analysis explores differences of the sub-types with respect to the clinical outcomes. However, the aforementioned method could discover subtypes which may not be related to survival or other clinical outcomes, as is evident in [VHP+10b]. This concern was first highlighted by [BT04] and later by [KMC+10].

To address the shortcomings of traditional supervised and unsupervised approaches [RFW+10] propose a Bayesian Infinite Mixture of Experts Model to cluster patients with respect to their survival outcomes. Their model, in addition to determining main effects of the genes, also gives an insight to their higher order interactions in different clusters. However, this is achieved at the cost of discretizing continuous variables which – of course – leads to loss

of information of continuous molecular data. Furthermore, their approach may suffer from non-interpretability when patient groups with different survival outcomes have near identical molecular profiles, thus failing to provide biological explanations for survival.[BT04] propose a semi-supervised clustering which combines both gene expression data and clinical end-point data. They first identify a set of genes that are significantly correlated with survival time (using univariate Cox regression), then subsequently apply an unsupervised clustering technique (Nearest Shrunken Centroids) with the obtained set of genes. Risk predictions are made by using the principal component scores of the above mentioned set of genes. Although successfully used in many applications,[BT04] approach also has some limitations. For example, the algorithm requires to pre-specify the number of disease subtypes, which can be difficult in practice. Furthermore, the principal components of a set of genes to predict continuous risk scores can be difficult to interpret. Finally, uni-variate gene selection can fail, if multiple genes have a joint significant effect on survival, but marginal effects are weak.[KMC+10] as a further development to this approach propose a Recursive Partition Mixture Model (RPMM) which successively fits models with varying number of clusters (K) and uses modified Bayesian Information Criterion to efficiently estimate K. Moreover the model also selects the optimal gene set M. The key idea for feature and model selection is to train the model on top-ranking genes and to check the separability of the survival curves on an independent test set. At the end that gene set M is selected which gives the lowest possible p-value. Although computationally attractive, the results of feature selection and cluster number determination are heavily dependent on how the whole data set is split into training and testing.

In this Chapter we try to overcome several of the above mentioned limitations of present

techniques. More specifically our proposed Survival based Bayesian Clustering (SBC) approach has the following features[AF17] :

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

4.2 Proposed Approach

Our SBC approach rests on the foundations of Bayesian model-based clustering and sparse bayesian survival curve estimation. We first motivate the use of these two methodologies: While a particular appealing property of model-based clustering (here using Gaussian mixtures) is to naturally deal with uncertainty regarding cluster assignment of patients, the Bayesian framework, in addition, allows for an elegant way to circumvent the model selection problem, i.e. to decide for a particular number of clusters. More specifically,in our present work, we build on previous work in the machine learning community on Infinite Gaussian Mixture Model (GMM)[Ras99]. Infinite GMMs are based on a Dirichlet Process (DP) prior over parameters[Nea00]. The DP priors define a probabilistic model for data generation and for cluster assignments. The most important characteristic of infinite GMM is Bayesian Model Averaging which allows us to estimate the posterior distribution over the number of clusters, thus avoiding the need to compare separately fitted GMMs[GR10], [Nea00].

In the past several attempts have been made to use DP models for clustering gene expression profiles by[MS02], [MYB04] and more recently by[YSM11]. [MYB04] compare the performance of the infinite model to the finite model case in a simulation setting and find it advantageous to use infinite model especially in the case of high noise. Motivated by these findings and its inherent flexibility we choose Dirichlet Process Gaussian Mixture Models (DPMM) for modelling the expression profiles in our work.

One of the key innovations of the present work is the additional inclusion of cluster-specific survival models in the DPMM. We use Accelerated Failure Time model (AFT) with the log-normal assumption[Roy01] to model the survival or progression free survival times of the patients. The choice of the AFT model as opposed to the Cox Proportional Hazards model was made due to the ease of casting the AFT model in a Bayesian setting. We model the AFT as a Bayesian LASSO[PC08] to identify potential biomarkers which are related to survival times. Apart from cluster-specific sparse survival models two further key innovations in our approach are:

- Data Integration : We extend our mixture model to more than one data source (e.g. gene expression plus miRNA expression). As opposed to existing work our approach thus combines multi-omics and survival information to cluster patients.

- Prediction: In contrast to unsupervised clustering methods, our model can be used to make survival as well as class predictions (sub-type) of new patients. In contrast to supervised methods we do not need to know patient sub-types in advance.

4.3 Method Details

4.3.1 Dirichlet Process Mixture Model

Dirichlet Process (DP) Mixture models belong to the broad category of Bayesian Non-Parametric methods. They allow for the inference of countable (possibly infinite) number of mixture components K. Dirichlet Process were first introduced by [Ant74] and [Fer73]. If we assume $X_1, X_2, \ldots X_N$ to be N data points drawn independently from some unknown distribution, where X_i can be multivariate or categorical, then the Dirichlet Process Prior models the density of X_i in the following hierarchical fashion:

$$X_i|\vartheta_i \sim F(\vartheta_i)$$

$$\vartheta_i|G \sim G$$

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

F is the conditional distribution of X_i which is parametrized by ϑ_i. G is the posterior mixture distribution which is mostly marginalized when inferring DP mixture model. G_0 is the base distribution and represents prior information about the parameter values. The parameter α is known as the concentration parameter and controls the number of clusters that we obtain from the posterior distribution. The marginalized prior representation was obtained by [BM73] by representing it as series of conditional distributions:

$$\vartheta_i|\vartheta_1, \ldots \vartheta_{i-1} \sim \frac{1}{i-1+\alpha}\sum_{j=1}^{i-1}\delta(\vartheta_j) + \frac{\alpha}{i-1+\alpha}G_0$$

DP can be also be thought of as a distribution over distributions. It is known that Dirichlet Process Prior can also be obtained by taking the limit of a finite mixture model with K, the number of clusters going to infinity i.e. $K \to \infty$. It was shown by[Nea00] that by introducing class labels c_i for each data point the old formulation of[BM73] can be re-written as following:

$$X_i | c_i, \vartheta_i \sim F(\vartheta_{c_i})$$

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

$$\vartheta_i \sim G_0$$

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

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

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

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

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

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

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

$$\vartheta_i|\vartheta_1, \ldots \vartheta_{-i} \sim \frac{1}{N-1+\alpha}\sum_{j\neq i}\delta(\vartheta_j) + \frac{\alpha}{N-1+\alpha}G_0$$

The parameter α (also know as the concentration parameter) controls the prior number of expected clusters. Following[GR10] it has been given an inverse gamma prior expressing the belief that apriori we do not expect a large number of clusters.

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

For further details one can refer to[Nea00].

4.3.2 Hierarchical Multivariate Gaussian Model

As a choice for the base distribution G_0, we use a hierarchical Gaussian Model. Our Hierarchical Multivariate Gaussian mixture model (referred to in this work as DPMM) follows closely the work of[GR10]. The conjugate Gaussian Mixture model can be described with the following sets of equations:

$$X_i|(c_i = j) \sim \mathcal{N}(\mu_j, S_j^{-1})$$

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

where X_i indicates a D- dimensional vector of measurements (e.g. gene expression profiles) for patient i. Furthermore, μ_j is the centre of cluster (or sub-type) j, described via a multivariate Gaussian with precision matrix S_j. The second equation constitutes a prior distri-

bution for μ_j which itself is a normal distribution with expectation ξ and scaled precision matrix ϱS_j. We regularize the precision matrix towards a diagonal matrix W as in [BF13] :

$$(S_j|\varphi, W) \sim \mathcal{W}(\varphi, (\varphi W)^{-1})$$

where $\mathcal{W}$ denotes a Wishart distribution with φ degrees of freedom. Empirical Bayes estimates as described in [GR10] are used as priors over the hyper-parameters (ξ, W etc.). We use a conjugate Dirichlet Process Gaussian Mixture Model as it allows for the possible marginalization of the cluster-specific parameters. More concretely, the joint distribution of the mean μ_j and the precision matrix S_j follows a Normal/Wishart distribution

$$(\mu_j, S_j) \sim \mathcal{NW}(\xi, \varrho, \varphi, \varphi W)$$

The parameter ϱ controls the strength of the dependence between the mean μ_j and the precision S_j; while φ controls the dependence between the hyperprior W and precision matrix S_j. We used the following distribution priors:

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

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

As φ controls the degrees of freedom in the Wishart distribution, it has been constrained as $\varphi > D - 1$. The hyper-parameters ξ and W are given priors based on the empirical Bayes

estimates μ_y and Σ_y from the data:

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

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

4.3.3 Bayesian LASSO penalized Accelerated Failure Time Model

The Accelerated Failure Time (AFT) model has been extensively used to model survival times of cancer patients, modelling either the survival probabilities or time to recurrence probabilities[Wei92]. In its most general form an AFT model is given by:

$$log(t_i) = \beta_0 + \beta^T X_i + \varepsilon_i, i = 1, \ldots N$$

where $log(t_i)$ is the log survival time (or progression free survival), and β is the vector of regression parameters. As most likely only a subset of features is truly associated to survival, we place a Laplacian prior over β which effectively induces a L1 penalty on the regression coefficients and penalizes small effects to exact zero. Following a Bayesian approach we place a a diffuse gamma-prior on the penalty strength parameter, λ, and evaluate its posterior.The hierarchical formulation of the Bayesian Lasso[PC08] is:

$$log(t_i)|\beta_0, X_i, \beta, \sigma^2 \sim \mathcal{N}(\beta_0 + \beta^T X_i, \sigma^2),$$

$$\beta|\sigma^2, \tau_1^2, \tau_2^2 \ldots, \tau_p^2 \sim \mathcal{N}(0_p, \sigma^2 D \quad),$$

$$D \quad = diag(\tau_1^2, \ldots \tau_p^2)$$

Our treatment of the censored patients follows closely to that of[STV06]. The key idea is to treat the censored outcomes $I_i = 0$ as yet another unknown parameter w_i and to use our probabilistic model to estimate the censored survival times. We augmented the survival times with pseudo variables w_i which are defined as follows:

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

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

For the case of censoring, w_i is assumed to be drawn from a left truncated normal distribution, with the left truncation at the censored survival time[STV06]. The Bayesian LASSO penalty amounts to placing a Laplacian prior on the coefficient matrix β of the following form :

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

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

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

The values for $r = 1$ and $\delta = 1.78$ were set as in[PC08]. The parameter σ^2 was given an inverse-gamma prior. The R-package *blasso* was used to sample parameters.

4.3.4 Bayesian Regularization

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

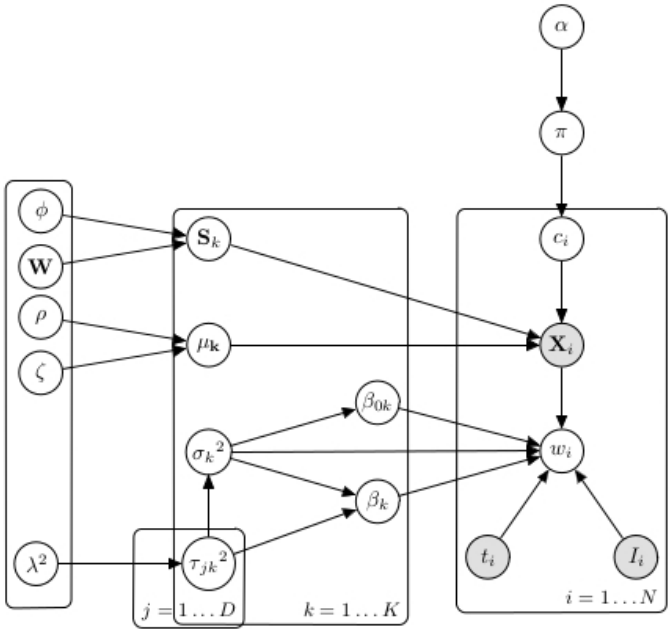

Figure 4.1: Graphical Model representation for **SBC**

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

4.3.5 Model fitting via Gibbs Sampling

SBC can be depicted as a graphical model, as shown in Fig4.1. The hierarchical model formulation and the use of conditionally conjugate model enables the use of a Gibbs sampling based algorithm for parameter estimation. The cluster indicator variables c_is are updated

using the following conditional distribution for those components which have non-zero elements i.e. $n_{-i,j} > 0$:

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

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

for all others combined we can sample from the conditional distribution as detailed below. As the assignment of a new cluster involves marginalization over mixture model parameters, this integral turns out to be non-tractable in our case. In order to circumvent this problem, we use the auxiliary variable method used in Algorithm 8 of[Nea00] with the number of auxiliary variables set to two as described in[GR10]. The conditional distribution of a data point to belong to a new cluster is as follows:

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

$$\propto \frac{\alpha}{N-1+\alpha}\int$$

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

As there is overall dependence between the parameters of the Gaussian mixture model (μ, S) and that of the BLASSO $(\beta_o, \beta, \sigma^2)$, the above integral is not easy to solve. This makes the overall Mixture Model non-conjugate and we resort to[Nea00]'s auxiliary variable approach to sample from the above distribution. The key idea is to able to approximate the above integral's value by drawing auxiliary parameters from the prior-distribution and considering the problem to be temporarily a finite mixture model. In our case, we found that $U = 2$ auxiliary parameters are sufficient for good convergence. The model is fitted us-

ing an alternating Gibbs update scheme for cluster-specific parameter set $(\mu_j, S_j, \beta_{oj}, \beta_j, \sigma_j^2)$ and the class labels $(c_1, \ldots c_N)$ each of which can now be sampled from it's conditional distribution. More details of the exact sampling algorithm can be found in Appendix B. For our Gibbs Sampling we use 100 burn-in iterations and 200 MCMC samples with samples being drawn every 5th iteration (thinning). To assess the convergence of our MCMC chain, we looked at the log-likelihood trace plots. To get estimates for cluster membership of patients we use the mode of marginal posterior distribution of each of the class labels from our Gibbs sampling.

4.3.6 Feature Importance

The hierarchical formulation of the SBC allows us to define a ranking over the discriminatory ability of each feature with respect to two clusters (a & b). We define the ranking r^i of a feature i as in [YH11]:

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

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

4.3.7 Data Integration

Our present model can be extended to integrate more than one -omics data source. In our present work, we use data sources which have continuous (Gaussian) values. These are then all modelled as described above. To combine several data sources $(v = 1 \ldots Q)$ we compare two different strategies: a) one in which we work with independently pre-filtered feature

sets from each of the data source $X_i^{(v)}$ and b) one in which we perform a Canonical Correlation Analysis (CCA) on the original (pre-filtered) features and then map data from each data source on the top canonical covariates ($X'^{(v)}_i$).

CCA is a classical data integration method ([Hot36]) which is used to extract concordant feature sets. Each of the canonical covariates is constructed to successively explain maximal correlation between linear feature combinations from two or more data sources. After we obtain the feature sets from the above mentioned two methods we assume that the complete model likelihood of feature sets of a patient given its cluster membership, can be factorized as:

$$p(X_i^{(1)}, X_i^{(2)}, \dots X_i^{(Q)} | c_i = j, \mu_j^{(1)}, S_j^{(1)}, \dots \mu_j^{(Q)}, S_j^{(Q)})$$

$$= \prod_{v=1}^{Q} \mathcal{N}((X_i^{(v)} | \mu_j^{(v)}, S_j^{-1(v)})$$

where $X_i^{(v)}$ denotes features of the i-th patient from the v-th data source with features which come either from a) or b). We further suppose a factorization of the AFT mode across data sources as:

$$log(t_i) | [X_i^{(1)}, \dots X_i^{(Q)}, c_i = j, \beta_{oj}^{(1)}, \beta_j^{(1)}, \sigma_j^{2(1)}, \dots \beta_{oj}^{(Q)}, \beta_j^{(Q)}, \sigma_j^{2(Q)}]$$

$$\sim \prod_{v=1}^{Q} \mathcal{N}(\beta_{oj}^{(v)} + \beta_j^{T(v)} X_i^{(v)}, \sigma_j^{2(v)})$$

This essentially means that each data source has its own cluster-specific AFT model as described Section 4.3.3 and a weight that depends on the likelihood of observing the clinical endpoint with features from that data source .The cluster indicator of a patient sample c_i, as in the one data source case, is given a Dirichlet Process Prior. We call this approach as in-

tegrative SBC or iSBC. As an illustration, we have shown in Fig 4.2 the iSBC model with $Q = 3$. To simplify notation, all parameters of the Hierarchical Multivariate Gaussian are represented by $\Theta_k^{(v)}$ and those of the the Bayesian LASSO penalized AFT model are denoted by $B_k^{(v)}$.

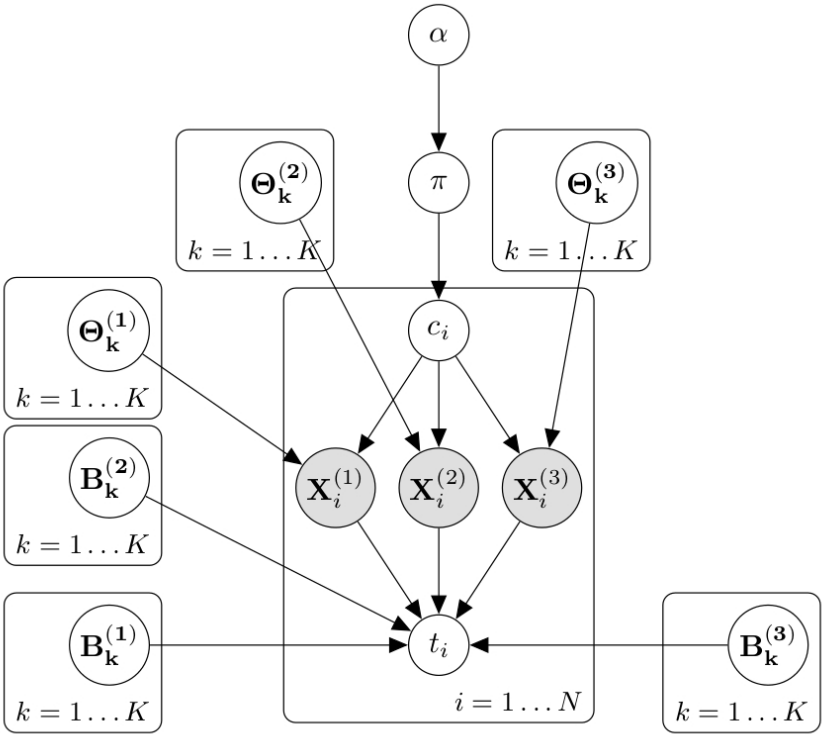

Figure 4.2: Graphical Model representation for **iSBC** with $Q = 3$ data sources.

4.3.8 Making Model Predictions

Given a already trained SBC model with parameters $[\vartheta_{1:N}^{(m)}, c_{1:N}^{(m)}]$ over M MCMC samples and molecular data X^* for test patient, we would like to solve two predictions problems a) survival prediction and b) prediction of cluster membership. For the sake of simpler nota-

tion, X^* is assumed to be of one specific -omics type, but the same approach also works for multi-omics data.

Survival prediction

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

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

where C_m denotes the number of clusters found in MCMC sample m. Notably, each MCMC sample corresponds to one full parameter set of our model. Hence β_{0jm}, β_{jm} denote the regression parameters in our AFT model for MCMC sample m. Moreover $v_{jm}(X^*)$ is the weight that is dependent on the likelihood of X^* to belong to cluster j for the MCMC sample m . The weights $v_{jm}(X^*)$ for the discovered clusters $j = 1 \ldots C_m$ in the MCMC sample m are proportional to their corresponding densities :

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

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

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

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

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

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

$$(\mu_u, S_u) \sim \mathcal{NW}(\xi, \varrho, \varphi, \varphi W)$$

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

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

Cluster membership

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

bility of the DP model:

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

where n_{jm} is the number of patients in the cluster j and μ_{jm}, S_{jm} are the corresponding cluster parameters of the Hierarchical Multivariate Gaussian model for the m-th MCMC sample, b is a normalization constant. Apart from the already discovered clusters, the Dirichlet Process prior also places non-zero probability for the test point to form a new cluster. This probability is given by:

$$p(c^* = c^{new}|X^*, \vartheta_{1:N}^{(1:m)}, c_{1:N}^{(1:m)}) = b\frac{\alpha}{N-1+\alpha}\int \mathcal{N}(X^*|\mu_o, S_o^{-1})d(G_o)$$

where (μ_o, S_o) are drawn from their prior distribution G_o given by the Hierarchical Multivariate Gaussian Model as described above. In-order to avoid solving this integral, we again use the auxiliary variable approach of[Nea00] to approximate the above probability. The details of the auxiliary variable approach are the same as for the above section on Survival prediction. This means that apart from the existing classes, the patient X^* can form a new cluster with the probability:

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

for auxilary variables $u = 1, 2$.

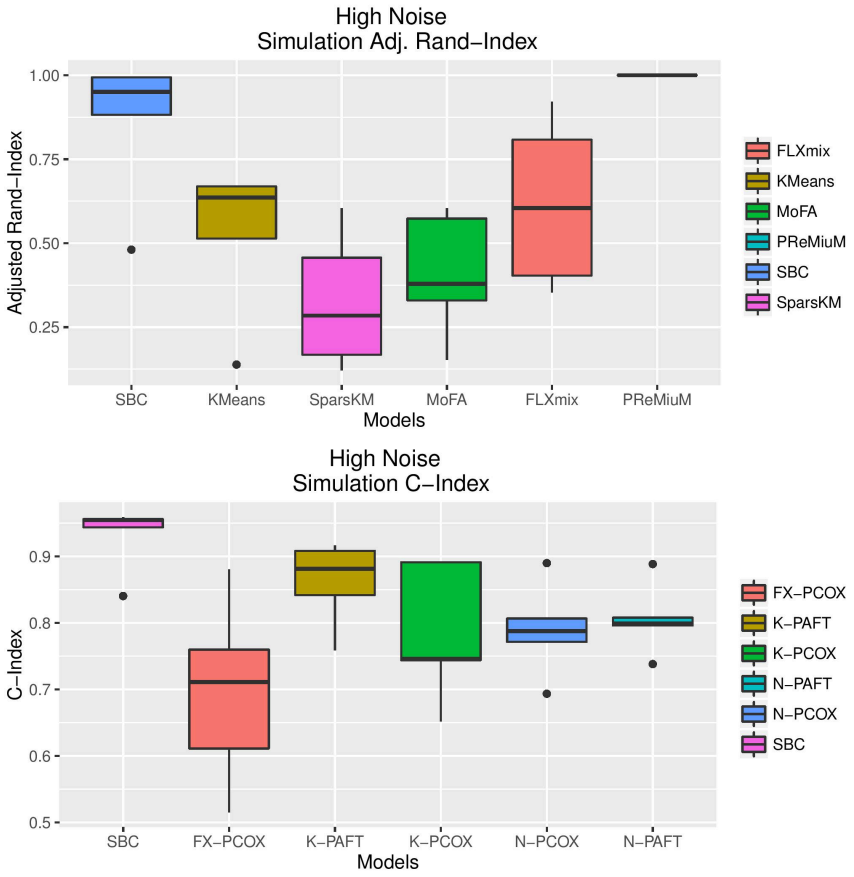

Figure 4.3: Simulation results on the training set using SBC and the high noise scenario and D=20

4.4 Simulation Study

We investigated the performance of our Survival based Bayesian Clustering (SBC) model in various simulation settings. We simulated molecular data as multivariate Gaussians with non-trivial correlation structure and varying degrees of overlap using the *MixSim* R-package. The package provides a list of D dimensional cluster-specific mean vectors and $D \times D$ cluster-specific full covariance matrices. Since SBC – as well as competing methods – are typically applied on a pre-filtered subset of all features ($10 \leq D \leq 60$), we also investi-

gated the robustness of our model when a certain fraction of noise features (20 percent and 50 percent of all features) were added, which did not contribute to the clustering structure. Finally, the molecular data was obtained by concatenating the relevant and noise features. For the survival data generation we used cluster-specific log-normal AFT models applied on the molecular data and then added cluster-specific Gaussian noise.

4.4.1 Data Generation

We simulated $N = 100$ data points, each for training the model and for testing it. We repeated the whole simulation process 10 times and compared our results with other competing approaches. We simulated the cluster-relevant features using cluster-specific parameters (mean vector and precision matrix) employing the *MixSim* R package[MCM12]. We have conducted the simulations in two scenarios to explore the effect of noise

1. The low noise scenario where 20 percent of the features were noise (uninformative for clustering) and there was 1 percent cluster overlap in the informative feature space.

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

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

4.4.2 Simulation results and comparisons

We initialized the model with k-means estimate by choosing k with the help of silhouette plots. The superior results of our SBC model in comparison to other methods (also in high noise setting) demonstrate the need of integrating the survival times in clustering. For comparison of our SBC model with other competing models we used FLXmix [GL08], k-means, Mixture of Factor Analyzers(MoFA) [MP00], PReMiuM [LHA+13], sparse k-means (SparsKM) and sparse hierarchical clustering (SparseHC) [WT12]. We give a short summary of the competing clustering methods :

1. FlXmix [GL08] is a curve clustering algorithm. FlexMix implements a general framework for fitting discrete mixtures of regression models. It allows the integration of Generalized linear models and penalized models. It uses an EM algorithm to estimate the parameters. We used FLXmix with *glmnet* R package for high-dimension regression. The clusters are initialized using a standard k-Means algorithm and the number of clusters are chosen based on the Bayesian Information Criterion (BIC).

2. PReMiuM [LHA+13] is a package for Bayesian clustering using a Dirichlet Process Mixture Model. It allows for both continuous/discrete variables response variables but does not deal with survival information. To implement our censored response variables, we disregarded censoring and considered the response as continuous. It also allows to make predictions. The number of clusters are discovered automatically using Dirichlet Process prior.

3. MoFA (Mixture of Factor Analyzers) [MP00] is a model-based density estimation to take into account noise in high dimensional data sets. We set the number of factors to be

two and selected the number of clusters using BIC.

4. SpaseHC (Sparse Hierarchical Clustering) and Sparse-KM (Sparse K-means Clustering) are two algorithms in the R package *'sparcl'*[WT12]. These two methods provide a principal way to deal with noisy data. The number of clusters are optimized by maximizing the average cluster silhouette width.

5. K-Means - For the case of two data sources we created a concatenated data matrix by joining the columns of the two data sources and running K-Means on the joint matrix. To choose the number of clusters we looked for clusters which maximized the average silhouette width of the clusters.

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

The two measures used to compare our results were the C-index[HCP+82] and Adjusted Rand Index. Rand Index measures the agreement between two clusterings, it ranges from 0 (no agreement) to 1 (full agreement), the adjusted Rand Index also corrects for chance groupings and can have negative values (indicating worse than chance agreement). The C-Index (or Concordance Index) is used to assess prediction performance in survival analysis and is akin to Area-Under-Curve (AUC) in the classification case. To compare purely unsupervised clustering methods, such as k-means against our SBC approach with respect to survival predictions on training data we applied a two-step strategy: first clustering and then fitting cluster-specific survival curves using either a lasso penalized AFT or Cox model. We call the corresponding algorithms as K-PCOX (K-means clustering followed by cluster-specific penalized Cox regression), N-PCOX (Penalized Cox regression disregarding any

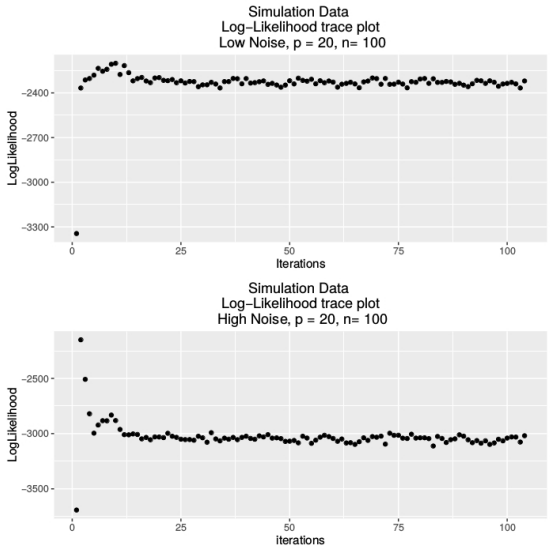

Figure 4.4: Likelihood trace plots during the burnin period for the low and high noise scenarios

clustering), K-PAFT (K-means clustering followed by cluster-specific penalized AFT) and N-PAFT (Penalized AFT disregarding any clustering).

Our SBC algorithm generally achieves a higher adjusted Rand Index, C-index than competing methods.

4.4.3 Assessing the convergence of the Gibbs Sampler

In-order to assess the convergence of the MCMC sampler, we use log-likelihood trace plots. In Fig. 4.4 we show two such plots for the case of low and high noise scenarios with $D = 20$. In all our simulations we found 100 burn-in iterations to yield a convergent MCMC chain. We then used 200 MCMC samples for our posterior estimation (with samples being taken every 5th iteration).

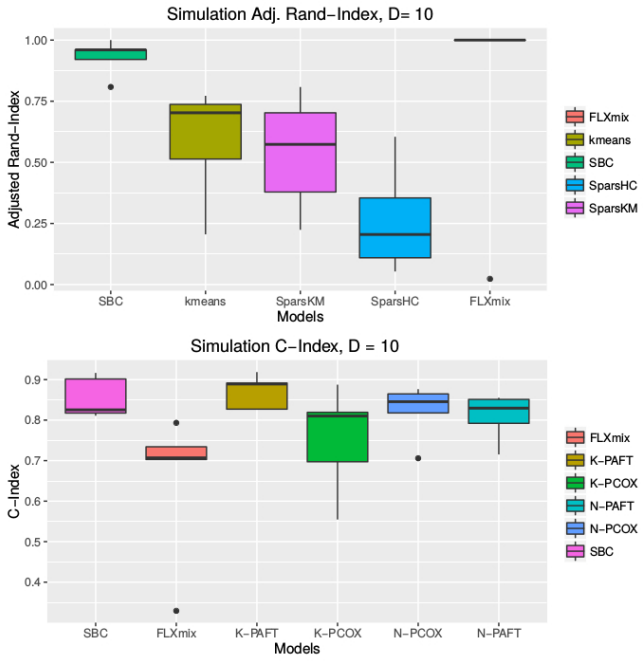

Figure 4.5: Simulation results on the training set using SBC and the high noise scenario and D = 10.

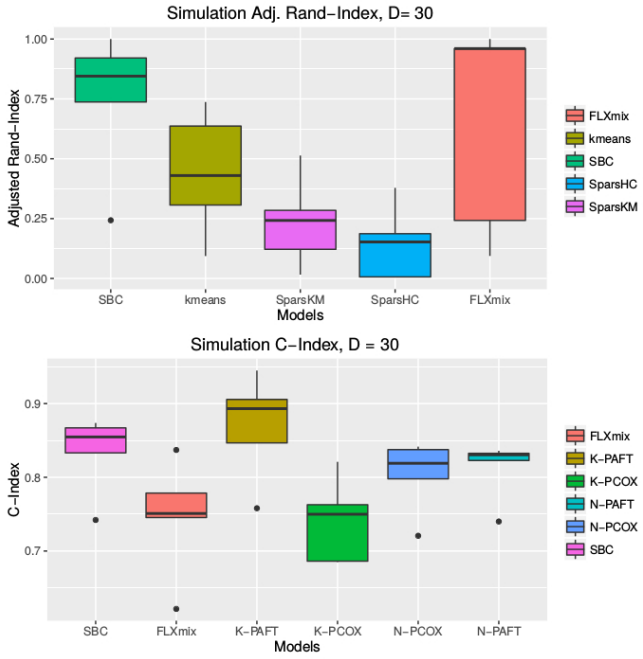

Figure 4.6: Simulation results on the training set using SBC and the high noise scenario and D = 30.

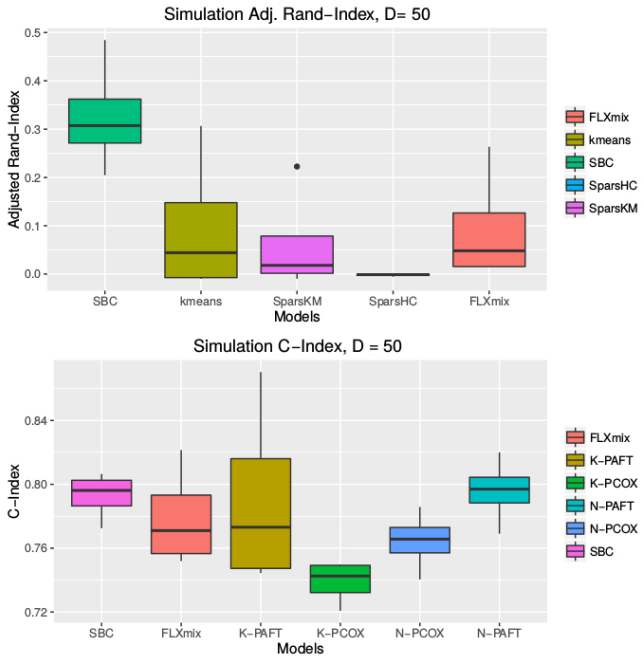

Figure 4.7: Simulation results on the training set using SBC and the high noise scenario and D = 50.

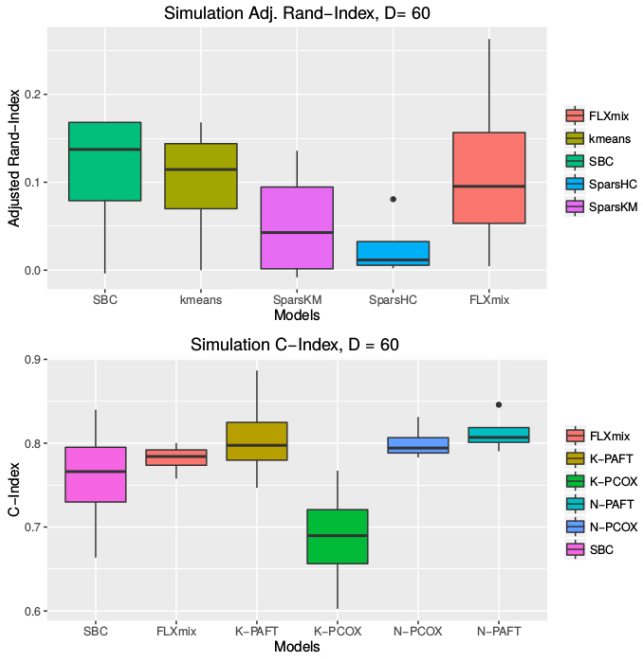

Figure 4.8: Simulation results on the training set using SBC and the high noise scenario and D = 60.

4.4.4 Effect of varying dimension D

We also varied the dimension (number of features) of the SBC, apart from the $D = 20$, we tested for $D = 10, 30, 50, 60$ shown in Fig.4.5,4.6,4.7,4.8. As expected, increasing dimension and fraction of irrelevant features had a negative influence on SBC performance, but altogether the advantage over competing methods still remained. This held true also for detecting truly relevant features.

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

4.4.5 Feature Importance from SBC model

As discussed above, our SBC model enables us to rank features on their ability to distinguish clusters. In our simulations we can compare the performance of the SBC to detect relevant features. From our SBC model we can get scores for the relevance of each feature which are either "relevant" or "non-relevant" and thus we can calculate the average Area Under the curve (AUC) for this classification. This has been shown in Fig.4.9 where we contrast this with penalized FlXmix[GL08]. The results represent 5 simulation repeats and

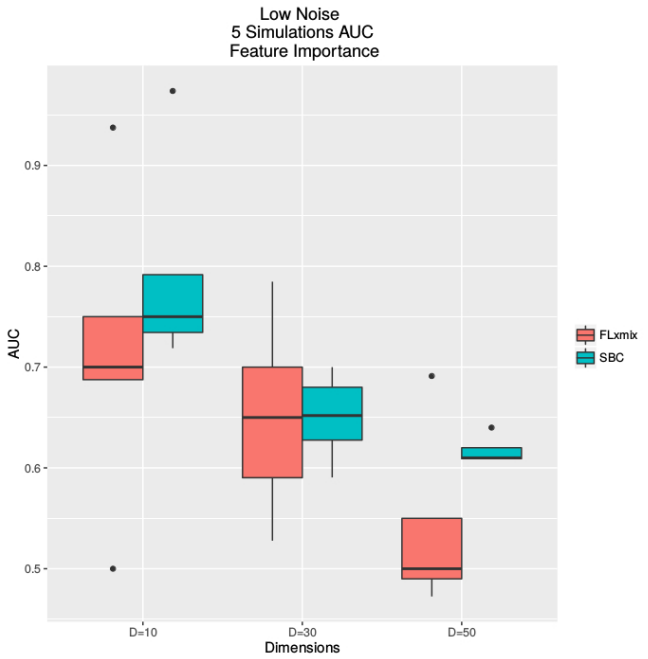

Figure 4.9: Simulation results on the training set for detecting feature importance in the low noise scenario

we restrict ourselves to the "Low Noise scenario". One can see a similar trend to the above simulations with increasing D leading to a deterioration of the model performance.

4.5 Real Data

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

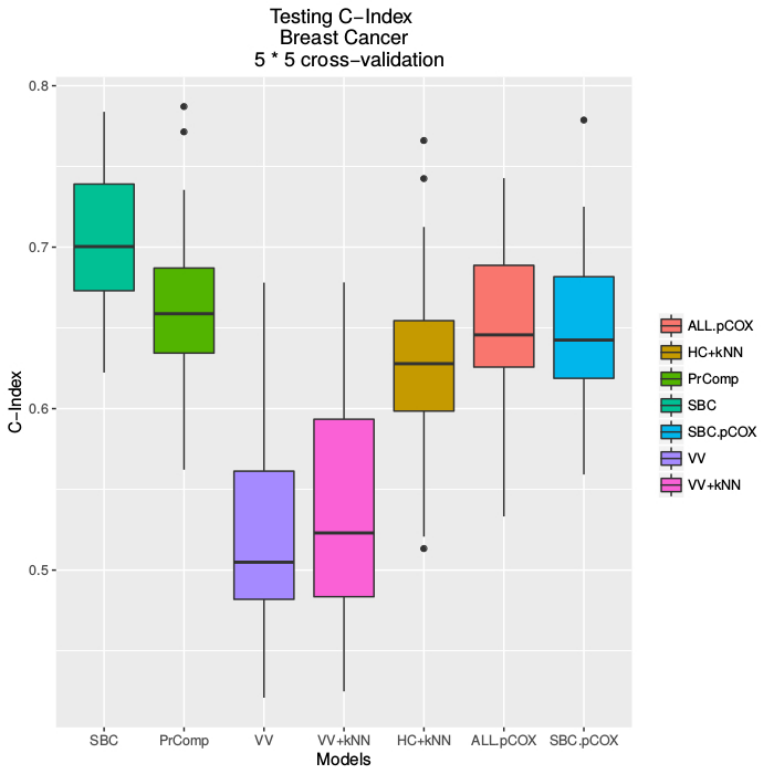

Figure 4.10: Results on the Breast Cancer data set. Box plots depict cross-validated C-indices for different methods.

4.5.1 Breast Cancer

We used the breast cancer microarray data set used in [VDVHV+02] and available through the `seventyGeneData` R-package. For the clinical endpoint we used "time to metastatis" along with the corresponding censoring indicator for metastatis. The authors classified the data into two groups, we call this clustering as Vijver classification (referred as VV). The 70-gene signature [VVDVDV+02] was used to compare with our approach to stratify 295 patients in terms of our clinical end-point. In order to reduce the dimensionality of the data we pre-filtered genes according to two criteria: a) using the most significant p-values from univariate cox-regression models and b) using a t-test between metastatic and non-metastatic groups. Tak-

ing the intersection of these two ranked sets we arrive at a pre-filtered list of genes which is subsequently referred to as 'SBC signature'. Notably, the same pre-filtering was also applied to two of the competing methods to ensure fair comparison (see below). Our SBC approach outperformed the following competing methods for survival-prediction during cross-validation procedure (measured using C-Index) (see Fig 4.10):

- An average linkage hierarchical clustering (HC) of patients on the training data (using the SBC signature) within the cross-validation procedure followed by k-nearest neighbour (k-NN) predictions for the cluster membership on the test data and survival predictions by a penalized Cox regression model (pCOX). This approach was taken in the spirit of van't Veer et al. (abbreviated as HC+kNN).

- The same setting, but with original grouping of patients according to Vijver et al. (VV) together with the 70 gene signature and then followed by k-NN together with pCOX (abbreviated as VV+kNN)

- Using classification by Vijver et al. on the training and test sets and building cluster-specific pCOX models (abbreviated as VV)

- Taking the first 20 principal components of the whole set of features on the training data, within the cross-validation procedure and using a pCOX. That means test data within the cross-validation procedure was first projected on the first 20 principal components constructed on the training data, and then survival predictions were performed via a pCOX model. (abbreviated as PrComp)

- A single L1-regularized Cox regression model (disregarding clustering) on a) the whole set of features (ALL.pCOX) and b) on the pre-filtered SBC features (SBC.pCOX)

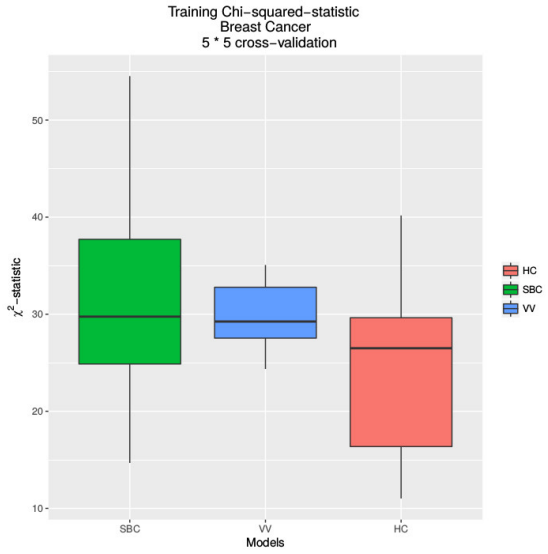

Figure 4.11: Cross-validation results for Breast Cancer. Log-rank statistic is based on the **recovered classes** from the SBC model on the training set

Figs.4.11,4.12 in addition indicate that SBC yields a separation of survival curves in different clusters that was at least as good as that obtained with competing stratification approaches (hierarchical clustering, original VV grouping). Instead of the hierarchical clustering, we also tried the k-means clustering and the results were similar. Depicted in these two figures is the test-statistic for the log-rank test comparing estimates of the hazard functions associated to the detected clusters. The test statistic is constructed by calculating the observed and expected number of events in each cluster at each observed time. A large value of the test statistic indicates a stronger deviance from the null hypothesis of no difference in the hazard functions of different clusters.

Next, we demonstrate the results obtained with our SBC method when training the model on a randomly chosen subset of 50% of the samples. Our SBC signature for this split comprised of 58-probe IDs. We obtained two clusters namely, "Good Prognosis" (me-

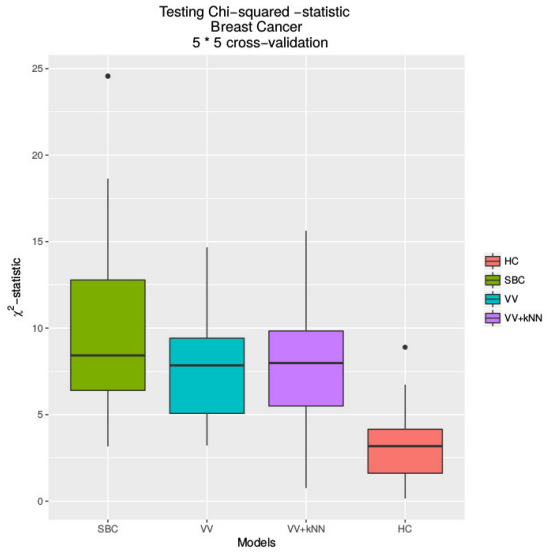

Figure 4.12: Cross-validation results for Breast Cancer. Log-rank statistic is based on the **predicted classes** from the SBC model on the test set.

dian time to distant metastasis 93 months) and "Bad Prognosis" (median time to distant metastasis 47 months). These two clusters yielded two well separated survival curves ($p = 1.7e-08$) on the training data set. We then used our method to predict class memberships and survival times of patients (see convergence diagnostic plot Fig. 4.14). On the testing set (the 50% of the samples not used for model training), this yielded two clusters which have significant differences in their survival curves (see Fig4.16). Further investigation of the two clusters obtained by our SBC method showed that the Bad prognosis group was significantly enriched ($p = 2.4e-15$, hypergeometric test) in the Estrogen Receptor negative (ER-) type. ER status has been long established as risk factor for metastatic breast cancer[PSDW84]. We also found significant enrichment ($p = 2.5e-05$, hypergeometric test) of the Good Prognosis cluster with the Luminal sub-type which has been reported to be associated with better prognosis[SPT+01]. Over-representation analysis of our SBC signature with respect to

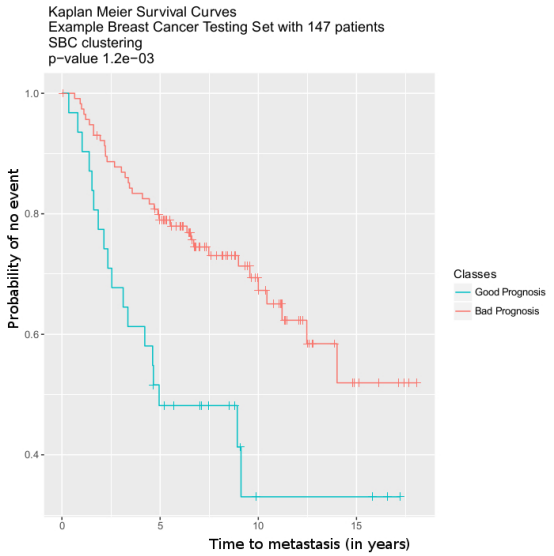

Figure 4.13: Results on the Breast Cancer test data set with the example training-testing split. Predicted classes from SBC. Crosses indicate censored outcomes. Clinical end point is time to metastasis.

Gene Ontology terms revealed the significant "Protein Methyltransferase Activity" (FDR <0.05). This process is indeed of known relevance for breast cancer[KCV+03].

A further ranking of the SBC genes w.r.t. their importance for clustering indicated a particular strong influence of E2F1 and TIMELESS. The gene E2F1 has been established to be related to breast cancer and is even prognostic for metastasis[HPB+03] while the circadian gene TIMELESS has been postulated as a risk factor for breast cancer tumorigenesis[FLZ+12]. Another important gene according to SBC was PGR (Progestrone Receptor), whose role in breast cancer has been long known[HM78]. Other noteworthy genes include Reticulon 3 (RTN3), which has been associated to cell apoptosis[LLK+09], and IGFBP5, which has been related to cell growth in breast cancer[SSC+92].

The results on the training data set are presented in Figure 4.15 where the molecular differences between the two SBC clusters are visually visible. The columns of the heat map are

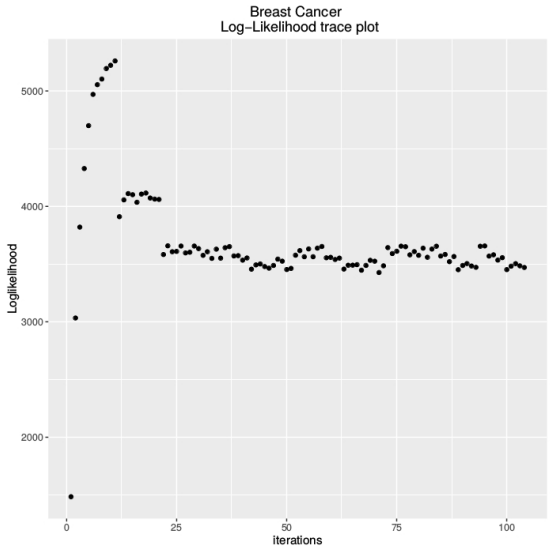

Figure 4.14: Log-likelihood trace plots for the Breast Cancer Data Set

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

Table 4.1: Breast Cancer Data Set Results on the example data-split

arranged according to the log-odds ratio of belonging to the two clusters. Enrichment results of our SBC derived "Good prognosis" and "Bad Prognosis" classes with respect to key factors in breast cancer progression are shown in Table4.2 and Table4.3. Gene Ontology (GO) enrichment analysis of the SBC signature was carried out via a conditional hypergeometric test (R-package GOstats[FG06]). Multiple-testing correction was applied using[BH95] method to control the False Discovery rate.

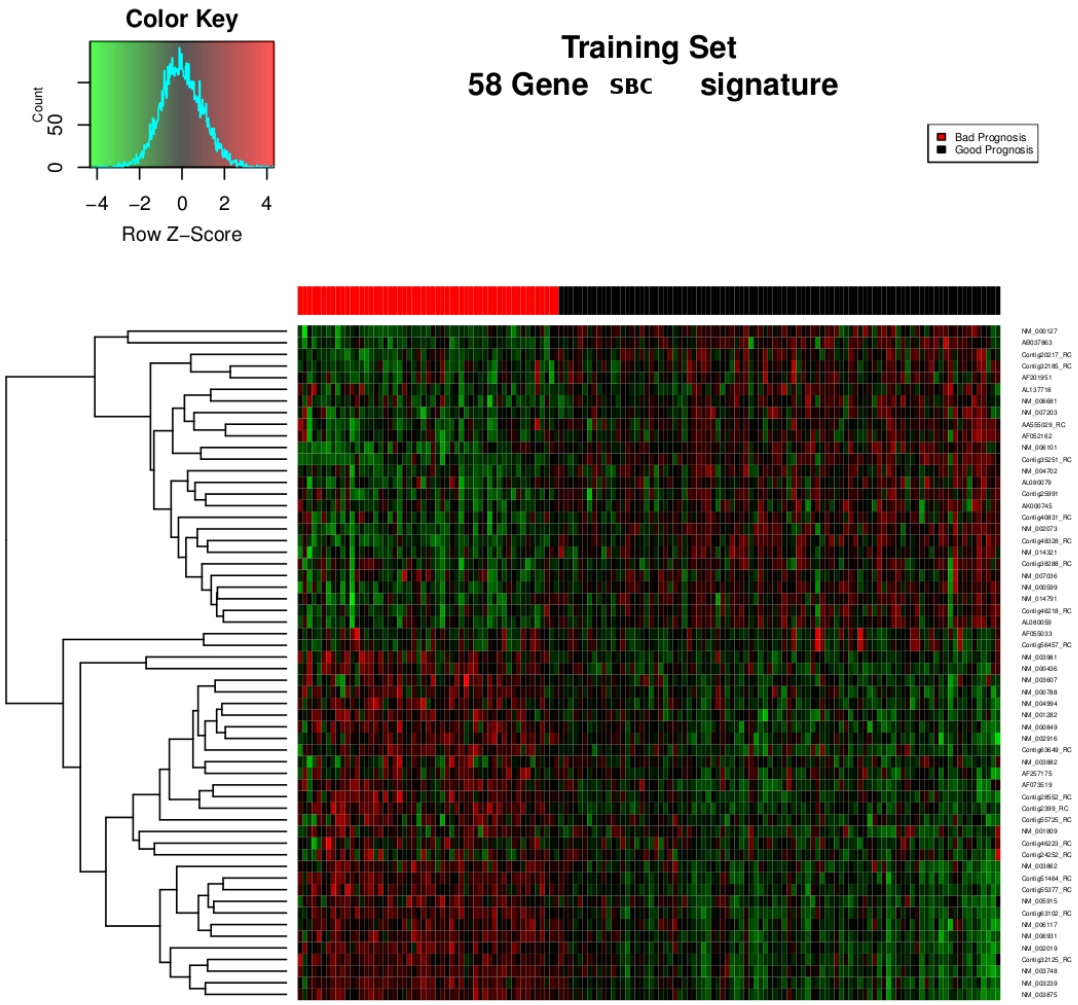

Figure 4.15: SBC on Breast Cancer training set

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

Table 4.2: Results on Breast Cancer Data set: Enrichment of SBC classes with ER status

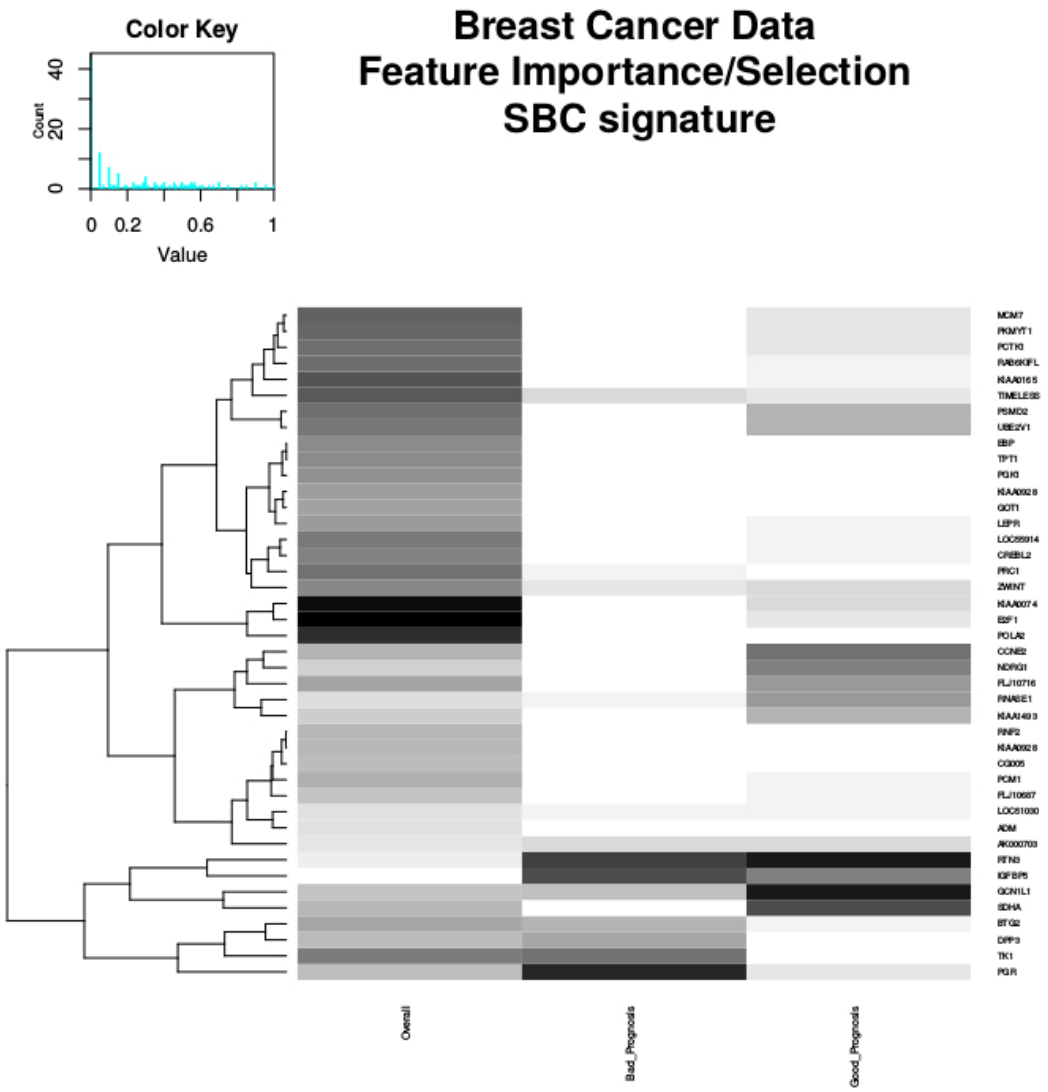

Figure 4.16: Feature Importance and Selection from SBC on the Breast Cancer data set. The leftmost column represents importance of feature on molecular data clustering, the two right columns represent strength of association to cluster specific survival times. Darker colours imply stronger effects.

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

Table 4.3: Results on Breast Cancer Data set:Association of SBC classes with breast cancer sub-types

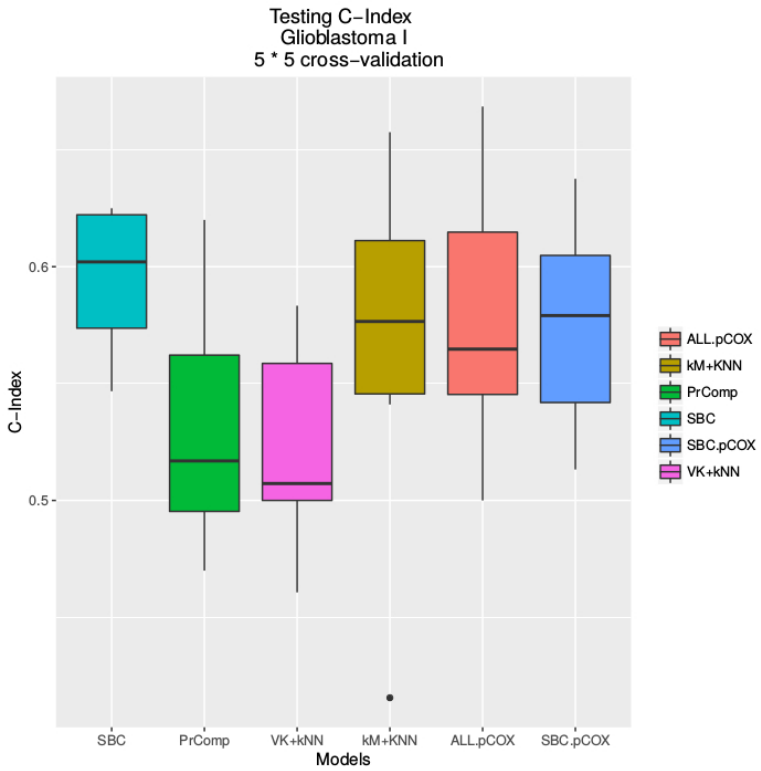

Figure 4.17: Results on the Glioblastoma I data set. Box plots depict cross-validated C-indices for different methods.

4.5.2 Glioblastoma I (Verhaak et al.)

We also applied our SBC model on the Glioblastoma Multiforme (GBM) microarray data from [VHP+10b]. The data were downloaded from https://tcga-data.nci.nih.gov/docs/publications/gbm_exp/.

We considered the "overall survival" as the clinical endpoint in our analysis. Overall, 196 patients (with survival information) were selected along with the original 840 gene [VHP+10b] signature which we used for comparison (henceforth known as the Verhaak signature). Using only the training data we filtered features based on their p-values from uni-variate Cox Regression models and chose the top genes as our SBC signature. For the Cross-validation

we used the same technique to arrive at the SBC model within each of the cross-validation loops. Our method was able to predict survival better than the following methods (see Fig.4.17):

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

In addition, Figs. 4.18 and 4.19 indicate a better separation of survival curves with SBC than achieved by original VK stratification, VK + kNN and kM+KNN.

For our example data-split we chose top 47 genes as the SBC signature and trained our SBC model. Using that we discovered four distinct clusters (see convergence diagnostic plot in Fig.4.21) with unequal numbers of patients (10, 5, 25, 58). These clusters showed molecular differences as well as significantly different survival curves also on the test set (Fig.4.20). We referred to the four clusters as "Good","Good Moderate","Bad Moderate" and "Worst" based on their respective mean survival times (830 days, 626 days, 380 days, 180 days). Looking at the patients in the "Best" prognosis cluster we find a high enrichment ($p=3.5e-05$, hypergeometric test) in the "Proneural" GBM sub-type defined by Verhaak et al. which has been reported in the literature to be linked with better survival[CGL+10]. As in the breast cancer data set, we again computed the feature importance of the SBC signature, one par-

ticular gene which has a higher contribution across all cluster comparisons (see Fig.4.22) is the "Programmed cell death 6" or PDCD6 gene. It has been known for its proapoptotic function and is thought to be involved in survival pathways in cancer[SXF+12]. Another interesting gene, which is assigned a high relevance by our method is TUSC4. TUSC4 has been established as a tumour suppressor gene regulating BRCA1 stability[PL14]. BRCA1 expression has been reported as a biomarker for GBM prognosis[VWC+15].

For the example data-split, we also report in Table 4.4 our results. We again use the log-likelihood trace plot to assess the convergence of our Gibbs sampling iterations as can be seen in Fig.4.21

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

Table 4.4: Glioblastoma I data set results for example data-split

There is a significant association between clusters discovered by our SBC and the ones reported by Verhaak et al., see Table 4.5 (p=$3.5e-05$, χ^2 test). We also note that the Best prognosis class exclusively contained samples from the Proneural Verhaak GBM class while the Good Moderate prognosis class was split between Classical and Mesenchymal sub-types.

To better understand our SBC signature we plot the feature importance of all the genes

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

Table 4.5: Results on Glioblastoma I: Association of SBC classes with GBM Verhaak sub-types

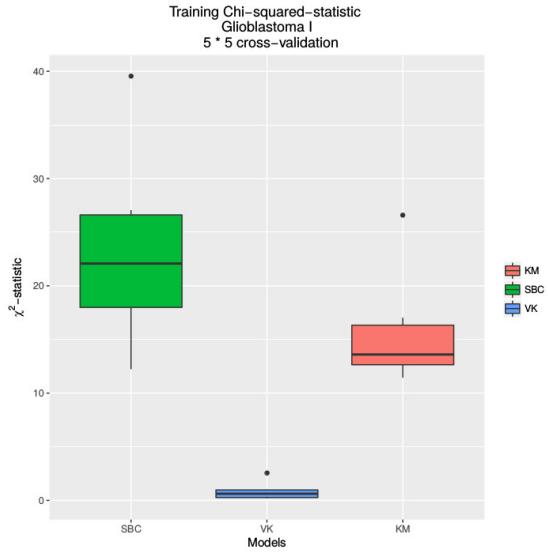

Figure 4.18: Cross-validation results for GBM I. Log-rank statistic is based on the **recovered classes** from the SBC model on the training set.

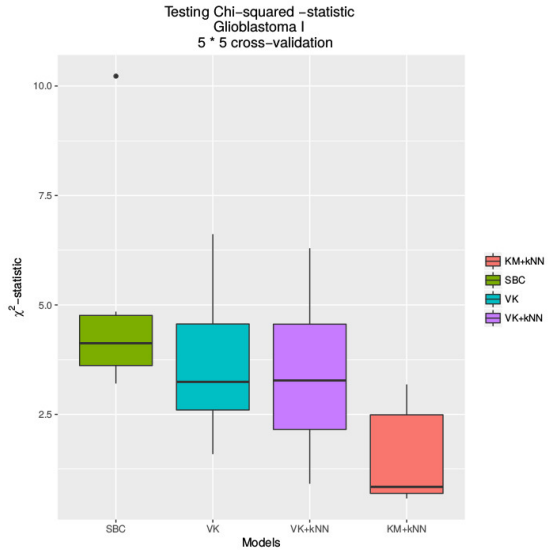

Figure 4.19: Cross-validation results for GBM I. Log-rank statistic is based on the **predicted classes** from the SBC model on the test set.

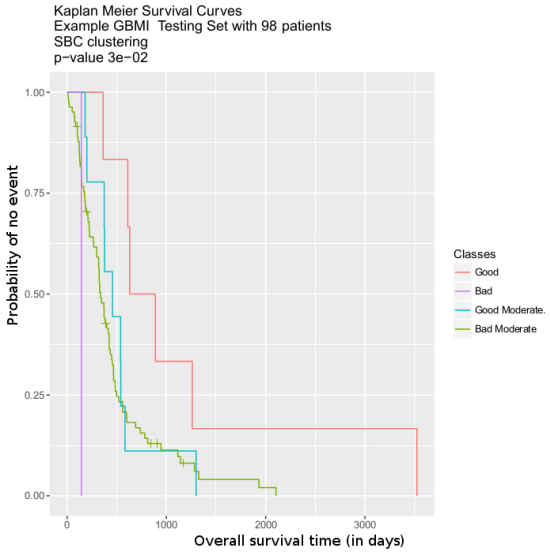

Figure 4.20: Results on Glioblastoma I test data set with example training-testing split. Predicted classes from SBC. Crosses indicate censored outcomes. Clinical end-point is overall survival.

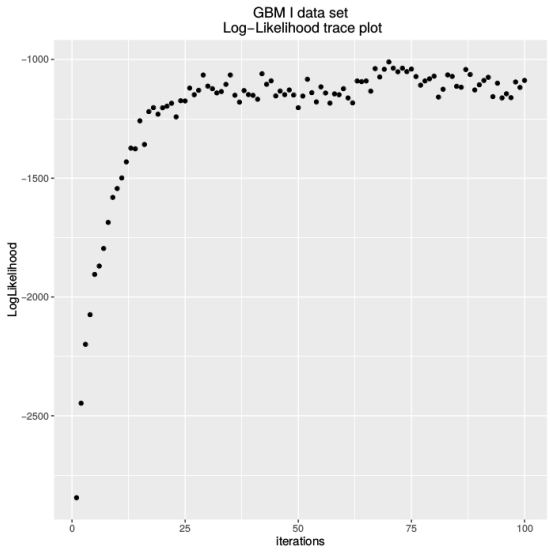

Figure 4.21: Log-likelihood trace plots for the Glioblastoma I Set

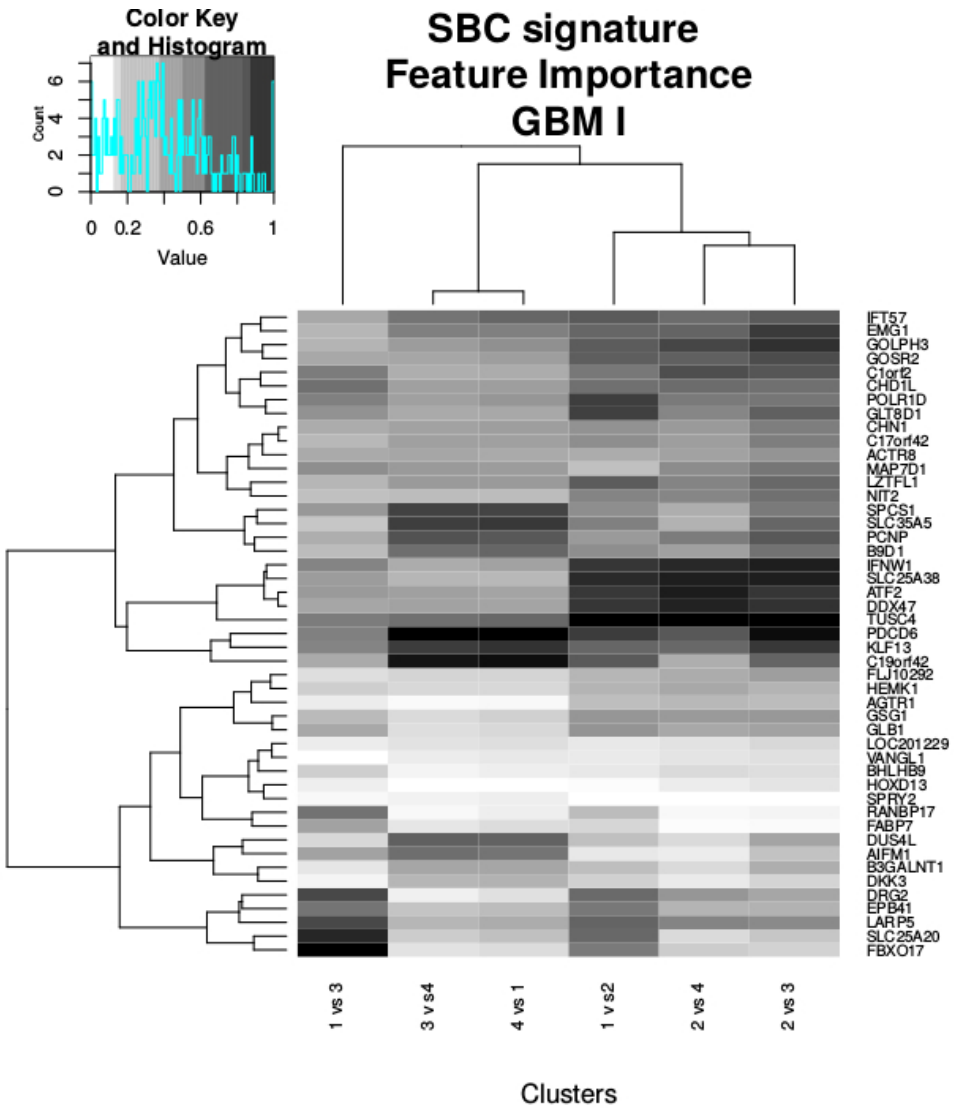

Figure 4.22: Results on Glioblastoma I (SBC):Feature importance of the SBC signature on the GBM-Verhaak data set in discriminating respective clusters

in our SBC signature to distinguish between different clusters in Fig. 4.22. Some genes which have higher contributions across all cluster comparisons (shown by a darker color in the heatmap in Fig. 4.22) were investigated to reveal interesting biological functions. Apart from the above mentioned genes, the gene SLC25A38 which is a member of the SLC25 gene family and also plays a prominent role as a SBC signature gene has been reported to suppress cell growth in human gliomas[WFS+11].

4.5.3 Glioblastoma II (TCGA-GBM)

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

- A combination of k-nearest neighbour cluster assignment followed by a cluster specific penalized Cox regression (abbreviated as KMkN) using the SBC signature with the concatenated matrix of mRNA and miRNA expression profiles for each patient. When CCA features are used the method is referred to as C.KMkN.

- The PrComp method, as defined above but this time applied to the concatenated data matrix of gene and miRNA expression profiles.

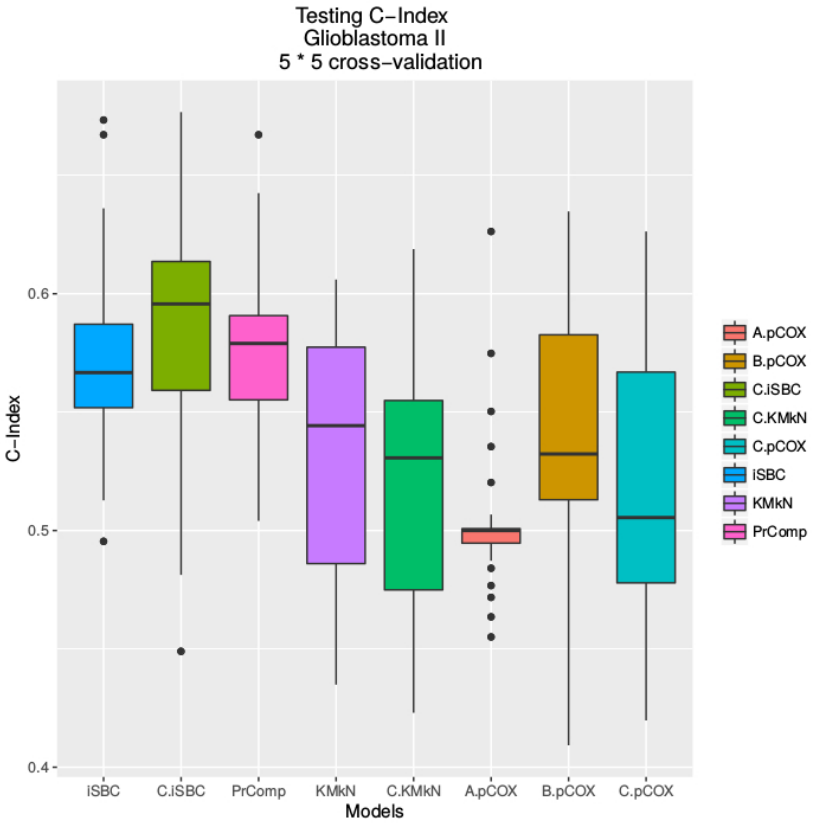

Figure 4.23: Results on the Glioblastoma II data set. Boxplots depict cross-validated C-indices for different methods.

- Single (disregarding clustering) Penalized Cox regression applied on the concatenated matrix with all the features, referred to as A.pCOX. When SBC features are used, it is referred to as B.pCOX. While when CCA features are used we refer to it as C.pCOX.

Our results (Fig.4.28) indicate at least as good prediction performance with our iSBC and C.iSBC methods than with competing ones (PrComp). At the same time Figs. 4.24 and 4.25 show that our methods separated survival curves better (after predicting cluster membership of test patients) than a k-means clustering approach or k-means plus kNN cluster membership predictions.

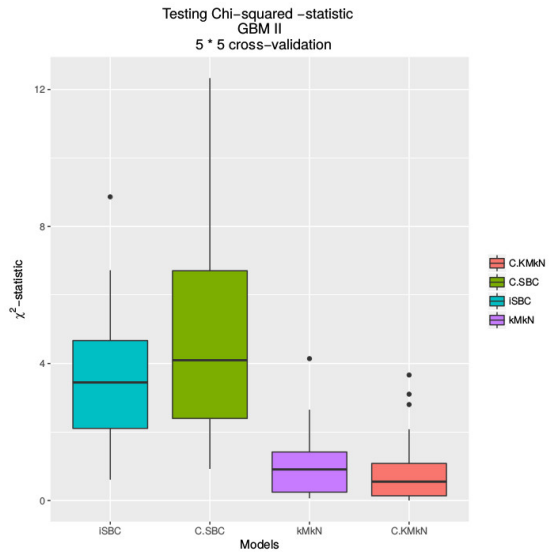

Figure 4.24: Cross-validation results for GBM II.Log-rank statistic is based on the **predicted classes** from the iSBC model on the test set

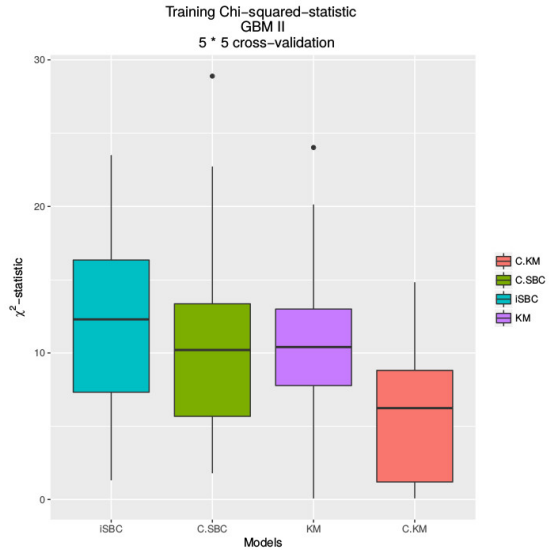

Figure 4.25: Cross-validation results for GBM II. Log-rank statistic is based on the **recovered classes** from the iSBC model on the training set

Delving deeper in the example data-split we selected 31 top ranking mRNAs and the top 31 miRNA probes as our iSBC signature. We then applied our iSBC method once with and once without projecting data on the top 10 canonical covariates. The CCA pre-processing leads to slight increase in the survival prediction (Fig. 4.28). In the following part we focus our discussion on the solution obtained without CCA preprocessing. Application of our iSBC approach lead to the discovery of 4 clusters (we call them, as before, "Worst","Good Moderate","Bad Moderate","Best" based on the prognosis) of unequal number of patients (2, 27, 54, 13). The clusters from our iSBC still result in clearly separable survival curves on both training and test data sets (see Fig.4.26). We further investigated cluster-specific enrichment with respect to somatic mutations. The mutation pattern found in genes included in our model is significantly related to the iSBC derived clusters ($p = 1e-05$, χ^2-test). An interesting observation was the mutual exclusive mutation pattern of TP53 and PTEN genes among the iSBC clusters, meaning that if TP53 was found mutated in one iSBC cluster, PTEN was never mutated in that cluster and vice-versa. This mutual exclusivity has also been reported in literature[KGM+02]. Over-representation analysis of the iSBC signature revealed the significant Gene Ontology term "negative regulation of G1/S transition of mitotic cell cycle" (FDR <0.05). This is highly interesting because cancer cells have an over-active cell cycle, leading to proliferation and hinting at possible mechanism for cancer progression. Looking at the most discriminatory features from our iSBC model (see Figs.4.28,4.29), we find that one important mRNA iSBC feature is "developmentally regulated GTP-binding protein 2" or DRG2 gene which has been shown to induce apoptosis in cancer cells[Jyx+12]. Another interesting and discriminatory gene is β-catenin (CTNNB1), which is a key protein in the Wnt signaling pathway. Deregulation of the Wnt pathway has

been associated with various cancers, including GBM[LLA+16]. Another discriminatory gene identified by iSBC is ADAM22, which has been shown to be under-expressed in high-grade gliomas[GNP+06]. An important miRNA feature miR-661 is known to activate the p53 pathway and suppresses tumour progression[HBPO14]. Furthermore we found miR-675, which has been linked to Gliomas[SWL+14] while miR-637 has been shown to inhibit tumorigenesis in various cancer types[ZHF+11] and is discussed as a prognostic marker in gliomas[QSL+15]. For the example data-split, we report in Table 4.6 our results. We again use the log-likelihood trace plot to assess the convergence of our Gibbs sampling iterations as can be seen in Fig.4.27. Our iSBC resulted in significantly different survival curves also on the test set (Fig.4.26) We conducted a cluster-specific somatic mutation enrichment analysis. For this purpose we looked for genes (which are part of our iSBC signature) which show cluster-specific somatic mutations. Somatic mutation data was only available for 23 patients out of 96 training patients. We obtained the mRNA signature from the SBC model, moreover miRNAs were also mapped to their gene targets using the 'multiMiR' package in R[RKT+14]. 57 unique genes were identified in this manner. 55 out of the 57 genes show the same pattern illustrated in Table4.7 where all of them show mutation exclusively in the best prognosis cluster of SBC. The interesting case is that of TP53 and PTEN genes which show a mutual exclusive behavior of somatic mutation as shown in Table 4.8 and Table 4.9.

In a similar manner as in GBM-Verhaak data set we plotted the feature importance of the mRNA and miRNA SBC signature in Fig.4.28 and Fig.4.29. We also explored features which had more contributions with respect to others (shown by darker colours in Fig.4.28 and Fig.4.29).

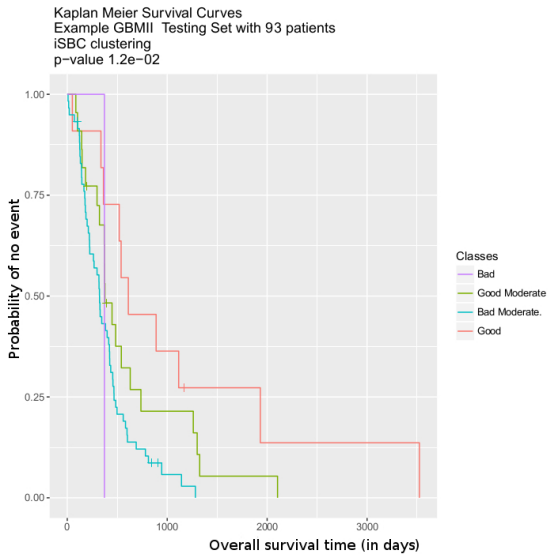

Figure 4.26: Results on Glioblastoma II data set with example training-testing split. Predicted classes from iSBC on the test set. Crosses indicate censored outcomes. Clinical end-point is overall survival.

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

Table 4.6: TCGA-GBM data set results for example data-split

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

Table 4.7: Results on Glioblastoma II (iSBC): Number of somatic mutations across iSBC defined clusters for signature genes except TP53 and PTEN

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

Table 4.8: Results on Glioblastoma II (iSBC):Number of somatic mutations across SBC defined clusters for TP53

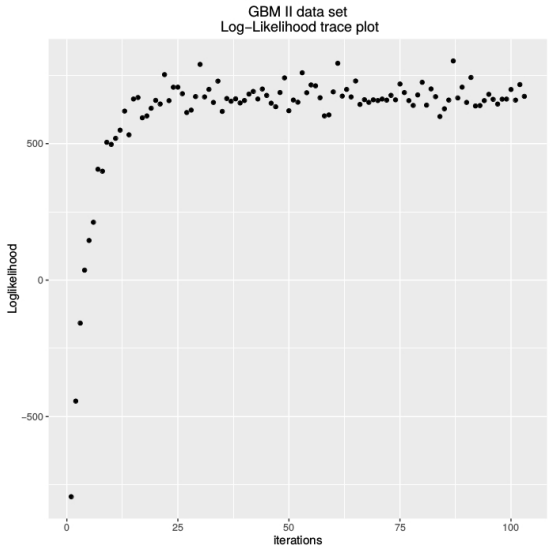

Figure 4.27: Log-likelihood trace plots for the Glioblastoma II Set

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

Table 4.9: Results on Glioblastoma II (iSBC): Number of somatic mutations across SBC defined clusters for PTEN

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

Table 4.10: Actual running times for SBC/iSBC on Real Data Sets

4.6 Running Times for SBC and iSBC

We present the actual running times of the SBC on the three data sets. We used 11.7 GB Intel 64 bit Xeon (R) 4x2.66 Ghz processor in Table 3.10.

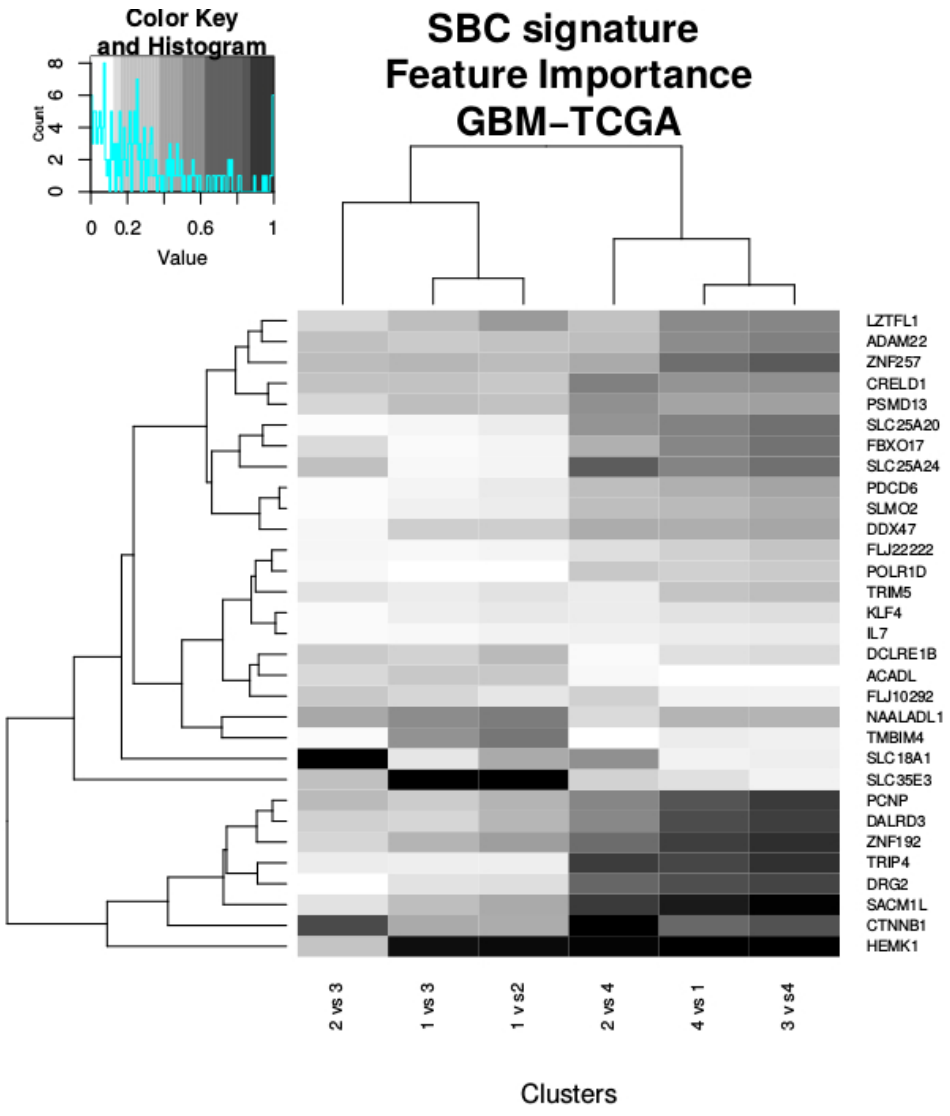

Figure 4.28: Results on Glioblastoma II (iSBC): Feature importance of the SBC signature on TCGA-GBM gene expression in discriminating respective clusters

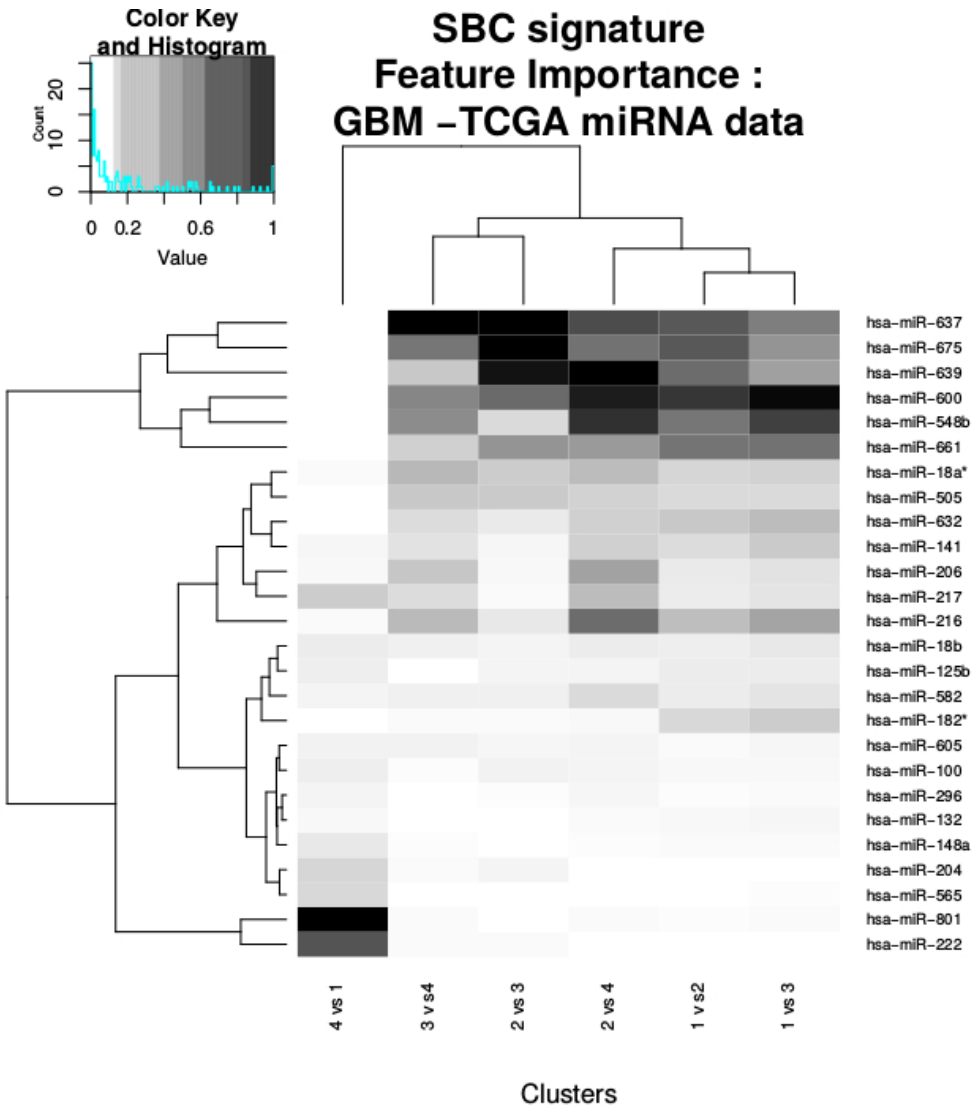

Figure 4.29: Results on Glioblastoma II (iSBC):Feature importance of the SBC signature on TCGA-GBM mi-RNA expression in discriminating respective clusters

4.7 Effect of Survival Data on SBC

4.7.1 Hierarchical Dirichlet Process Mixture Model

SBC distinguishes itself from traditional clustering algorithms for patient level microarray data by including clinical end-points as a very important source of information. In this section we explored the scenario when we ignore the clinical end-point information and use a very similar Hierarchical Dirichlet Process Mixture Model (hDPMM) (as shown graphically in Fig.4.30) to cluster the patients based on just their molecular profiles (gene expression). The difference between SBC and hDPMM is the absence of parameters for modelling the survival information in hDPMM. We can also make predictions from hDPMM in a very similar manner as described for SBC.

For the sake of comparability, we used the corresponding SBC signatures as the feature set for the Breast Cancer data set and Glioblastoma I data set. A survival model (penalized Cox PH) is then fitted on top the clustering obtained above to stratify the survival curves and to make predictions.

4.7.2 Breast Cancer Data Set

The hDPMM model on the breast cancer data set yields no clusters. The results are presented in Table 4.11 As can be seen, stratifying patients according to the SBC yields much better predictions for survival than using hDPMM. Thus we can conclude that survival information plays a vital role in obtaining the ”Good prognosis” and ”Bad prognosis” clusters which were obtained from our original SBC model.

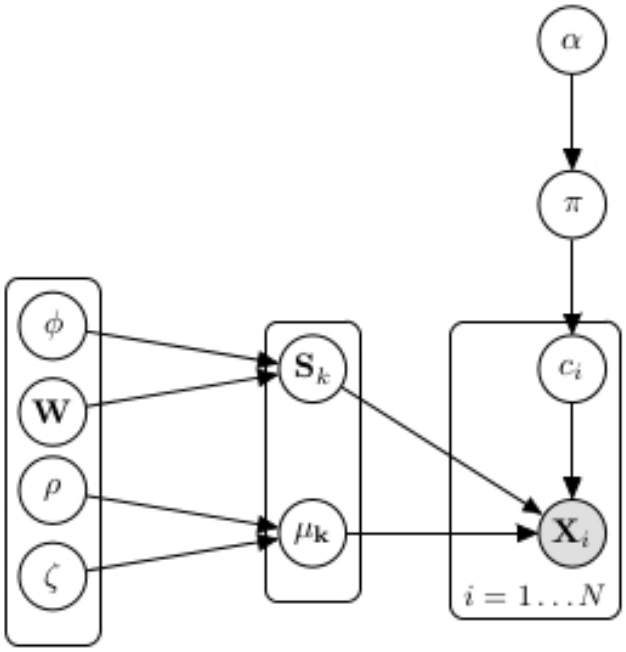

Figure 4.30: Graphical Model representation for hDPMM

4.7.3 Glioblastoma I Data Set

The application of the hDPMM model on the Gliobalstoma I data set yields 3 clusters, one which contains the majority of the data points. The three clusters contain 94,3 and 1 data points respectively for the training data set. For the prediction, the hDPMM places all the test points in one cluster. The results comparing hDPMM with SBC are presented in Table 4.12. Again, we can make a strong case that stratifying patients according to the SBC yields much better predictions for survival than using hDPMM and that the original 4 clusters obtained from the SBC model are heavily influenced by the clinical end-points.

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

Table 4.11: Breast Cancer Data Set Results with hDPMM

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

Table 4.12: Glioblastoma I Data Set Results with hDPMM

4.8 Effect of CCA pre-processing on iSBC

4.8.1 Glioblastoma II Data Set

Here we report the results for CCA pre-processed iSBC on the example training-testing split for GBMII data. We used a non-penalized CCA approach to pre-process the Glioblastoma II Data set with the initial set of 31 mRNA and 31 miRNAs. We chose the top 10 Canonical Correlates and obtained the corresponding transformed data matrices for mRNA and miRNA each now containing 10 features each. The results on this transformed Glioblastoma II data set are shown in Table 4.13 which show marginal benefits on the prediction performance of the integrative SBC model using features derived from applying CCA to our SBC signature data sets. We still get four clusters as before with 57, 18, 15 and 6 data points respectively. Furthermore, we obtain feature importance for our new set of features (as shown in Fig. 4.33 and Fig.4.34) and plot the factor loading matrices in Fig.4.31

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

Table 4.13: Glioblastoma II Data Set Results with new feature sets derived from CCA

and Fig.4.32. The corresponding correlation values for the features are also shown.

4.9 Conclusion

In this Chapter we have introduced a novel fully Bayesian clustering algorithm (SBC) which takes in clinical end-points of patients along with heterogeneous -omics data to perform two tasks in one - a) patient sub-group identification on training data and b) prediction of patient sub-group and survival time on testing data. Our method was based on the motivation of discovering clusters of patients using their distinct molecular signatures and strong survival curve separability. Another important motivation was the predictive utility of our approach along with biological interpretability. We have shown with simulations and real data that our method outperforms ad-hoc algorithms like k-means followed by fitting cluster-specific survival models. Furthermore, our SBC yields clearly better results than a hierarchical Gaussian DPMM without survival information, indicating the relevance of the clinical outcome in our model. We believe the ability of SBC to identify patient-subgroups differing in survival constitutes an advantage compared to existing approaches like[VVDVDV+02, VHP+10b]). Furthermore, SBC is principally able to take into account more than -omics data source. Our assumed cluster specific factorization of the complete likelihood essentially weighs features inversely to their noise level. The CCA preprocessing approach

explored here is a refinement in that context, which could potentially also allow for combining discrete with continuous data types, as e.g. shown in [WTH09]. In future research we want to explore this aspect further and see, how CCA or similar latent factor approaches could be integrated better into our SBC method.

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

One of they key challenges for any clustering algorithm for biological data is to explore the biological underpinnings of the obtained clusters. In this regard we have found that certain sub-types from our model are particularly enriched in certain biological markers (for example ER status for breast cancer) and also correlate strongly with some sub-types in the well established classification schemes for example of [SPT+01], [VHP+10b] . Altogether we see SBC as a step towards a more clinically relevant dissection of patient heterogeneity.

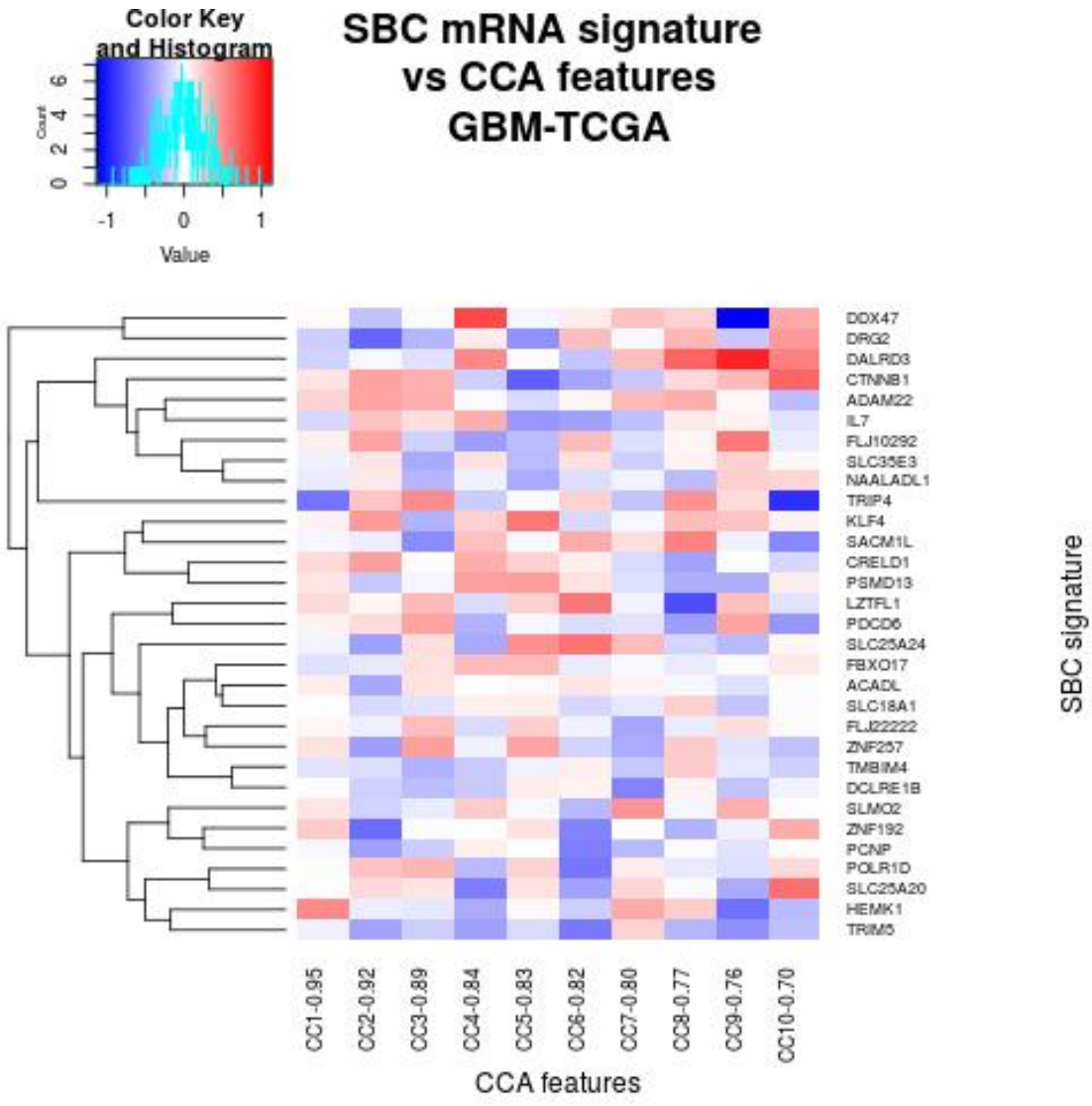

Figure 4.31: Factor Loading Matrix between CCA features and the original SBC mRNA signature. Canonical covariates are named as CC1-xx to CC10-xx, where 'xx' indicates the respective canonical correlation

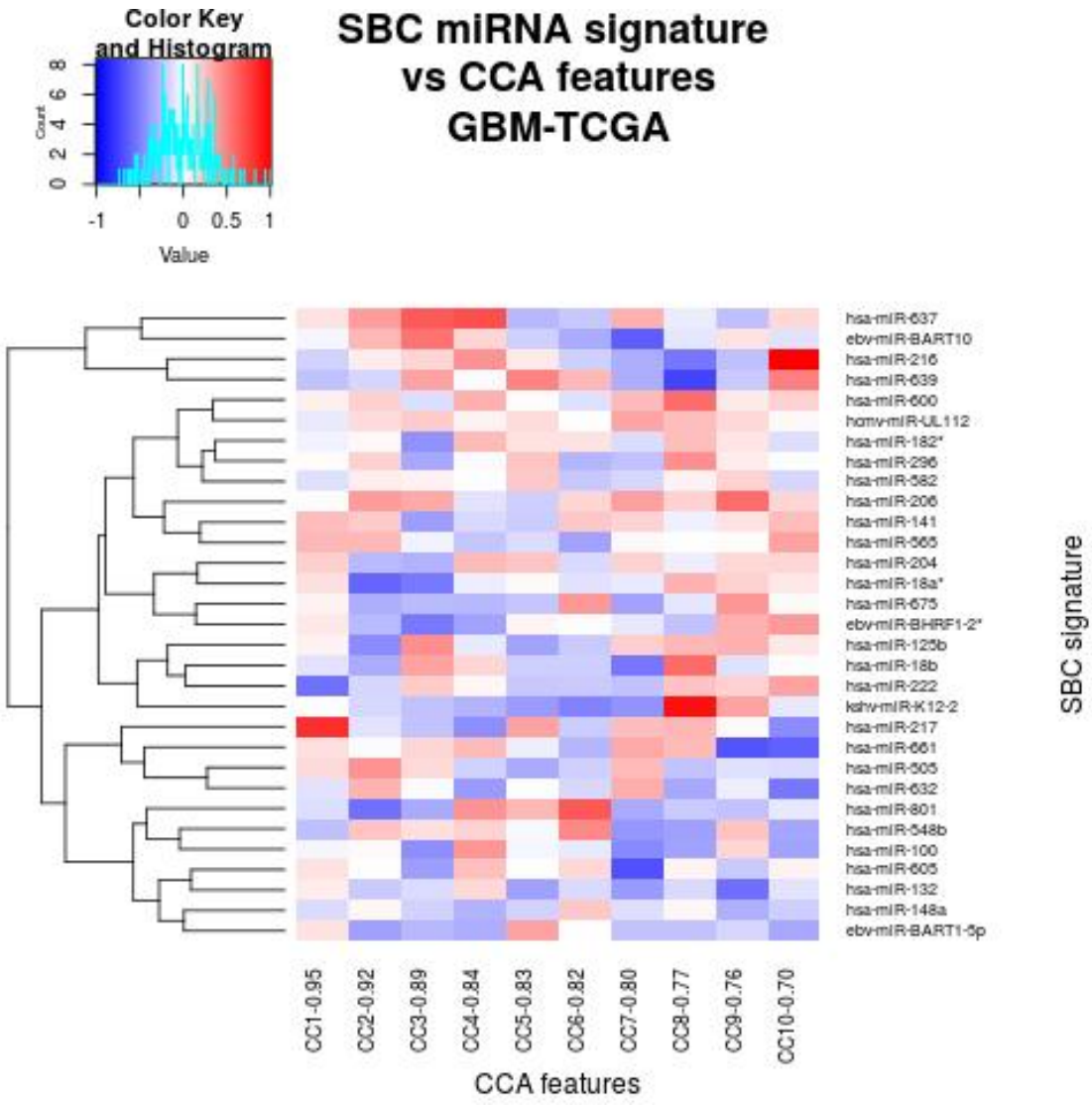

Figure 4.32: Factor Loading Matrix between CCA features and the original SBC miRNA signature. Canonical covariates are named as CC1-xx to CC10-xx, where 'xx' indicates the respective canonical correlation.

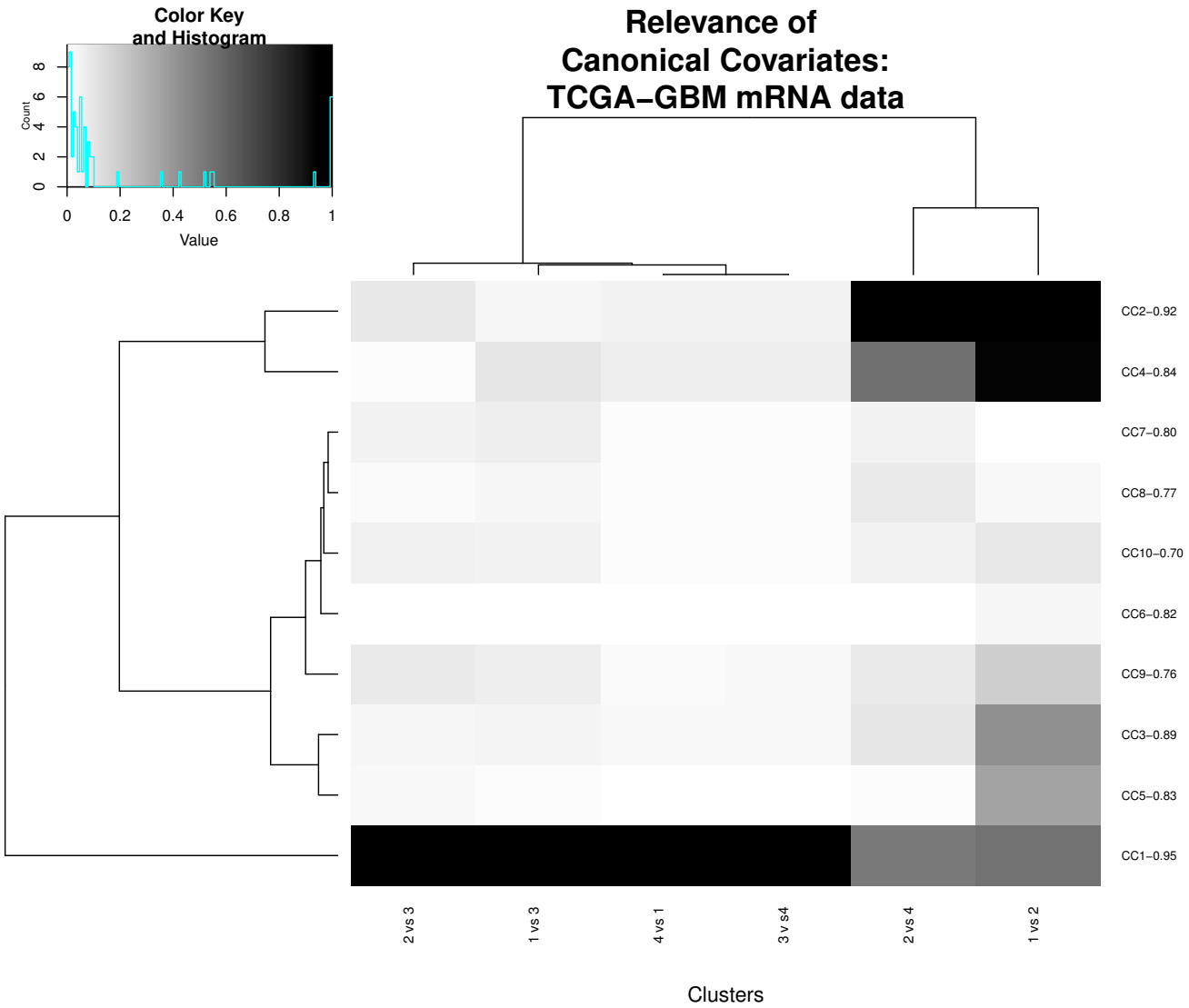

Figure 4.33: Feature Importance of the new CCA features derived from the mRNA-SBC signature.Canonical covariates are named as CC1-xx to CC10-xx, where 'xx' indicates the respective canonical correlation

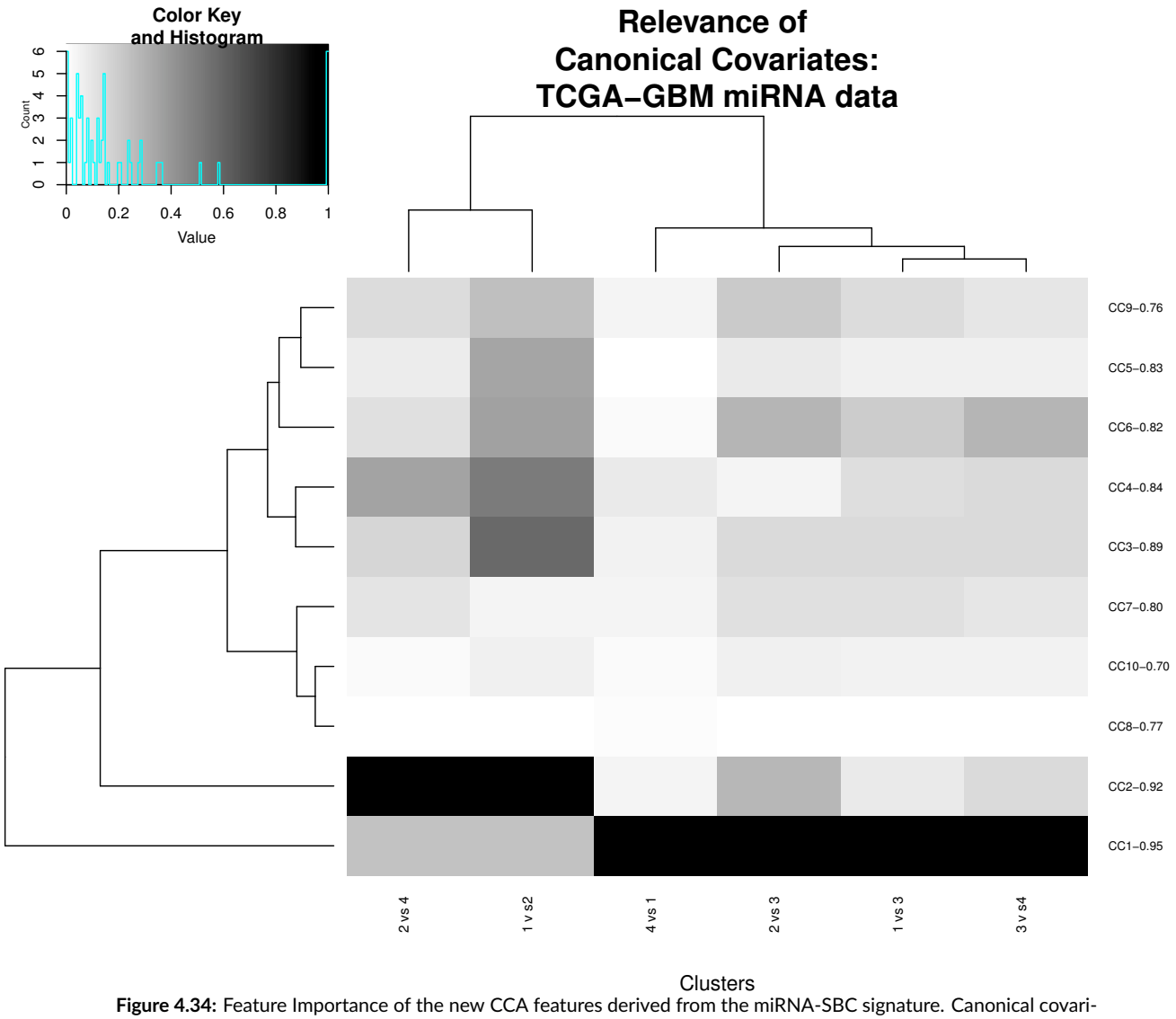

Figure 4.34: Feature Importance of the new CCA features derived from the miRNA-SBC signature. Canonical covariates are named as CC1-xx to CC10-xx, where 'xx' indicates the respective canonical correlation

"The mind is not only capable of knowing [innate ideas], but further of finding them in itself; and if it had only the simple capacity to receive knowledge it would not be the source of necessary truth"

Gottfried Wilhelm Leibniz

5

Applications of SBC to IDENTIREST data

As explained in Chapter 4, Survival based Bayesian Clustering (SBC) potentially provides a clinically relevant method for dissecting unknown heterogeneity at sample and patient level. One of the goals in the IDENTIREST project was also to better understand the heterogeneity of the samples which could be clinically meaningful. For this purpose, we apply our SBC method on the cohort of samples from the IDENTIREST project. In the first part

of this chapter, we use the SBC model trained on the Verhaak cohort to make predictions on IDENTIREST Central Samples. In the second part we use a modified version of SBC (hDPMM) to explore the heterogeneity in the entire cohort of IDENTIREST samples (Peripheral and Central samples) in an unbiased way.

5.1 Predictions of SBC on Central Samples

In Chapter 4, we had trained our SBC model on Glioblastoma patients from the Verhaak cohort [VHP+10a] with the goal to stratify GBM patients with respect to their molecular profiles as well as their overall survival times. To recapitulate: Using the cohort of 98 patients, we had trained our SBC model based on the 47 gene signature to obtain 4 different patient strata (we referred the strata as following: Cluster1 or Good Prognosis, Cluster 2 or Good Moderate prognosis, Cluster 3 as Bad Moderate prognosis and Cluster 4 as Bad Prognosis) (see Section 4.5.2). Also, in the same section we used 98 other samples from the Verhaak cohort as validation set and obtained SBC predictions. The predictions of the SBC led to 4 predicted strata which were significantly separated in their Overall Survival (OS) curves.

We now use this trained SBC model on the Verhaak cohort to make predictions on the IDENTIREST cohort of patients. For this purpose, we use only those IDENTIREST patients which have unique Central Sample biopsy along with overall survival information. The rationale behind the choice of central samples is as follows: As has been repeatedly shown in Chapter 3 (Sections 2 & 3), Central samples are molecularly distinct from the Peripheral samples and can be considered similar to GBM cancer samples. This hypothesis is further strengthened in Chapter 3 (Section 4) where Central Samples were satisfactorily classified in the four molecular subtypes defined by Verhaak et al. while the Periphery sam-

ples tend to fall predominantly into one Mesenchymal class. The choice of patients which have unique Central Samples is made to avoid the problem (as mentioned in the last section) of aggregating multiple samples for the same patient (which is necessary for the application of SBC). This leaves us with 37 Central samples.

Out of the 47 gene SBC signature, expression values for 7 genes were not available in the IDENTIREST samples. Hence, we used the R-package `'impute'` [HTN+11] to perform imputation of the gene expression values based on its k-Nearest Neighbors. Next, we corrected for the batch effect between the 98 sample Verhaak data set and 37 sample IDENTIREST cohort. We achieve this batch effect correction by simply adding the difference of means between feature values in both data sets, this means that the IDENTIREST validation dataset is translated to the mean of the Verhaak training set. This is the same technique that we adopted in Chapter 3 (Section 5) and its graphical depiction can be seen in Fig. 5.1. After imputation and batch correction we are ready to make predictions on the IDENTIREST samples (based on the gene expression and overall survival data).

The SBC predicts 3 different strata for the 37 Central samples, with 25 patients belonging to Cluster 4 (or Good Moderate prognosis), 7 samples belonging to Cluster 3 (or Bad Moderate Prognosis) while 5 samples belong to Cluster 2 (or Bad Prognosis group). The OS Kaplan Meier curves are significantly separated (p-value of $4e-03$). The PCA plot along with predicted classes for the 37 Central IDENTIREST samples is shown in Fig.5.2, along with the corresponding Kaplan Meier curves. The predicted C-Index on the IDENTIREST cohort was 0.56 (similar to that obtained on the Verhaak prediction cohort, see Chapter 4).

In order to further validate the prediction clusters on the IDENTIREST cohort, we use

the CNV (genomics) data for the 37 samples to check for significant differences. One important characteristic of each of the genomic samples is whether it contains GBM-specific genomic alterations or not. These GBM-specific alterations can be of many types, a detailed description is provided in Appendix C. Thus the samples were divided into a) those having typical GBM mutations and b) those not having typical GBM mutation and c) others. Many samples within the IDENTIREST cohort exhibit typical GBM genomic mutations and other do not. We looked whether certain predicted SBC clusters were enriched in typical GBM mutations containing samples. Results can be seen in Table 5.1. There is significant association between the predicted SBC classes and Typical GBM mutation status (p-value $1e-03$, χ^2 test). All samples in the predicted Bad prognosis group contained Typical GBM mutations. We further looked into the CNV data for the 1070 genes (which were selected in Chapter 3, Section3). We ask the question whether the CNV data is associated with predicted SBC class labels. We fit gene-specific generalized linear models to the predicted SBC cluster label with gender, age as additional covariates apart from CNV data. Out of the 1070 genes, 43 genes have significant (FDR 0.01) association with predicted SBC class labels. This association can also be seen in the CNV heatmap of the 43 genes for the 37 IDENTIREST samples in Fig. 5.3.

We further investigated the relationship between the gene expression of the 47 SBC signature and the 43 CNV genes which showed significant association to the predicted SBC cluster. There is no overlap between these two sets of genes, so we investigated the 43 CNV genes for their expression. Out of the 43 CNV genes, 18 of them showed a strong association (FDR 0.01) between their gene expression values and copy number changes. This association was established, as before, by fitting gene-specific linear models to the outcome vari-

able of gene expression with CNV, age, gender being the covariates. This subset of 18 genes thus shows consistent patterns of gene expression and CNV with respect to the predicted SBC clusters as shown in Fig. 5.4. After establishing consistent information in CNV and gene expression data in the 18 aforementioned genes, we checked for correlation of these 18 CNV genes and the 47 SBC gene signature. This correlation has been plotted in Fig.5.5. 65% of the correlations shown are significant (FDR 0.05), these significant correlations represent trans-effects of CNVs . Thus we can make a strong point that there is a strong connection 47 SBC gene signature and the 43 CNV genes which were used to validate the clustering.

Thus, we have been able to validate the results of the SBC model trained on the Verhaak cohort of 98 samples on the independent validation set of our IDENTIREST cohort. We not only get significantly different predicted cluster-specific OS curves, but also these clusters are interpretable using CNV data. To summarize, we have used both the gene expression as well as CNV data for the IDENTIREST cohort of Central samples to validate the original SBC model trained on the Verhaak cohort.

	With Typical GBM mutations	Without Typical GBM mutations
Cluster 3 Bad Moderate Prognosis	0	7
Cluster 2 Bad Prognosis	5	0
Cluster 4 Good Moderate Prognosis	12	13

Table 5.1: Results on CNV data for SBC predicted classes on IDENTIREST Central Samples

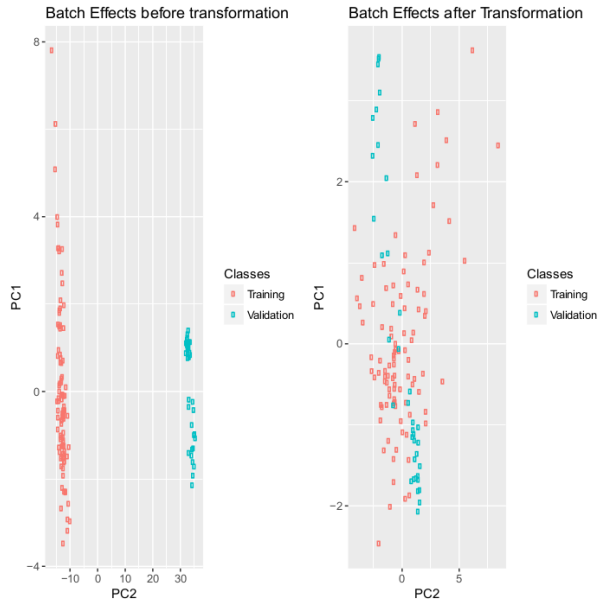

Figure 5.1: The batch effect is clearly observed on the left between the training and validation data sets. On the right is the PCA plot after batch correction

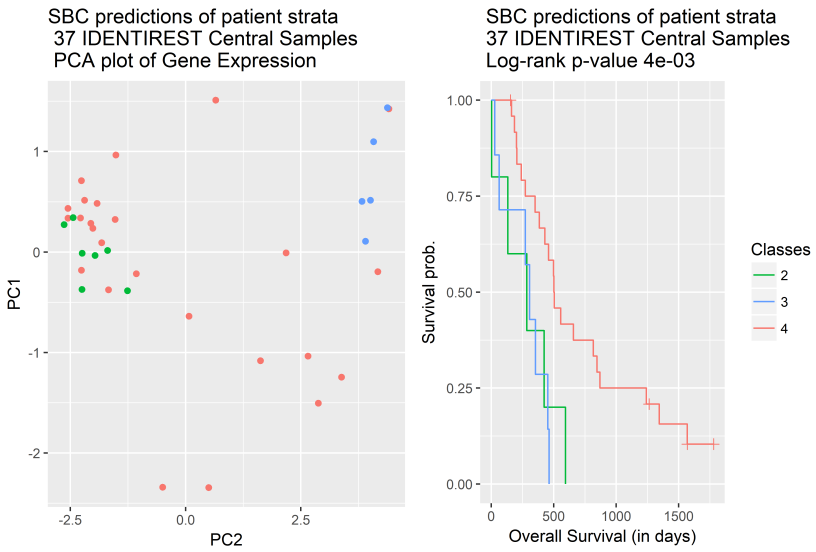

Figure 5.2: Results of prediction of SBC model on Central Cells. The left figure shows a PCA of the gene expression data of the Central Cells with three predicted classes. The right figure shows the different KM curves with the log-rank p-value of the predicted strata

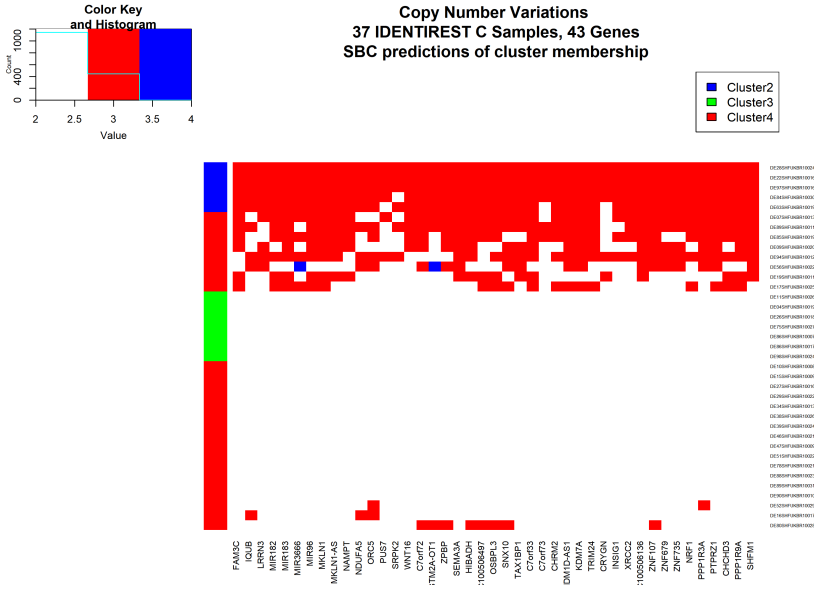

Figure 5.3: Results of interpretation of predicted SBC clusters using CNV data. The heatmap shows the CNVs of top 43 associated genes. The central IDENTIREST samples are arranged according to hierarchical clustering. Labels on the left are SBC predicted clusters.

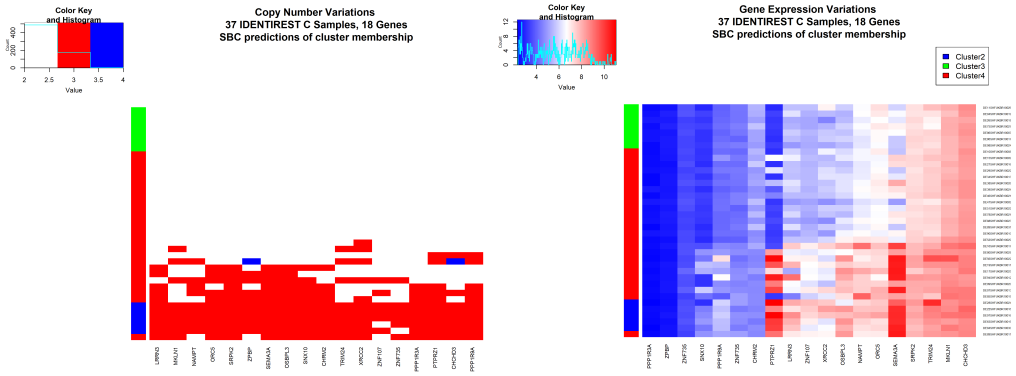

Figure 5.4: Association between CNV data and gene expression profiles. Both left and right figures have samples arranged according to hierarchical clustering in the same order. Left figure shows the CNV changes while the right figure shows the gene expression.

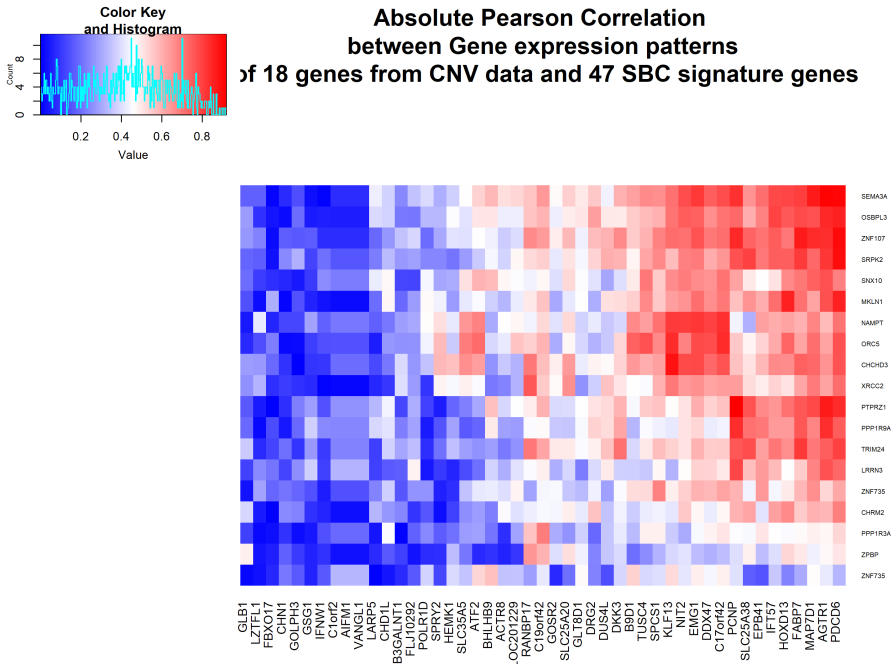

Figure 5.5: Correlation between gene expression values of the 18 CNV genes (on the rows) and 47 SBC signature genes.

5.2 Modelling sample heterogeneity using h-DPMM

In this section we make an attempt at exploring the Periphery and Central IDENTIREST samples heterogeneity in a more unbiased way. In Chapter 3 (Section 3.2) we have presented the heterogeneity present within the transcriptomics data and later on explored the Periphery samples in greater detail (see Section 3.5). Here we use a variation of SBC algorithm, h-DPMM to explore this heterogeneity (h-DPMM is described in the Chapter 4 and reproduced in Fig.5.6). As SBC normally works in the case of one sample per patient (with a corresponding unique clinical end-point) multiple samples from the same patient pose a challenge to SBC model. We believe that hDPMM model retains some of the strengths of the SBC model and is a potent tool to explore heterogeneity in the gene expression pro-

files of the samples. As the h-DPMM does not use survival data, we avoid the problem of modelling samples with different molecular data yet having the same survival time as they belong to the same patient.

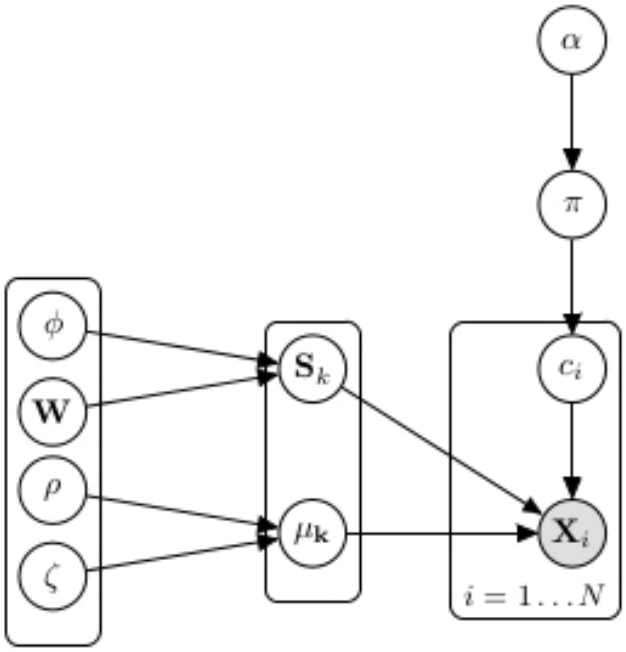

Figure 5.6: Graphical Model representation for hDPMM

The microarray data chosen contained ALL the 260 IDENTIREST samples along with ALL 27,148 microarray Features which could be annotated. An unbiased filter based approach was taken to come up with a set of signature genes (50). These were genes top 50 genes selected on the basis of their moderated F-statistics from R-package `limma` . The F-statistic measures the variations of the genes without any specific contrast. Our model hDPMM discovers 3 distinct clusters (Average Silhouette Index 0.23) in these 260 patients based on their expression of 50 gene signature. As one can qualitatively see in Fig.5.7, the clusters obtained from hDPMM are associated to the P and C cell types. This analysis

shows that results from our model confirm the initial hypothesis about the differences between the Peripheral and Central cell types. We note that this heterogeneity was obtained without taking the survival times (or PFS) of the patients into account, also the the list of 50 features was obtained without any reference to any contrast.

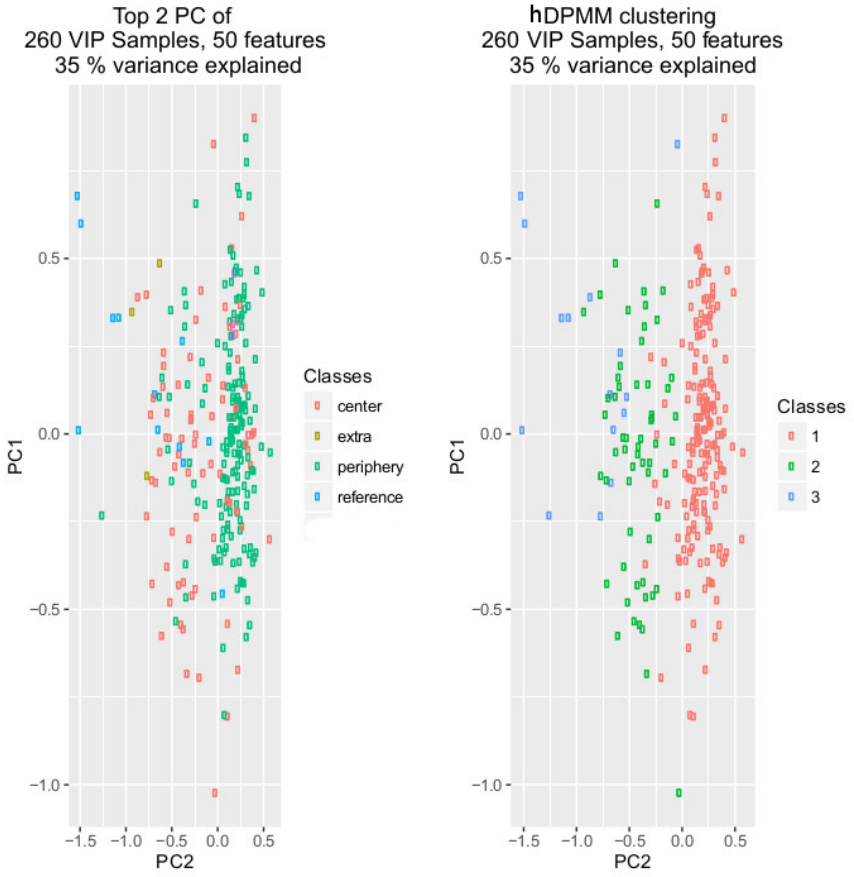

Figure 5.7: Results of hDPMM on 220 IDENTIREST samples presented in terms of PCA plots. On the left the labels come from classification of th cell according to the surgeon . On the right the same PCA has labels according to clusters obtained from the hDPMM. There are 3 hDPMM clusters

We can see that the heterogeneous clusters obtained on the IDENTIREST samples from hDPMM have a strong relationship to their corresponding surgeon annotation (Periphery or Central). We now show that using the genomics data from the IDENTIREST patients,

we can further validate the hDPMM clustering: We have genomics data available for 178 of the 260 IDENTIREST samples. All these 178 samples fall in two DPMM cluster (i.e. DPMM cluster I with 151 samples and DPMM cluster II with 27 samples). DPMM cluster III samples have no corresponding available genomics data. As there is no overlap between the 50 gene signature used here and the 1072 genes which show copy number changes (see Chapter 3), we look for enrichment of samples in each hDPMM cluster with certain genomic characteristics. Qualitatively, it can be seen from Fig.5.8 that the gene expression pattern from the hDPMM clustering agrees to the corresponding genomic mutations. The typical GBM aberrations containing samples are defined in Appendix C. Quantitatively, we find that there is significant enrichment of DPMM Cluster I with Non-Typical GBM samples (p-value of $2.2e-16$). The frequency of CNV calls are also significantly different in DPMM Cluster I samples and DPMM cluster II samples. (p-value $7.5e-16$). Thus we have shown that the genomic data (CNV) can be used to validate the clustering obtained from hDPMM using transcriptomics data.

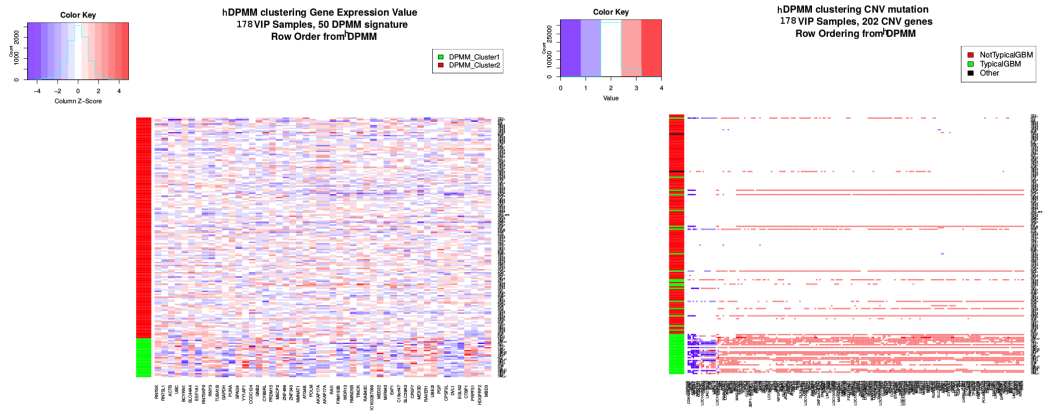

Figure 5.8: Results of hDPMM on 178 IDENTIREST samples which also have corresponding genomic data. On the left is the gene expression with the samples being arranged according to the hDPMM. On the right we have the same sample ordering but the corresponding CNV data being shown. The

To summarize, we have explored the *sample* heterogeneity (jointly for Peripheral and Central Cells) using the modified SBC (hDPMM). The clustering obtained from the hDPMM model on the transcriptomics data agrees well with the respective genomic changes, thus strengthening our belief that the clusters that we obtain are indeed biologically distinct.

"Reading furnishes the mind only with materials of knowledge; it is thinking that makes what we read ours."

John Locke

6

Statistical foundations of the SBC

6.1 Introduction

Survival-based-Bayesian Clustering (as introduced in Chapter 4 and later applied in Chapter 5)is fundamentally an instance of a non-parametric Bayesian model. Bayesian non-parametric models are becoming increasingly important in Biostatistics, an excellent resource in this domain is[HHMW10]. We first explain the general idea of Bayesian non-parametric

models.

Probability models usually form the backbone of many statistical problems. Data is seen as realizations of a collection of random variables $X_1, X_2, \ldots X_n$, where X_i itself could be a vector of random variables corresponding to data that are collected on the i-th experimental unit in a sample of n units from some underlying population. Usually, the working assumption is that the X_is are independently drawn from some underlying probability distribution G. Bayesian models address the uncertainty that exists about G. Let g denote the probability density function (p.d.f.) of G. A statistical model arises when G is known to be a member g_v from a family $G = \{g_v : v \in \Theta\}$ labeled by a set of parameters v from an index set Θ. Models which are described through a vector v of a finite number of real values are referred to as finite-dimensional or *parametric models.* This implies that they can be described as $G = \{g_v : v \in \Theta \subset \mathrm{R}^p\}$: The aim of the analysis is then to use the observed sample to determine a plausible value for v, or a set which may contain v.

The above assumption which constrains the statistical model to a specific set of parameters may be too restrictive in many situations. Such an assumption on the parametric form of the distribution may also limit the scope of this type of inference. These problems are encountered in many areas of statistical modelling in biology where the sample size n is quite small. Hence, to allow for greater flexibility and robustness against mis-specification of a parametric statistical model, we need to consider models where the class of densities is so large that it can no longer be indexed by a finite dimensional parameter v, and we therefore require parameters v in an infinite dimensional space[HHMW10]. Inference of such high dimensional parameter space in an Bayesian setting involves placing priors over the infinite-dimensional parameter v, such priors are known as Bayesian nonparametric (BNP) priors.

This is also then taken to be the definition of Bayesian non parametric models. That is, we define BNP priors as probability models for infinite-dimensional parameters and refer to the entire inference model as a BNP model. The two most popular of BNPs in Statistical Modelling are: Dirichlet Processes and Gaussian Processes. While Dirichlet Process specify such a non-parametric prior over probability distributions [Teh11], Gaussian Processes are used as a prior over an unknown functions [Ras04] .

The above described infinite-dimensional parameters of interest can, generally, be viewed as functions. These functions of interest include probability distributions (such as estimating the distribution of $p(X)$ where X is the molecular data from patients), or they could be conditional trends, e.g. mean regression functions ($p(y|X)$, the probability of observing an outcome y for a patient, given the molecular data X). The unique aspect about the SBC is the fact that it can be viewed as BNP where either i) we are interested to estimate the probability distribution and hence the clustering property of $p(X, y)$ or ii) we are interested in estimating $p(y|X)$ as a non-parametric function. This fact is reflected in the ability of SBC to make two kinds of predictions (both survival predictions as well as class or stratum predictions). As our current model combines aspects from both explorative clustering models as well as prediction models, we take a look at SBC from both points of views. It will be shown in the following sections that SBC can be viewed independently from both these perspectives. In this regard, we also mention that SBC is a generative approach modelling thus estimating $p(y, X)$ rather than $p(y|X)$.

The joint modelling (y, X) comes at a cost: As the dimension of X (molecular data) increases, the SBC concentrates more on fitting a distribution for X rather than $p(y|X)$. In order for such an approach to be competitive to a discriminatory approach appropriate

feature selection techniques need to be employed. Thus feature selection is another vital aspect to SBC which has been explained in detail in last part of this chapter.

6.2 SBC as a clustering model

Our present SBC model has similarities to the work of Müller[MEW96] in the area of Bayesian curve fitting. The same authors provide a R-package 'PPMx' implementing their idea along with some later modifications[MQR11]. The key idea is to model $z_i = (y_i, X_i)$ as joint random variable with a DP prior. Thus the SBC model simply reduces to a BNP prior on the joint distribution of the response and the covariates representable by the following hierarchy:

$$\begin{aligned} y_1, \ldots y_n | X, \phi &\overset{ind}{\sim} p(y_i | X_i, \phi_i) \\ X_1, \ldots X_n | \gamma &\overset{ind}{\sim} p(X_i | \gamma_i) \\ (\phi_i, \gamma_i), \ldots (\phi_n, \gamma_n) &\overset{ind}{\sim} G \\ G &\sim DP(\alpha G_0) \end{aligned}$$

with $G_0 = G_0\phi \times G_0\gamma$, this equation leads to the following joint probability model (with $\Lambda = (\gamma, \phi)$)

$$p(y, X|\Lambda) \sim \sum_{j=1}^{k} n_j p(y, X|\Lambda_j^*) + \alpha \int p(y, X|\Lambda) dG_0(\Lambda)$$

The above integral is difficult to calculate because of the complex dependency between γ and ϕ. The two sets of parameters are locally independent (cluster wise), but globally dependent. This present model was used in[MEW96] and then re-branded in[MQR11]. This tricky

dependency which has been highlighted in [MQ10] forces us to use Auxiliary variable approximation in our solution (see Chapter 4). While the Bayesian curve clustering works only for extremely low dimensional data sets with continuous outcome, our SBC approach has been developed primarily for survival output y and high-dimensional correlated input variable X. The small sample size along with high-dimensional input also forces the SBC to adopt a more robust, hierarchical prior over the parameter space $\Lambda = (\gamma, \phi)$

It is also worth noting that the above model can be cast as what is known as a 'Product Partition Mixture Model' (PPM). In such models we model the partition ϱ_n probability where $\varrho_n = (S_1, \ldots S_{k_n})$denote a partition of the n experimental units into k_n number of S_j subsets.

$$p(\varrho_n) \propto \prod_{j=1}^{n_k} c(S_j)$$

where $c()$ is the cohesion function. In a simple DP model this is given as $c(S_j) = \alpha(|S_j| - 1)$. The SBC model can be seen as a modification of the simple DP prior, making the cohesion function dependent on X, such that $p(\varrho_n|X)$ becomes a function of X in the following fashion:

$$p(\varrho_n|X) \propto \prod_{j=1}^{n_k} g(X_j^*)c(S_j)$$

$X_j^* = (X_i)_{i \in S_j^*}$. The function $g()$ is a similarity function between X points in the same cluster. The form of $g()$ induced by our SBC model can be seen in [MQR11]. A modification to the above cohesion function was suggested by Park & Dunson [PD10]. The authors motivate their method as the modelling task for $p(\varrho_n|X)$ in a different manner than the earlier approach of [MEW96]. These two approaches have been compared with each other along with Hierarchical Mixture of Experts, FlexMix in [MQ10]. Their proposed model uses the following modified

cohesion function:

$$c(X_j^*, S_j^*) = \alpha(k_j - 1)! \int \prod_{i \in S_j^*} f_2(X_j|v) dG_{ov}(v)$$

where f_2 describes the likelihood model for X_i, the base measure G_{ov} is chosen conjugate to f_2 such that the integral can be calculated. Given this, the prior over cluster assignment ϱ_n can be written:

$$p(\varrho_n) \propto \prod_{j=1}^{n_k} c(X_j^*, S_j)$$

[PD10] develop a 'Generalized Polya Urn' Scheme now to update the parameters ϕ (just as the Polya Urn Scheme which was used to update the joint parameter Λ in the present SBC method). Hence, this method provides an alternative strategy for updating the parameters φ_i of the conditional distribution $p(y_i|X_i, \varphi_i)$, avoiding the computation of parameter γ_i for the distribution of $p(X_i|\gamma_i)$, which in many cases may not be that important. We have described this method as an alternative to the SBC sampling approach so that one can appreciate the subtleties within the SBC model along with its different proposed variants in the literature [MQR11] [PD10] . Sampling for the parameter γ_i for the distribution of $p(X_i|\gamma_i)$ is done explicitly in SBC as it also allows us to rank high-dimensional features in γ_i in terms of their importance to clustering. This ranking is useful later on when we interpret our SBC model.

From the above description it's easy to see that probability models $p(\varrho_n)$ can systematically be generalized to $p(\varrho_n|X)$ and hence can be termed as covariate-based random partition models or covariate based product partition model (also known as PPMx models in Müller [MQR11]) . The implied conditional distribution $p(\varrho_n|X)$ defines a probability model

for ϱ_n, indexed by covariates X, such an approach becomes particularly convenient when the model is combined with the sampling model for observed data y as $(p(y_i|X_i, \varphi_i))$[MQR08]. One important limitation of such a approach, especially in the case of multi-modal data for X is the specification of a probability model for different types of the co-variates. For a long list of mixed kind of data formats, it becomes increasingly challenging to define meaningful probability distributions. Such situations are quite common also for many biomedical applications where X_i could also include patient data like treatment history, age, ethnicity, insurance coverage,location etc. These details also have to be included in the model without the explicit assumption of any clustering pattern within it.

6.3 SBC as a predictive model

Another point of view is to look at SBC from a purely predictive perspective. In this context, similarities can be seen with the works of Hannah et al.[HBP11] and Shahbaba et al.[SN09]. These authors provide a different terminology to their work such as Mixture of Generalized Linear Models or Non Linear Modelling using Dirichlet Process Mixture Model. Their approach can be summarized as follows: By modelling the joint distribution of the response variable, y, and of the covariates, x, non-parametrically using Dirichlet process mixtures, along with keeping the relationship between y and x linear within each component of the mixture, the overall relationship between y and x becomes non-linear if the mixture contains more than one component, with different regression coefficients[BIB15]. Thus, the idea of cluster-specific linear relationships that we have in our current SBC approach provides an elegant approach for modelling of an overall non-linear relationship. In other words, SBC can be seen as a new class of methods for nonparametric regression. Given a

data set of input-response pairs, the SBC produces a global non-linear model of the joint distribution through a mixture of local linear models[HBPII]. The key difference between the SBC approach and that of Hannah et al. and Shahbaba et al. is the high dimensionality of x which is modelled explicitly using a hierarchical model as well as the nature of the output being censored survival times in SBC rather than continuous or categorical as used other works

The predictive model can be understood as a set of the following equations as written below, where f_x represents the assumed distribution of X (multivariate normal in our case) and f_y is the conditional distribution that $y|X$ follows (truncated univariate normal in the case of SBC), the form of f_y can be varied, thus giving rise to generalized linear models or even classification models:

$$
\begin{aligned}
P &\sim DP(\alpha G_o) \\
\Theta = (\Theta_{i,x}, \Theta_{i,y})|P &\sim P \\
X_i|\Theta_{i,x} &\sim f_x(.|\Theta_{i,x}) \\
Y_i|\Theta_{i,y} &\sim f_y(.|X_i, \Theta_{i,y})
\end{aligned}
$$

The Dirichlet process prior on the distribution P clusters the covariate-response pairs (x, y). In a predictive setting, the training phase involves observing both x and y. The posterior distribution of P allows the data to cluster to nearby covariates that exhibit the same kind of relationship to their response. During the predictive or test phase, when the response is not observed, predictive expectation $E(f_y(Y_i^*|X_i^*)$ can be understood by clustering the covariates based on the training data and then predicting the response according to the model

associated with the covariates cluster[HBP11]. Therefore, the DP prior acts as a kernel for the covariates thereby calculating the distance between two points by the probability that the hidden parameter Θ is shared.

This kind of modelling provides a useful alternative for predictive models. Many other powerful predictive models, such as Generalized Linear Models (GLMs) and Gaussian processes (GP), make assumptions about data dispersion and homoscedasticity. Over-dispersion occurs when the data variance is positively correlated with the predicted model mean. The modelling technique described above successfully creates classes of models that account for over-dispersion. A model is homoscedastic when the response variance is constant across all covariates; a model is heteroscedastic when the response variance changes with the covariates[HBP11]. Models like traditional GLMs or GPs are homoscedastic and hence give poor data-fits when this assumption is not met. In contrast, the modelling framework described by SBC captures heteroscedasticity when mixtures of linear models (or GLMs) are used. The mixture model setting allows variance to be modeled by a separate cluster-specific parameter or by a collection of clusters in a single covariate location. As a result of this approach we end up with smoothly transitioning heteroscedastic posterior response distributions.

Another popular technique that is related to SBC is 'Mixture of Experts'. Mixture of experts model was the first attempt at using a collection of simple linear models to build a non linear model and was proposed by Jacobs et al.[JJNH91]. This model was introduced as a supervised learning procedure for models that consist of many local experts, each specialized for a subset of data. A gating network decides which expert should be used for a given data point, the parameters of which are learned from the training data. These approaches

provided fixed number of experts and came with the risk of over-fitting and complicated model selection. Rasmussen and Ghahramani[RG02] extended the number of experts to potentially infinite by defining the gating network to be based on an input-dependent adaptation of Dirichlet process. Later, Meeds and Osindero[MO06] proposed an alternative view of the Mixture of Experts model by proposing a joint mixture of experts model over covariates and response variable.

A good way to emphasize the importance of Mixture of Experts Model or generally mixture of predictive models is by looking at the Fig.6.1. The left plot shows data points drawn from two classes denoted red and blue, in which the background color (which varies from pure red to pure blue) denotes the true probability of the class label. The center plot shows the result of fitting a single logistic regression model using maximum likelihood, in which the background color denotes the corresponding probability of the class label. Because the color is a near-uniform purple, we see that the model assigns a probability of around 0.5 to each of the classes over most of input space. The right plot shows the result of fitting a mixture of two logistic regression models, which now gives much higher probability to the

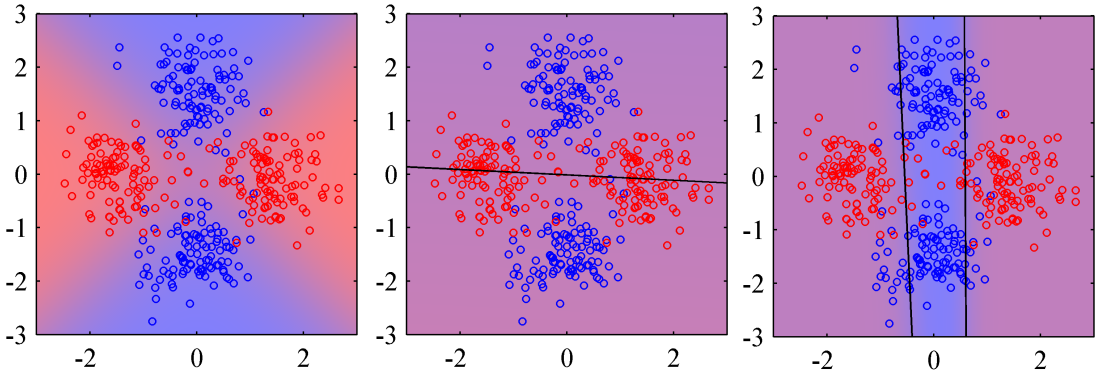

Figure 6.1: Depiction of mixture of experts models for prediction. Image reproduced from Christopher Bishop's book " Pattern Recognition and Machine Learning"[Bis06]. Explanation of the figure is contained in the text

correct labels for many of the points in the blue class. Thus identifying the distribution of points X helps us correctly define two different prediction models which then accurately capture the overall complicated relationship between X and y (the class labels).

While discussing SBC as a predictive model, one should also note that it is a generative model. Generative models generate all possible values of data X given the target variable y by concentrating on modelling the generation process of the data, i.e. $p(X|y)$, this in turn means that we estimate the joint model $p(X, y)$ as $p(X, y) = p(X|y)p(y)$. The other class of predictive models are known as discriminative models which provide a model only for the target variable $p(y|X)$. Both generative models and discriminative models can be used for the task of predictions. Generally speaking, generative models have some advantages over discriminative models such as providing a framework to handle missing or partially labeled data. They can also, potentially, augment small quantities of expensive labeled data with large quantities of cheap unlabeled data. Two highly successful examples of predictive models which are generative are: Naive Bayes Classifier and Latent Dirichlet Allocation (LDA)[BNJ03] both of which have been successfully used in applications like document labeling and image analysis. These generative models perform well in classifying documents (like e-mail spam detection) with previously unknown patterns.

At the same time that generative models can be quite successful in many problems, they can also be computationally intensive. Another critical drawback of generative models is that finding a good estimate for the joint distribution of (X, y) does not in necessarily translate into a good estimate of decision boundaries. On the other hand discriminative models are often preferred for their good class separation abilities and are often computationally fast. However, using a generative model in our SBC approach allows estimate of

the hidden structure in the data X. This leads to a more interpretable prediction model. Our SBC approach is based on the intuitive assumption that different regions in the data X can have different predictive models and hence modelling $p(X, y)$ allows us to discover the pattern in the data.

SBC approach also has similarities to curve clustering approaches such as FlexMix[GL08] which implements a general framework for fitting discrete mixtures of regression models. As it is parametric (the number of curves need to specified beforehand) and non-Bayesian it uses EM algorithm for parameter estimation. It is fairly flexible as regressors and responses may be multivariate with arbitrary dimension. Our SBC model has been compared to FlexMix; the non-parametric and Bayesian nature of SBC makes it not only better equipped for noisy Biological data, it also yields predictive advantage. To overcome limitations of a parametric mixture of regression models Bayesian profile regression (PR)[MPJR10] was proposed. In its original formulation it is closely related to the augmented response model described before[MQR11]. It's a non-parametric two-step procedure that flexibly models covariates X (using a DP process prior) that identifies "important" cluster specific covariates and then then connects them to a data model. This model is an alternative to regression models, non-parametrically linking a response vector to covariate data through cluster membership[MPJR10]. The initial Profile Regression (PR) model which was motivated from epidemiological applications considered target variable y as a binary outcome, this was later extended in PReMiuM[LHA+15] which offers many forms of $f_y(Y_i|X_i, \ldots)$ like binary, categorical, count and continuous responses. PReMiuM also allows to account for the missing values for the covariates. It has also been discussed in Barcella et al.[BIB15] PR (or PReMiuM) can be equivalently represented as a Product Partition Mixture Model. As opposed

to FlexMix or PR, our SBC approach only uses a survival model, where the target variables y can be censored, also there is considerably more focus in our SBC approach to robustly model the high-dimensional input X.

6.4 Variable Selection

Irrespective of whether we consider SBC as a clustering or a predictive model, it has been shown that the success such models requires relatively fewer covariates (the dimension D of the vector X). This is necessary because as the number of covariates increase, their influence on partition probabilities outweighs any information the response y provides for clustering [QMP15]. This often results in a large number of clusters with very few (even single) observations hence resulting in a poor model fit and out-of-sample prediction as shown in [QMP15]. Therefore feature/covariate selection plays an important role in SBC and similar statistical models. Generally, the covariates are chosen based on their importance in two different areas [BIB15]:

1. relevance in determining the clustering of the observations X, i.e. its effect in distinguishing different clusters.

2. effect on the level of a response/target variable through the likelihood of $f_y(Y_i|X_i, \ldots)$

In our SBC model, we adopt two different feature selection (feature importance) strategies for the above described two tasks. Apart from variable selection another possible technique to avoid the problem of poor model fit is to decouple clustering of the observations X_i with the clustering of the response model $f_y(Y_i|X_i, \ldots)$, this was the approach adopted by Wade et al. [WDPT14] by proposing a enriched Dirichlet process prior over the two different kinds

of partition. Again, this scheme was proposed to down weight a covariate's influence on clustering. Variable selection is the much more common alternative approach that attempts to accommodate a large number of covariates. In particular, Quintana et al.[QMP15] propose a variable selection technique based on the PPMx model by introducing introducing $v_{j,l}$ binary indicators for the jth cluster and lth covariate. Covariates can be active $v_{j,l} = 1$ or inactive $v_{j,l} = 0$. The overall prior probability for the partition ϱ_n then can be written as:

$$p(\varrho_n|X, \mathrm{v}) \propto \prod_{j=1}^{n_k} c(S_j)(\prod_{l=1}^{D} g(X_{j,l}^{v_{j,l}}))$$

A hierarchical logistic prior is used for $p(v_{j,l})$. Such a hierarchical prior allows for the sharing of the information about important covariates across clusters. Another approach was taken by Chung and Dunson[CD09] which use Bernoulli variable η_{jl} ($l = 1 \ldots D$) instead of $v_{j,l}$ and for clusters $j = 1 \ldots K$) which takes values either 0 (feature omitted) or 1 (feature selected) for a particular cluster j. Until this point, this approach is identical to that of Quintana[QMP15], however Chung et al. use a different set of hierarchical priors over η

$$\eta_{jl} \sim Bernoulli(\varkappa_j)$$

To borrow information across mixture components, they use the sparseness-favoring prior of[CD09]:

$$\varkappa_j \sim 1(w_j = 0)\delta_0(\varkappa_j) + 1(w_j = 1)Beta(\varkappa_j; a, b)$$

$$w_j \sim Bernoulli(0.5)$$

The posterior for η_{jl} can be then be updated from its conditional distribution. Feature selection again plays an important part in PR[MPJR10] and closely resembles the techniques described above. The probability that a data point x_i is factorized as a product over its features in the cluster j:

$$x_i|\eta_{j1}, \dots \eta_{jD} \sim \prod_{d=1}^{d=D} p(x_{id}|\eta_{jd})$$

In the next step, variable selection is then performed by replacing the distribution of each covariate as following[BIB15]:

$$p(x_{id}|\eta_{jd}, \pi_d) = \pi_d p(x_{id}|\eta_{jd}) + (1 - \pi_d) r_d(x_{id})$$

$\pi_d \in (0, 1)$ is a continuous weight and $r_d(x_{id})$ indicates the proportion of times covariate d takes value x_{id}. Here. π_d indicates that covariate d is informative in terms of clustering. The authors then define for π_d either a Beta hyperprior distribution for each d or alternatively a mixture of a Beta distribution and Dirac measure. This then borrows information over the clusters as mentioned above.

The three variable selection techniques describe above allow for the feature selection of important covariates in clustering of the observation, however they do not address the question of which covariates are most related to the target variable. This concern was addressed by Barcella et al.[BIBML16] where they introduce a spike and slab penalty to model the

relationship between X and y:

$$
\begin{aligned}
y_1, y_2, \ldots y_n | X, \Theta, \lambda &\overset{ind}{\sim} \mathbb{N}(y_i | x_i \vartheta_i^t, \lambda) \\
x_i | \eta_{j1}, \ldots \eta_{jD} &\sim \prod_{d=1}^{d=D} p(x_{id} | \eta_{jd}) \\
(\vartheta_i, \eta_i), \ldots (\vartheta_n, \eta_n) | G &\overset{ind}{\sim} G \\
G &\sim DP(\alpha G_0)
\end{aligned}
$$

where Θ is matrix of parameters with n rows and D columns. Now the base distribution $G_0 = G_0\vartheta \times G_0\eta$ is then given a spike and lab prior as follows:

$$
G_0 = \prod_{d=1}^{D} \{\pi_d \delta(\vartheta_d) + (1 - \pi_d)\mathbb{N}(\vartheta_d | \mu_d, \tau_d)] Beta(\eta_d | a, b)\}
$$

Thus we have given a brief overview of variable selection in statistical models similar to SBC (broadly termed as covariate dependent Dirichlet Process Mixture Models or DPMMx). The major objective is to identify the most important covariates for the partition of the observations. From a statistical point of view variable selection in DPMMx like models is particularly necessary as higher dimensional covariates can dominate the DPMMx likelihood. Solutions proposed in the literature often involve introduction of latent variables which mitigate the effect of specific covariates in determining the partition . In other applications it may also be important to identify those covariates that best explain a response variable[BIB15]. The variable selection introduced in Barcella et al.[BIBML16] applies variable selection in regression settings by specifying spike and slab distributions as base measures. In our SBC approach we use another popular method for Bayesian variable selection, viz.,

Bayesian LASSO[PC08].

7

Conclusions

7.1 Overview

The present work is an attempt to better understand disease heterogeneity in patients (and tissue samples) by leveraging the large amounts of molecular data that is presently being made available from large patient cohorts. In the context of cancer, this large amounts of molecular data has not only provided new insights into the biological diversity of hu-

man cancers[SML15] but also enabled the possibilities to discover previously unknown disease subtypes. A better understanding of heterogeneity within the disease can then be made use of to deliver better personalized medicine solutions. Therefore, our work is useful in identifying/classifying patient strata based on certain discriminating molecular features between patient sub-groups, in other words defining sub-groups based on sub-group specific biomarkers can be understood from the perspective of refining molecular disease taxonomies. Such taxonomies can potentially play a role in two important ways:

- Development of targeted therapies against certain patient sub-groups. An example of such an approach in GBM is the use of targeted cancer immunotherapy for blocking the PD-1/PD-L1 pathway[XHIY17]. The success of such a targeted approach relies heavily on the ability to understand the right patient strata which need to be treated with immunotherapy.

- Better understanding of clinical therapy response vs. non-response. Typically, GBM (or other cancers) is treated with chemotherapy after surgery, but a considerable fraction of patients is chemotherapy resistant. Looking into therapy resistance from the perspective of molecular strata, which are also clinically supported, might be a way to understand and potentially overcome therapy resistance in the future.

From a clinical point of view, the work in this thesis can also be thought of as a supplementary decision support system. For example, in the concrete case of our RIvsRU study, such a classifier could be useful for clinicians to decide targeted radio therapy for patients. In clinical practice, such a classifier can be translated into a prognostic toolkit (a customized gene assay) which can be used in a cost-effective manner. This prognostic toolkit will only

need the gene expression of a handful of genes which are biomarkers for progression. Such a prognostic toolkit, however, would first require a prospective validation for the RIvsRU classifier. Furthermore, the prognostic toolkit would be subject to clinical trials for regulatory approval.

7.2 Achievements summary

The thesis introduces the concept of patient stratification, personalized medicine and the use of modern high-throughput data in the introductory chapter. This is further motivated by giving a case study example of Glioblastoma. Glioblastoma has been used throughout the thesis as one of the primary disease test cases. This is in part because of the IDENTIREST project which aimed at studying the recurrence of Glioblastoma and devising new therapeutic approaches.

Next, the thesis provides a brief introduction to the vast field of -omics data and its use in personalized medicine. The use of -omics data in statistical modelling comes with a whole range of issues, most notably that of high dimensionality and correlation. This is touched upon in Chapter 2 along with standard paradigms for machine learning methods for patient stratification. The goal in this chapter was to motivate the need to look at multi-omics data sets in the context of patient stratification. A detailed qualitative review for many such statistical and machine learning techniques is then provided which highlights the challenges and strategies in multi-omics data integration for patient stratification.

Next, the use of actual statistical methods on Glioblastoma patients is explored in the context of the IDENTIREST project in Chapter 3. After the initial project description along with the Pilot Study, the focus was on the key question of exploring heterogeneity in

the samples using multi-omics data (mRNA and CNV). We were able to see differences between the Central and Peripheral samples on both kinds of omics data (mRNA and CNV). We found consistent patterns between mRNA and CNV data for some of the genes which showed strong differential expression between Peripheral and Central Samples.

Following on that, we were also able to show the differences between the Central and Peripheral samples using previously established classification scheme of Verhaak et al. [VHP+10b]. As a final application of Machine Learning based classifiers we showed that using transcriptomics data we were able to successfully predict the site for tumor recurrence. We show the good predictive performance first via cross-validation scheme and then on an independent validation set. We next provided biological interpretation and visualization of our four pathway signature. Notably, our signature also contained strong survival information (PFS and OS), thus adding another layer of interpretation.

This RIvsRU classifier is envisioned as a decision support system for the clinicians. The vision is to enable preventive radio-therapy, which is targeted against the location of most likely tumor recurrence. To enable this vision there are several steps needed:

- A prospective clinical validation of the RIvsRU classifier, which demonstrates comparable prediction performance to what we observed so far. Currently, we estimate that around 100 patients are needed for this purpose.

- After the prospective validation a clinical study would then need to show the efficacy of such a targeted radio therapy on the recurrence of GBM in patients when compared to standard of care.

- If indeed the benefits of targeted radio therapy can be demonstrated, the classifier

would then be needed to be translated into clinical practice using a cost-effective customized assay. In this case, the model parameters have to be translated from the current gene expression assay platform to the much smaller, customized assay platform. The final goal there would be the development of such a customized assay as a prognostic toolkit, which can be used in day-to-day clinical practice.

- As with any other diagnostic/prognostic tools, there needs to be rigorous clinical trials approved by the regulatory agencies (FDA,EMA) before it can be brought into the market.

- Finally, such a prognostic tool would also need a Cost-benefit analysis taking into account the potential benefit on the health of patients in relation to the overall costs for developing such a prognostic tool.

Hence, we can understand the clinical potential of our RIvsRU classifier and the future steps that need to be taken in order to translate it into a clinically useful prognostic tool.

The Survival Based Bayesian Clustering[AF17] is introduced in Chapter 4. This technique is a fully Bayesian clustering algorithm which takes in clinical end-points of patients along with heterogeneous -omics data and accomplishes two key tasks in one:

- clinically relevant patient sub-group identification on training data and

- prediction of patient subgroup and survival time on testing data.

Our SBC algorithm was motivated by the need to approach the problem of patient stratification taking into account patient specific survival risk models. Effectively, we get clinically and biologically relevant patient subgroups out of our approach. Another important motivation was the predictive utility of our method that we demonstrated using cross-validation

results on two important cancer data sets. We also compared the SBC method to ad-hoc techniques and found that SBC outperformed the competing methods. The key ability of SBC to identify patient-subgroups differing in survival constitutes an advantage compared to existing approaches. Furthermore, SBC is also principally able to take into account more than one -omics data source (mRNA and miRNA). Moreover, we also demonstrate that certain sub-types from our model are particularly enriched in certain biological markers (for example ER status for breast cancer) and also correlate strongly with some sub-types in the well established classification schemes. This coupled with the ability of SBC to identify sub-group specific biomarkers which have also been reported in the literature makes SBC a potent novel tool in the area of patient stratification and a vital step towards a more clinically relevant dissection of patient heterogeneity[AF17].

As a follow up on the practical applications of SBC, we demonstrate its utility in the context of the IDENTIREST project in Chapter 5. We explore the validation of the SBC model trained on Verhaak data set to our IDENTIREST cohort. We were able to predict potentially clinically relevant patient strata (with respect to Overall survival). We were also successful in finding distinct genomic patterns (CNV) in the predicted patient strata which serves yet again to validate our prediction results. As a second application, we use a variation of SBC (hDPMM) to better understand the heterogeneity of the samples (and patients) in an unbiased manner. We used the hDPMM to analyse the full cohort of IDENTIREST samples. The clustering that we obtained from fitting the hDPMM was enriched in the surgeon defined classes (Peripheral and Central). Also using the genomics data (using CNVs) we further tried to validate the hDPMM clusters. Thus in both the sections of this Chapter we have used multi-omics data sets (gene expression and CNV data) from the

IDENTIREST cohort to validate our results.

We conclude the thesis with an elaborate statistical description of our SBC method in Chapter 6. We first motivate SBC as a general generative Bayesian non-parametric model. We next explain two fundamentally distinct ways to look at SBC- a) from a non-parametric clustering point of view and b) from a non-parametric predictive modelling view. In this context, we contrast SBC with other popular machine learning techniques like Mixture of Experts and Bayesian Profile Regression. Similarities between Generalized Linear Models and SBC are also discussed. In the end, we also provide a brief overview of variable selection in models similar to SBC. These theoretical perspectives on the SBC also point to the statistical shortcomings and possible workarounds for such shortcomings. Overall, this chapter gives an insight to the statistically involved nature of our proposed SBC approach and directions for future methodological improvements in the same.

Our SBC method can be judged to be of value from both statistics/ machine learning perspective as well as clinical/medical perspective. The methodological statistical developments are driven from the needs of clinical applicability and we show that the results obtained can be interpreted from a biological point of view. As such the SBC can be seen as an important tool which can help/augment clinical decision making. In a wider context, we believe that this thesis is a step closer towards the goal of achieving personalized medicine solutions using molecular -omics and clinical patient data. This thesis fills an important void in the scientific literature on the need to explore patient heterogeneity using multi -omics data in combination of clinical data. We have also shown that the application of machine learning techniques on such data can indeed be a crucial part of the puzzle in the field of patient stratification. Therefore, this work is of value for health care data scientists,

biologists studying disease mechanisms and medical doctors treating patients.

7.3 Future directions

Looking ahead in the future, this thesis can be a good starting point for further explorations of the fusion and interplay of clinical and molecular data for the development of predictive machine learning patient stratification algorithms. There are broadly two different directions one could take: a) looking into various novel sources of molecular/clinical data apart from ones used in this thesis (e.g. bioimaging data, electronic health records, wearable mobile technologies etc.) and b) looking into different algorithms for patient stratification (e.g. deep learning approaches, matrix factorization or graph-based approaches). The stratification algorithms then need to translated into clinical practice by testing them in clinical studies. Success in clinical studies of these stratification techniques will be the ultimate criterion for their wide-scale applicability, acceptability and adoption.

A

Publication List

Here is a list of my peer-reviewed works that formed the basis of this thesis:

i Ahmad, Ashar, and Holger Fröhlich. "Towards clinically more relevant dissection of patient heterogeneity via survival-based Bayesian clustering." Bioinformatics 33.22 (2017): 3558-3566. [AF17]

ii Ahmad, Ashar, and Holger Fröhlich. "Integrating Heterogeneous omics Data via Sta-

tistical Inference and Learning Techniques." Genomics and Computational Biology 2.1 (2016): 32. [AF16]

The following is the list of other published peer-reviewed works. These works were not directly related to this thesis but were published during the course of my PhD:

i Schmitt, Ina, Oliver Kaut, Hassan Khazneh, Laura deBoni, Ashar Ahmad, Daniela Berg, Christine Klein, Holger Fröhlich, and Ullrich Wüllner. "L-dopa increases α-synuclein DNA methylation in Parkinson's disease patients in vivo and in vitro." Movement Disorders 30, no. 13 (2015): 1794-1801.

ii Sharma, Amit, Oliver Kaut, Anna Pavlova, Holger Fröhlich, Ashar Ahmad, Ina Schmitt, Osman El-Maarri, Johannes Oldenburg, and Ullrich Wüllner. "Skewed X-chromosome inactivation and XIST locus methylation levels do not contribute to the lower prevalence of Parkinson's disease in females." Neurobiology of aging 57 (2017): 248-e1.

iii Narimani, Zahra, Hamid Beigy, Ashar Ahmad, Ali Masoudi-Nejad, and Holger Fröhlich. "Expectation propagation for large scale Bayesian inference of non-linear molecular networks from perturbation data." PloS one 12, no. 2 (2017): e0171240.

The above list can also be accessed for actual version of the published works under: `https://scholar.google.de/citations?user=AOWOdDwAAAAJ&hl=en`

B

Sampling Algorithm for SBC

The appendix explicitly lists the individual steps for the Gibb's sampling used for model fitting SBC. We have defined the prior distribution of the parameters in Chapter 4. The goal of fitting a hierarchical Bayesian model (like SBC) is to obtain the joint distribution of all the parameters and hyper parameters in the model. Gibb's sampling is used when the overall joint distribution of the parameters is difficult to obtain, however based on the

hierarchical structure, conditional distributions of each of the variables can be obtained (conditioned on the variables in the Markov blanket of that variable). The Hierarchical structure of the SBC (as shown in figure 4.1) provides a Bayesian Network representation of our parameters and is used to define the conditional distribution for each individual parameter (represented as one node). It can be shown that iteratively drawing samples from the conditional distribution for each parameter constitutes a Markov Chain and the stationary distribution of that Markov chain is the joint distribution [GCSR04].

Now we describe in detail the overall sampling scheme. Here are some notations: Let $\Theta_k = \{\mu_k, S_k\}$ be the parameters of the hierarchical Gaussian Mixture model (GMM), $B_k = \{\beta_{ok}, \beta_k, \sigma_k^2\}$ be the parameters of the AFT bayesian LASSO (BLASSO) model.

Correspondingly, we have hyper-parameters of the hierarchical GMM $H_1 = \{\xi, \varrho, W, \varphi\}$, and of the AFT BLASSO $H_2 = \{\lambda, \tau^2\}$. If we have $k = 1 \ldots K^-$ occupied clusters at any moment, each containing n_k data points, the total number of parameters are $\Phi_k = \{\Theta_k, B_k\}_{k=1\ldots K^-}$ and $H = \{H_1, H_2, \alpha\}$ hyper-parameters apart from cluster-indicator variables $c_1, c_2, \ldots c_N$. The gaol is to sample each of these variables by following at iterative strategy which can be summarized as follows:

- Update all parameters $\Phi_{k_{1\ldots K}}$
- Update all the hyper-parameters H
- Update all the cluster indicator variables $c_1, c_2, \ldots c_N$
- Update DP concentration parameter α

As described in Chapter 4, let $G_o(\alpha, H)$ be the overall prior distribution over the parameter space $\{\Phi\}$. While updating the cluster-indicator variable, we follow the Algorithm 8, as

described in Neal[Neaoo] with the Auxiliary variables set to two $U = 2$:

Algorithm 1 SBC Gibb's Algorithm

procedure UPDATE CLUSTER INDICATOR c_i(U =2)
    For each $i = 1 \ldots N$
    Let k^- be the distinct number of active clusters
    Check if the present cluster assignment also has other data points:
    if $c_i = c_j$ for some $j \neq i$ then
        Draw Φ_{k^-+1} and Φ_{k^-+2} independently from $G_0(\alpha, H)$
    end if
    Check if the present cluster assignment was the singleton data point:
    if $c_i \neq c_j$ for all $j \in 1 \ldots N$ then $c_i = k^- + 1$,
        Assign $\Phi_{k^-+1} = \Phi_{c_i}$
        And Draw Φ_{k^-+2} from $G_0(\alpha, H)$
    end if
    Sample c_i from the following distribution
    if $k = 1, 2 \ldots k^-$ then

$$p(c_i = k) \propto \frac{n_{-i,k}}{N - 1 + \alpha} \mathcal{N}(w_i | \mathrm{B}_k) \mathcal{N}(\mathrm{X}_i | \Theta_k)$$

    end if
    if $k = k^- + 1$ or $k = k^- + 2$ then

$$p(c_i = k) \propto \frac{\alpha}{N - 1 + \alpha} \mathcal{N}(w_i | \mathrm{B}_k) \mathcal{N}(\mathrm{X}_i | \Theta_k)$$

    end if
    $i \leftarrow i + 1$.
    goto *top*.
end procedure

C

GBM specific mutations

Here is a list the typical genomic mutations in GBM

- Gains on Chromosome 7
- Losses on Chromosome 1p, 6q, 9p, 10q and 13q
- Loss of Heterozygosity on Chromosome 10, 14q, 17p13.3

References

[AED+00] Alizadeh, Ash A. ; Eisen, Michael B. ; Davis, R E. ; Ma, Chi ; Lossos, Izidore S. ; Rosenwald, Andreas ; Boldrick, Jennifer C. ; Sabet, Hajeer ; Tran, Truc ; Yu, Xin u. a.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. In: *Nature* 403 (2000), Nr. 6769, S. 503–511

[AF16] Ahmad, Ashar ; Fröhlich, Holger: Integrating Heterogeneous omics Data via Statistical Inference and Learning Techniques. In: *Genomics and Computational Biology* 2 (2016), Nr. 1, S. 32

[AF17] Ahmad, Ashar ; Fröhlich, Holger: Towards clinically more relevant dissection of patient heterogeneity via survival-based Bayesian clustering. In: *Bioinformatics* 33 (2017), Nr. 22, S. 3558–3566

[Ant74] Antoniak, Charles E.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. In: *The annals of statistics* (1974), S. 1152–1174

[BCH+12] Barillot, Emmanuel ; Calzone, Laurence ; Hupe, Philippe ; Vert, Jean-Philippe ; Zinovyev, Andrei: *Computational systems biology of cancer*. CRC Press, 2012

[BF13] Bouriga, Mathilde ; Féron, Olivier: Estimation of covariance matrices based on hierarchical inverse-Wishart priors. In: *Journal of Statistical Planning and Inference* 143 (2013), Nr. 4, S. 795–808

[BFM18] Beerenwinkel, Niko ; Fröhlich, Holger ; Murphy, Susan A.: Addressing the Computational Challenges of Personalized Medicine (Dagstuhl Seminar 17472). In: *Dagstuhl Reports* Bd. 7 Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018

[BH95] Benjamini, Yoav ; Hochberg, Yosef: Controlling the false discovery rate: a practical and powerful approach to multiple testing. In: *Journal of the royal statistical society. Series B (Methodological)* (1995), S. 289–300

[BIB15] Barcella, William ; Iorio, Maria ; Baio, Gianluca: A comparative review of variable selection techniques for covariate dependent Dirichlet process mixture models. In: *Canadian Journal of Statistics* (2015)

[BIBML16] Barcella, William ; Iorio, Maria D. ; Baio, Gianluca ; Malone-Lee, James: Variable selection in covariate dependent random partition models: an application to urinary tract infection. In: *Statistics in medicine* 35 (2016), Nr. 8, S. 1373–1389

[Bis06] Bishop, Christopher M.: *Pattern recognition and machine learning*. springer, 2006

[BKH+02] Beer, David G. ; Kardia, Sharon L. ; Huang, Chiang-Ching ; Giordano, Thomas J. ; Levin, Albert M. ; Misek, David E. ; Lin, Lin ; Chen, Guoan ; Gharib, Tarek G. ; Thomas, Dafydd G. u. a.: Gene-expression profiles predict survival of patients with lung adenocarcinoma. In: *Nature medicine* 8 (2002), Nr. 8, S. 816–824

[BLSK16] Bunte, Kerstin ; Leppaaho, Eemeli ; Saarinen, Inka ; Kaski, Samuel: Sparse group factor analysis for biclustering of multiple data sources. In: *Bioinformatics* 32 (2016), Nr. 16, S. 2457–2463. http://dx.doi.org/10.1093/bioinformatics/btw207. – DOI 10.1093/bioinformatics/btw207

[BM73] Blackwell, David ; MacQueen, James B.: Ferguson distributions via Pólya urn schemes. In: *The annals of statistics* (1973), S. 353–355

[BNJ03] Blei, David M. ; Ng, Andrew Y. ; Jordan, Michael I.: Latent dirichlet allocation. In: *Journal of machine Learning research* 3 (2003), Nr. Jan, S. 993–1022

[BPD08] Boulesteix, Anne-Laure ; Porzelius, Christine ; Daumer, Martin: Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. In: *Bioinformatics* 24 (2008), Januar, Nr. 15, S. 1698–1706. http://dx.doi.org/10.1093/bioinformatics/btn262. – DOI 10.1093/bioinformatics/btn262. – ISSN 1367–4803, 1460–2059

[Bre01] Breiman, Leo: Random forests. In: *Machine learning* 45 (2001), Nr. 1, S. 5–32

[BS09] Binder, Harald ; Schumacher, Martin: Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. In: *BMC Bioinformatics* 10 (2009), S. 18. http://dx.doi.org/10.1186/1471-2105-10-18. – DOI 10.1186/1471–2105–10–18

[BT04] Bair, Eric ; Tibshirani, Robert: Semi-supervised methods to predict patient survival from gene expression data. In: *PLoS Biol* 2 (2004), Nr. 4, S. e108

[BTB+09] Barbie, David A. ; Tamayo, Pablo ; Boehm, Jesse S. ; Kim, So Y. ; Moody, Susan E. ; Dunn, Ian F. ; Schinzel, Anna C. ; Sandy, Peter ; Meylan, Etienne ; Scholl, Claudia u. a.: Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. In: *Nature* 462 (2009), Nr. 7269, S. 108–112

[BVDG11] Bühlmann, Peter ; Van De Geer, Sara: *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011

[CD09] Chung, Yeonseung ; Dunson, David B.: Nonparametric Bayes conditional distribution modeling with variable selection. In: *Journal of the American Statistical Association* 104 (2009), Nr. 488, S. 1646–1660

[CF12a] Cun, Yupeng ; Fröhlich, Holger: Biomarker gene signature discovery integrating network knowledge. In: *Biology* 1 (2012), Nr. 1, S. 5–17

[CF12b] Cun, Yupeng ; Fröhlich, Holger: Biomarker Gene Signature Discovery Integrating Network Knowledge. In: *Biology* 1 (2012), Februar, Nr. 1, S. 5–17. http://dx.doi.org/10.3390/biology1010005. – DOI 10.3390/biology1010005

[CF13] Cun, Yupeng ; Fröhlich, Holger: Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics. In: *PLoS ONE* 8 (2013), September, Nr. 9, S. e73074. http://dx.doi.org/10.1371/journal.pone.0073074. – DOI 10.1371/journal.pone.0073074. – ISSN 1932–6203

[CGL+10] Cooper, Lee A. ; Gutman, David A. ; Long, Qi ; Johnson, Brent A. ; Cholleti, Sharath R. ; Kurc, Tahsin ; Saltz, Joel H. ; Brat, Daniel J. ; Moreno, Carlos S.: The proneural molecular signature is enriched in oligodendrogliomas and predicts improved survival among diffuse gliomas. In: *PloS one* 5 (2010), Nr. 9, S. e12548

[CKB+14] Chalise, Prabhakar ; Koestler, Devin C. ; Bimali, Milan ; Yu, Qing ; Fridley, Brooke L.: Integrative clustering methods for high-dimensional molecular data. In: *Translational cancer research* 3 (2014), Nr. 3, S. 202

[CP12] Choi, Hyungwon ; Pavelka, Norman: When One and One Gives More than Two: Challenges and Opportunities of Integrative Omics. In: *Frontiers in Genetics* 2 (2012), Januar. http://dx.doi.org/10.3389/fgene.2011.00105. – DOI 10.3389/fgene.2011.00105. – ISSN 1664–8021

[CV95] Cortes, Corinna ; Vapnik, Vladimir: Support vector machine. In: *Machine learning* 20 (1995), Nr. 3, S. 273–297

[CXHY13] Chen, Xiaojun ; Xu, Xiaofei ; Huang, Joshua Z. ; Ye, Yunming: TW-k-means: automated two-level variable weighting clustering algorithm for multiview data. In: *Knowledge and Data Engineering, IEEE Transactions on* 25 (2013), Nr. 4, S. 932–944. http://dx.doi.org/10.1109/tkde.2011.262. – DOI 10.1109/tkde.2011.262

[Dab05] Dabney, Alan R.: ClaNC: point-and-click software for classifying microarrays to nearest centroids. In: *Bioinformatics* 22 (2005), Nr. 1, S. 122–123

[DGO+09] Daemen, Anneleen ; Gevaert, Olivier ; Ojeda, Fabian ; Debucquoy, Annelies ; Suykens, Johan A. ; Sempoux, Christine ; Machiels, Jean-Pascal ; Haustermans, Karin ; Moor, Bart D.: A kernel-based integration of genome-wide data for clinical decision support. In: *Genome Medicine* 1 (2009), April, Nr. 4, S. 39. http://dx.doi.org/10.1186/gm39. – DOI 10.1186/gm39. – ISSN 1756–994X

[DH04] Ding, Chris ; He, Xiaofeng: K-means clustering via principal component analysis. In: *Proceedings of the twenty-first international conference on Machine learning* ACM, 2004, S. 29

[DHKW+08] Desmedt, Christine ; Haibe-Kains, Benjamin ; Wirapati, Pratyaksha ; Buyse, Marc ; Larsimont, Denis ; Bontempi, Gianluca ; Delorenzi, Mauro ; Piccart, Martine ; Sotiriou, Christos: Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. In: *Clinical cancer research* 14 (2008), Nr. 16, S. 5158–5165

[DRR+13] Doucette, Tiffany ; Rao, Ganesh ; Rao, Arvind ; Shen, Li ; Aldape, Kenneth ; Wei, Jun ; Dziurzynski, Kristine ; Gilbert, Mark ; Heimberger, Amy B.: Immune heterogeneity of glioblastoma subtypes: extrapolation from the cancer genome atlas. In: *Cancer immunology research* 1 (2013), Nr. 2, S. 112–122

[EK14] Ellis, Hayley P. ; Kurian, Kathreena M.: Biological rationale for the use of PPARγ agonists in glioblastoma. In: *Frontiers in oncology* 4 (2014), S. 52

[FBB+18] Fröhlich, Holger ; Balling, Rudi ; Beerenwinkel, Niko ; Kohlbacher, Oliver ; Kumar, Santosh ; Lengauer, Thomas ; Maathuis, Marloes H. ; Moreau, Yves ; Murphy, Susan A. ; Przytycka, Teresa M. u. a.: From hype to reality: data science enabling personalized medicine. In: *BMC medicine* 16 (2018), Nr. 1, S. 150

[Fer73] Ferguson, Thomas S.: A Bayesian analysis of some nonparametric problems. In: *The annals of statistics* (1973), S. 209–230

[FG06] Falcon, Seth ; Gentleman, Robert: Using GOstats to test gene lists for GO term association. In: *Bioinformatics* 23 (2006), Nr. 2, S. 257–258

[FHT10] Friedman, Jerome ; Hastie, Trevor ; Tibshirani, Rob: Regularization paths for generalized linear models via coordinate descent. In: *Journal of statistical software* 33 (2010), Nr. 1, S. 1

[FLZ+12] Fu, Alan ; Leaderer, Derek ; Zheng, Tongzhang ; Hoffman, Aaron E. ; Stevens, Richard G. ; Zhu, Yong: Genetic and epigenetic associations of circadian gene TIMELESS and breast cancer risk. In: *Molecular carcinogenesis* 51 (2012), Nr. 12, S. 923–929

[FSA99] Freund, Yoav ; Schapire, Robert ; Abe, Naoki: A short introduction to boosting. In: *Journal-Japanese Society For Artificial Intelligence* 14 (1999), Nr. 771-780, S. 1612

[GCBI04] Gautier, Laurent ; Cope, Leslie ; Bolstad, Benjamin M. ; Irizarry, Rafael A.: affy—analysis of Affymetrix GeneChip data at the probe level. In: *Bioinformatics* 20 (2004), Nr. 3, S. 307–315

[GCSR04] Gelman, A ; Carlin, J ; Stern, H ; Rubin, D: *Bayesian Data Analysis*. Boca Raton, Florida : Chapman & Hall/CRC, 2004

[GL08] Grun, Bettina ; Leisch, Friedrich: FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. (2008)

[GNP+06] Giovanna, MD ; Ng, Ken ; Paradiso, Lucy ; Godde, Nathan J. ; Kaye, Andrew ; Novak, Ulrike: ADAM22, expressed in normal brain but not in high-grade gliomas, inhibits cellular proliferation via the disintegrin domain. In: *Neurosurgery* 58 (2006), Nr. 1, S. 179–186

[GPF+11] Gade, Stephan ; Porzelius, Christine ; Faelth, Maria ; Brase, Jan ; Wuttig, Daniela ; Kuner, Ruprecht ; Binder, Harald ; Sueltmann, Holger ; Beissbarth, Tim: Graph based fusion of miRNA and mRNA expression data improves clinical outcome prediction in prostate cancer. In: *BMC Bioinformatics* 12 (2011), Nr. 1, S. 488. http://dx.doi.org/10.1186/1471-2105-12-488. – DOI 10.1186/1471–2105–12–488. – ISSN 1471–2105

[GR10] Görür, Dilan ; Rasmussen, Carl E.: Dirichlet process gaussian mixture models: Choice of the base distribution. In: *Journal of Computer Science and Technology* 25 (2010), Nr. 4, S. 653–664

[Gro01] Group, Biomarkers Definitions W.: Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. In: *Clinical Pharmacology and Therapeutics* 69 (2001), Nr. 3, 89–95. http://dx.doi.org/10.1067/mcp.2001.113989. – DOI 10.1067/mcp.2001.113989. – ISSN 1532–6535

[GRS+10] Glas, Martin ; Rath, Barbara H. ; Simon, Matthias ; Reinartz, Roman ; Schramme, Anja ; Trageser, Daniel ; Eisenreich, Ramona ; Leinhaas, Anke ; Keller, Mihaela ; Schildhaus, Hans-Ulrich u. a.: Residual tumor cells are unique cellular targets in glioblastoma. In: *Annals of neurology* 68 (2010), Nr. 2, S. 264–269

[GST+06] Gevaert, Olivier ; Smet, Frank D. ; Timmerman, Dirk ; Moreau, Yves ; Moor, Bart D.: Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. In: *Bioinformatics* 22 (2006), Juli, Nr. 14, S. e184–e190. http://dx.doi.org/10.1093/bioinformatics/btl230. – DOI 10.1093/bioinformatics/btl230. – ISSN 1367–4803, 1460–2059

[GW08] Ginsburg, Geoffrey S. ; Willard, Huntington F.: *Genomic and personalized medicine*. Bd. 1. Academic Press, 2008

[HAA+10] Hudson (Chairperson), Thomas J. ; Anderson, Warwick ; Aretz, Axel ; Barker, Anna D. ; Bell, Cindy ; Bernabé, Rosa R. ; Bhan, M. K. ; Calvo, Fabien ; Eerola, Iiro ; Gerhard, Daniela S. ; Guttmacher, Alan u. a.: International network of cancer genome projects. In: *Nature* 464 (2010), April, Nr. 7291, S. 993–998. http://dx.doi.org/10.1038/nature08987. – DOI 10.1038/nature08987. – ISSN 0028–0836, 1476–4687

[HBP11] Hannah, Lauren A. ; Blei, David M. ; Powell, Warren B.: Dirichlet process mixtures of generalized linear models. In: *The Journal of Machine Learning Research* 12 (2011), S. 1923–1953

[HBPO14] Hoffman, Y ; Bublik, DR ; Pilpel, Y ; Oren, M: miR-661 downregulates both Mdm2 and Mdm4 to activate p53. In: *Cell Death & Differentiation* 21 (2014), Nr. 2, S. 302–309

[HCP+82] Harrell, Frank E. ; Califf, Robert M. ; Pryor, David B. ; Lee, Kerry L. ; Rosati, Robert A.: Evaluating the yield of medical tests. In: *Jama* 247 (1982), Nr. 18, S. 2543–2546

[HHMW10] Hjort, Nils L. ; Holmes, Chris ; Müller, Peter ; Walker, Stephen G.: *Bayesian nonparametrics*. Bd. 28. Cambridge University Press, 2010

[HHR11] Hawkins, R. D. ; Hon, Gary C. ; Ren, Bing: Next-generation genomics: an integrative approach. In: *Nature Reviews Genetics* (2011), Januar. http://dx.doi.org/10.1038/nrg2795. – DOI 10.1038/nrg2795. – ISSN 1471–0056, 1471–0064

[HIP+03] Huang, Erich ; Ishida, Seiichi ; Pittman, Jennifer ; Dressman, Holly ; Bild, Andrea ; Kloos, Mark ; D'Amico, Mark ; Pestell, Richard G. ; West, Mike ; Nevins, Joseph R.: Gene expression phenotypic models that predict the activity of oncogenic pathways. In: *Nature Genetics* 34 (2003), Juni, Nr. 2, S. 226–230. http://dx.doi.org/10.1038/ng1167. – DOI 10.1038/ng1167. – ISSN 1061–4036

[HM78] Horwitz, KATHRYN B. ; McGuire, WL: Estrogen control of progesterone receptor in human breast cancer: correlation with nuclear processing of estrogen receptor. In: *Journal of Biological Chemistry* 253 (1978), Nr. 7, S. 2223–8

[Hot36] Hotelling, Harold: Relations between two sets of variates. In: *Biometrika* 28 (1936), Nr. 3/4, S. 321–377

[HPB+03] Han, Sehwan ; Park, Kyeongmee ; Bae, Byung-Noe ; Kim, Ki H. ; Kim, Hong-Joo ; Kim, Young-Duck ; Kim, Hong-Yong: E2F1 expression is related with the poor survival of lymph node-positive breast cancer patients treated with fluorouracil, doxorubicin and cyclophosphamide. In: *Breast cancer research and treatment* 82 (2003), Nr. 1, S. 11–16

[HTN+11] Hastie, Trevor ; Tibshirani, Robert ; Narasimhan, Balasubramanian ; Chu, Gilbert ; Narasimhan, Maintainer B. ; Bioinformatics, Microarray biocViews: Package 'impute'. (2011)

[IBC+03] Irizarry, Rafael A. ; Bolstad, Benjamin M. ; Collin, Francois ; Cope, Leslie M. ; Hobbs, Bridget ; Speed, Terence P.: Summaries of Affymetrix GeneChip probe level data. In: *Nucleic acids research* 31 (2003), Nr. 4, S. e15–e15

[Iri03] Irizarry, R A.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. In: *Biostatistics* 4 (2003), S. 249–264

[JJB12] Jensen, Peter B. ; Jensen, Lars J. ; Brunak, Søren: Mining electronic health records: towards better research applications and clinical care. In: *Nature Reviews Genetics* 13 (2012), Nr. 6, S. 395

[JJNH91] Jacobs, Robert A. ; Jordan, Michael I. ; Nowlan, Steven J. ; Hinton, Geoffrey E.: Adaptive mixtures of local experts. In: *Neural computation* 3 (1991), Nr. 1, S. 79–87

[JLR07] Johnson, W E. ; Li, Cheng ; Rabinovic, Ariel: Adjusting batch effects in microarray expression data using empirical Bayes methods. In: *Biostatistics* 8 (2007), Nr. 1, S. 118–127

[Joh67] Johnson, Stephen C.: Hierarchical clustering schemes. In: *Psychometrika* 32 (1967), Nr. 3, S. 241–254

[Jyx+12] Jie, Chen ; yong, Shen bai ; xing, Deng xia ; Qian, Zhan ; hong, Peng cheng: SKP1-CULLIN1-F-box (SCF)-mediated DRG2 degradation facilitated chemotherapeutic drugs induced apoptosis in hepatocellular carcinoma cells. In: *Biochemical and biophysical research communications* 420 (2012), Nr. 3, S. 651–655

[KBF+12] Kormaksson, Matthias ; Booth, James G. ; Figueroa, Maria E. ; Melnick, Ari u. a.: Integrative model-based clustering of microarray methylation and expression data. In: *The Annals of Applied Statistics* 6 (2012), Nr. 3, S. 1327–1347. http://dx.doi.org/10.1214/11-aoas533. – DOI 10.1214/11–aoas533

[KCV+03] Kleer, Celina G. ; Cao, Qi ; Varambally, Sooryanarayana ; Shen, Ronglai ; Ota, Ichiro ; Tomlins, Scott A. ; Ghosh, Debashis ; Sewalt, Richard G. ; Otte, Arie P. ; Hayes, Daniel F. u. a.: EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. In: *Proceedings of the National Academy of Sciences* 100 (2003), Nr. 20, S. 11606–11611

[KGH08] Kauffmann, Audrey ; Gentleman, Robert ; Huber, Wolfgang: arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. In: *Bioinformatics* 25 (2008), Nr. 3, S. 415–416

[KGM+02] Kurose, Keisuke ; Gilley, Kristie ; Matsumoto, Satoshi ; Watson, Peter H. ; Zhou, Xiao-Ping ; Eng, Charis: Frequent somatic mutations in PTEN and TP53 are mutually exclusive in the stroma of breast carcinomas. In: *Nature genetics* 32 (2002), Nr. 3, S. 355–357

[KGS+12] Kirk, Paul ; Griffin, Jim E. ; Savage, Richard S. ; Ghahramani, Zoubin ; Wild, David L.: Bayesian correlated clustering to integrate multiple datasets. In: *Bioinformatics* 28 (2012), Nr. 24, S. 3290–3297

[KLR+14] Kristensen, Vessela N. ; Lingjærde, Ole C. ; Russnes, Hege G. ; Vollan, Hans Kristian M. ; Frigessi, Arnoldo ; Børresen-Dale, Anne-Lise: Principles and methods of integrative genomic analyses in cancer. In: *Nature Reviews Cancer* 14 (2014), Mai, Nr. 5, S. 299–313. http://dx.doi.org/10.1038/nrc3721. – DOI 10.1038/nrc3721. – ISSN 1474–175X

[KMC+10] Koestler, Devin C. ; Marsit, Carmen J. ; Christensen, Brock C. ; Karagas, Margaret R. ; Bueno, Raphael ; Sugarbaker, David J. ; Kelsey, Karl T. ; Houseman, E A.: Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. In: *Bioinformatics* 26 (2010), Nr. 20, S. 2578–2585

[KVK13] Klami, Arto ; Virtanen, Seppo ; Kaski, Samuel: Bayesian canonical correlation analysis. In: *The Journal of Machine Learning Research* 14 (2013), Nr. 1, S. 965–1003

[KVLK15] Klami, Arto ; Virtanen, Seppo ; Leppaaho, Eemeli ; Kaski, Samuel: Group Factor Analysis. In: *Neural Networks and Learning Systems, IEEE Transactions on* 26 (2015), Nr. 9, S. 2136–2147

[LCB+04] Lanckriet, G ; Cristianini, N ; Bartlett, P ; Ghaoui, L E. ; Jordan, M: Learning the Kernel Matrix with Semidefinite Programming. In: *J. Machine Learning Research* 5 (2004), S. 27–72

[LCMM10] Lê Cao, Kim-Anh ; Meugnier, Emmanuelle ; McLachlan, Geoffrey J.: Integrative mixture of experts to combine clinical factors and gene markers. In: *Bioinformatics* 26 (2010), Nr. 9, S. 1192–1198

[LHA+13] Liverani, Silvia ; Hastie, David I. ; Azizi, Lamiae ; Papathomas, Michail ; Richardson, Sylvia: PReMiuM: an R package for profile regression mixture models using Dirichlet processes. In: *arXiv preprint arXiv:1303.2836* (2013)

[LHA+15] Liverani, Silvia ; Hastie, David I. ; Azizi, Lamiae ; Papathomas, Michail ; Richardson, Sylvia: PReMiuM: An R package for profile regression mixture models using Dirichlet processes. In: *Journal of statistical software* 64 (2015), Nr. 7, S. 1

[LLA+16] Lee, Yeri ; Lee, Jin-Ku ; Ahn, Sun H. ; Lee, Jeongwu ; Nam, Do-Hyun: WNT signaling in glioblastoma and therapeutic opportunities. In: *Laboratory Investigation* 96 (2016), Nr. 2, S. 137–150

[LLH+04] Lapointe, Jacques ; Li, Chunde ; Higgins, John P. ; Rijn, Matt van d. ; Bair, Eric ; Montgomery, Kelli ; Ferrari, Michelle ; Egevad, Lars ; Rayford, Walter ; Bergerheim, Ulf u. a.: Gene expression profiling identifies clinically relevant subtypes of prostate cancer. In: *Proceedings of the National Academy of Sciences of the United States of America* 101 (2004), Nr. 3, S. 811–816

[LLK+09] Lee, Jung T. ; Lee, Tae-Jin ; Kim, Cheol-Hee ; Kim, Nam-Soon ; Kwon, Taeg K.: Over-expression of Reticulon 3 (RTN3) enhances TRAIL-mediated apoptosis via up-regulation of death receptor 5 (DR5) and down-regulation of c-FLIP. In: *Cancer letters* 279 (2009), Nr. 2, S. 185–192

[LS99] Lee, Daniel D. ; Seung, H. S.: Learning the parts of objects by non-negative matrix factorization. In: *Nature* 401 (1999), Nr. 6755, S. 788–791

[LS01] Lee, Daniel D. ; Seung, H S.: Algorithms for non-negative matrix factorization. In: *Advances in neural information processing systems*, 2001, S. 556–562

[MCM+06] Moffat, Bradford A. ; Chenevert, Thomas L. ; Meyer, Charles R. ; Mckeever, Paul E. ; Hall, Daniel E. ; Hoff, Benjamin A. ; Johnson, Timothy D. ; Rehemtulla, Alnawaz ; Ross, Brian D.: The functional diffusion map: an imaging biomarker for the early prediction of cancer treatment outcome. In: *Neoplasia* 8 (2006), Nr. 4, S. 259–267

[MCM12] Melnykov, Volodymyr ; Chen, Wei-Chen ; Maitra, Ranjan: MixSim: an R package for simulating data to study performance of clustering algorithms. In: *Journal of Statistical Software* 51 (2012), Nr. 12, S. 1–25

[MEW96] Müller, Peter ; Erkanli, Alaattin ; West, Mike: Bayesian curve fitting using multivariate normal mixtures. In: *Biometrika* 83 (1996), Nr. 1, S. 67–79

[MFB+08] McLendon, Roger ; Friedman, Allan ; Bigner, Darrell ; Van Meir, Erwin G. ; Brat, Daniel J. ; M. Mastrogianakis, Gena ; Olson, Jeffrey J. u. a.: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. In: *Nature* 455 (2008), Oktober, Nr. 7216, S. 1061–1068. http://dx.doi.org/10.1038/nature07385. – DOI 10.1038/nature07385. – ISSN 0028–0836, 1476–4687

[MGKP08] Maragos, Petros ; Gros, Patrick ; Katsamanis, Athanassios ; Papandreou, George: Cross-Modal Integration for Performance Improving in Multimedia: A Review. In: Maragos, Petros (Hrsg.) ; Potamianos, Alexandros

(Hrsg.) ; Gros, Patrick (Hrsg.): *Multimodal Processing and Interaction*. Boston, MA : Springer US, 2008. – ISBN 978–0–387–76315–6 978–0–387–76316–3, S. 1–46

[MO04] Madeira, Sara C. ; Oliveira, Arlindo L.: Biclustering algorithms for biological data analysis: a survey. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 1 (2004), Nr. 1, S. 24–45. http://dx.doi.org/10.1109/tcbb.2004.2. – DOI 10.1109/tcbb.2004.2

[MO06] Meeds, Edward ; Osindero, Simon: An alternative infinite mixture of Gaussian process experts. In: *Advances in Neural Information Processing Systems*, 2006, S. 883–890

[MP00] McLachlan, Geoffrey ; Peel, David: Mixtures of factor analyzers. In: *Finite Mixture Models* (2000), S. 238–256

[MPJR10] Molitor, John ; Papathomas, Michail ; Jerrett, Michael ; Richardson, Sylvia: Bayesian profile regression with an application to the National Survey of Children's Health. In: *Biostatistics* 11 (2010), Nr. 3, S. 484–498

[MQ10] Müller, Peter ; Quintana, Fernando: Random partition models with regression on covariates. In: *Journal of statistical planning and inference* 140 (2010), Nr. 10, S. 2801–2808

[MQR08] Müller, Peter ; Quintana, Fernando ; Rosner, Gary: Bayesian clustering with regression / Working paper. 2008. – Forschungsbericht

[MQR11] Müller, Peter ; Quintana, Fernando ; Rosner, Gary L.: A product partition model with regression on covariates. In: *Journal of Computational and Graphical Statistics* 20 (2011), Nr. 1, S. 260–278

[MS02] Medvedovic, Mario ; Sivaganesan, Siva: Bayesian infinite mixture model based clustering of gene expression profiles. In: *Bioinformatics* 18 (2002), Nr. 9, S. 1194–1206

[MS17] Mathur, Sunil ; Sutton, Joseph: Personalized medicine could transform healthcare. In: *Biomedical reports* 7 (2017), Nr. 1, S. 3–5

[MTD12] Montanaro, Lorenzo ; Treré, Davide ; Derenzini, Massimo: Changes in ribosome biogenesis may induce cancer by down-regulating the cell tumor suppressor potential. In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1825 (2012), Nr. 1, S. 101–110

[MTMG03] Monti, Stefano ; Tamayo, Pablo ; Mesirov, Jill ; Golub, Todd: Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. In: *Machine learning* 52 (2003), Nr. 1, S. 91–118

[MYB04] Medvedovic, Mario ; Yeung, Ka Y. ; Bumgarner, Roger E.: Bayesian mixture model based clustering of replicated microarray data. In: *Bioinformatics* 20 (2004), Nr. 8, S. 1222–1232

[Nea00] Neal, Radford M.: Markov chain sampling methods for Dirichlet process mixture models. In: *Journal of computational and graphical statistics* 9 (2000), Nr. 2, S. 249–265

[OLX+17] Olmez, Inan ; Love, Shawn ; Xiao, Aizhen ; Manigat, Laryssa ; Randolph, Peyton ; McKenna, Brian D. ; Neal, Brian P. ; Boroda, Salome ; Li, Ming ; Brenneman, Breanna u. a.: Targeting the mesenchymal subtype in glioblastoma and other cancers via inhibition of diacylglycerol kinase alpha. In: *Neuro-oncology* (2017)

[PC08] Park, Trevor ; Casella, George: The bayesian lasso. In: *Journal of the American Statistical Association* 103 (2008), Nr. 482, S. 681–686

[PD10] Park, Ju-Hyun ; Dunson, David B.: Bayesian generalized product partition model. In: *Statistica Sinica* (2010), S. 1203–1226

[PHD+04] Pittman, Jennifer ; Huang, Erich ; Dressman, Holly ; Horng, Cheng-Fang ; Cheng, Skye H. ; Tsou, Mei-Hua ; Chen, Chii-Ming ; Bild, Andrea ; Iversen, Edwin S. ; Huang, Andrew T. ; Nevins, Joseph R. ; West, Mike: Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. In: *Proceedings of the National Academy of Sciences of the United States of America* 101 (2004), Januar, Nr. 22, S. 8431–8436. http://dx.doi.org/10.1073/pnas.0401736101. – DOI 10.1073/pnas.0401736101. – ISSN 0027–8424, 1091–6490

[PJZ+08] Parsons, D W. ; Jones, Siân ; Zhang, Xiaosong ; Lin, Jimmy Cheng-Ho ; Leary, Rebecca J. ; Angenendt, Philipp ; Mankoo, Parminder ; Carter, Hannah ; Siu, I-Mei ; Gallia, Gary L. u. a.: An integrated genomic analysis of human glioblastoma multiforme. In: *Science* 321 (2008), Nr. 5897, S. 1807–1812

[PKC+06] Phillips, Heidi S. ; Kharbanda, Samir ; Chen, Ruihuan ; Forrest, William F. ; Soriano, Robert H. ; Wu, Thomas D. ; Misra, Anjan ; Nigro, Janice M. ; Colman, Howard ; Soroceanu, Liliana u. a.: Molecular subclasses of

high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. In: *Cancer cell* 9 (2006), Nr. 3, S. 157–173

[PLi14] Peng, Yang ; Lin, Shiaw-Yih: TUSC4 functions as tumor suppressor by regulating BRCA1 stability and functions. In: *Cancer Research* 74 (2014), Nr. 19 Supplement, S. 1573–1573

[PSDW84] Parl, Fritz F. ; Schmidt, B P. ; Dupont, William D. ; Wagner, Rüdiger K: Prognostic significance of estrogen receptor status in breast cancer in relation to tumor stage, axillary node metastasis, and histopathologic grading. In: *Cancer* 54 (1984), Nr. 10, S. 2237–2242

[PWCG01] Pavlidis, P ; Weston, J ; Cai, J ; Grundy, W: Gene functional classification from heteregoneous data. In: *Proc. 5th Int. Conf. Computational Molecular Biology*, 2001, S. 242–248

[QMP15] Quintana, Fernando A. ; Müller, Peter ; Papoila, Ana L.: Cluster-Specific Variable Selection for Product Partition Models. In: *Scandinavian Journal of Statistics* 42 (2015), Nr. 4, S. 1065–1077

[QSL+15] Que, T ; Song, Y ; Liu, Z ; Zheng, S ; Long, H ; Li, Z ; Liu, Y ; Wang, G ; Zhou, J ; Zhang, X u. a.: Decreased miRNA-637 is an unfavorable prognosis marker and promotes glioma cell growth, migration and invasion via direct targeting Akt1. In: *Oncogene* (2015)

[Ras99] Rasmussen, Carl E.: The Infinite Gaussian Mixture Model. In: *NIPS* Bd. 12, 1999, S. 554–560

[Ras04] Rasmussen, Carl E.: The Infinite Gaussian Mixture Model. In: Solla, S. A. (Hrsg.) ; Leen, T. K. (Hrsg.) ; Müller, K.-R. (Hrsg.): *Advances in Neural Information Processing Systems 12*, 2004, S. 554–560

[RFW+10] Raman, Sudhir ; Fuchs, Thomas J. ; Wild, Peter J. ; Dahl, Edgar ; Buhmann, Joachim M. ; Roth, Volker: Infinite mixture-of-experts model for sparse survival regression with application to breast cancer. In: *BMC bioinformatics* 11 (2010), Nr. 8, S. 1

[RG02] Rasmussen, Carl E. ; Ghahramani, Zoubin: Infinite mixtures of Gaussian process experts. In: *Advances in neural information processing systems*, 2002, S. 881–888

[RK90] Rousseeuw, Peter J. ; Kaufman, L: Finding groups in data. In: *SeriesinProbability&MathematicalStatistics 1990 34 (1)* (1990), S. 111–112

[RKT+14] Ru, Yuanbin ; Kechris, Katerina J. ; Tabakoff, Boris ; Hoffman, Paula ; Radcliffe, Richard A. ; Bowler, Russell ; Mahaffey, Spencer ; Rossi, Simona ; Calin, George A. ; Bemis, Lynne u. a.: The multiMiR R package and database: integration of microRNA–target interactions along with their disease and drug associations. In: *Nucleic acids research* 42 (2014), Nr. 17, S. e133–e133

[Roy01] Royston, P: The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. In: *Statistica Neerlandica* 55 (2001), Nr. 1, S. 89–104

[RV11] Rudy, Jason ; Valafar, Faramarz: Empirical comparison of cross-platform normalization methods for gene expression data. In: *BMC bioinformatics* 12 (2011), Nr. 1, S. 467

[SBK13] Sun, Jiangwen ; Bi, Jinbo ; Kranzler, Henry R.: Multi-view biclustering for genotype-phenotype association studies of complex diseases. In: *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on* IEEE, 2013, S. 316–321

[SBK14] Sun, Jiangwen ; Bi, Jinbo ; Kranzler, Henry R.: Multi-view singular value decomposition for disease subtyping and genetic associations. In: *BMC genetics* 15 (2014), Nr. 1, S. 73. http://dx.doi.org/10.1186/1471-2156-15-73. – DOI 10.1186/1471–2156–15–73

[SFF+15] Serra, Angela ; Fratello, Michele ; Fortino, Vittorio ; Raiconi, Giancarlo ; Tagliaferri, Roberto ; Greco, Dario: MVDA: a multi-view genomic data integration methodology. In: *BMC bioinformatics* 16 (2015), Nr. 1, S. 1. http://dx.doi.org/10.1186/s12859-015-0680-3. – DOI 10.1186/s12859–015–0680–3

[SGG+13] Savage, Richard S. ; Ghahramani, Zoubin ; Griffin, Jim E. ; Kirk, Paul ; Wild, David L.: Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data. In: *arXiv preprint arXiv:1304.3577* (2013)

[SML15] Song, Qingxuan ; Merajver, Sofia D. ; Li, Jun Z.: Cancer classification in the genomic era: five contemporary problems. In: *Human genomics* 9 (2015), Nr. 1, S. 27

[SMO+03] Shannon, Paul ; Markiel, Andrew ; Ozier, Owen ; Baliga, Nitin S. ; Wang, Jonathan T. ; Ramage, Daniel ; Amin, Nada ; Schwikowski, Benno ; Ideker, Trey: Cytoscape: a software environment for integrated models of biomolecular interaction networks. In: *Genome research* 13 (2003), Nr. 11, S. 2498–2504

[SMS+12] Shen, Ronglai ; Mo, Qianxing ; Schultz, Nikolaus ; Seshan, Venkatraman E. ; Olshen, Adam B. ; Huse, Jason ; Ladanyi, Marc ; Sander, Chris: Integrative subtype discovery in glioblastoma using iCluster. In: *PloS one* 7 (2012), Nr. 4, S. e35236

[SN09] Shahbaba, Babak ; Neal, Radford: Nonlinear models using Dirichlet process mixtures. In: *The Journal of Machine Learning Research* 10 (2009), S. 1829–1850

[SNC77] Sanger, Frederick ; Nicklen, Steven ; Coulson, Alan R.: DNA sequencing with chain-terminating inhibitors. In: *Proceedings of the national academy of sciences* 74 (1977), Nr. 12, S. 5463–5467

[SOL09] Shen, Ronglai ; Olshen, Adam B. ; Ladanyi, Marc: Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. In: *Bioinformatics* 25 (2009), Nr. 22, S. 2906–2912. http://dx.doi.org/10.1093/bioinformatics/btp659. – DOI 10.1093/bioinformatics/btp659

[Som17] Somasundaram, Kumaravel: *Advances in Biology and Treatment of Glioblastoma*. Springer, 2017

[SPT+01] Sørlie, Therese ; Perou, Charles M. ; Tibshirani, Robert ; Aas, Turid ; Geisler, Stephanie ; Johnsen, Hilde ; Hastie, Trevor ; Eisen, Michael B. ; Rijn, Matt van d. ; Jeffrey, Stefanie S. u. a.: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. In: *Proceedings of the National Academy of Sciences* 98 (2001), Nr. 19, S. 10869–10874

[SRT+02] Shipp, Margaret A. ; Ross, Ken N. ; Tamayo, Pablo ; Weng, Andrew P. ; Kutok, Jeffery L. ; Aguiar, Ricardo C. ; Gaasenbeek, Michelle ; Angelo, Michael ; Reich, Michael ; Pinkus, Geraldine S. u. a.: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. In: *Nature medicine* 8 (2002), Nr. 1, S. 68–74

[SS02] Schölkopf, B ; Smola, A: Learning with kernels. In: *Cambridge: MIT Press. Schölkopf, B., Mika, S., Burges, C. J., P. Knirsch, K.-R. M., Rätsch, G., & Smola, A. J* (2002), S. –2000–81

[SSC+92] Sheikh, M S. ; Shao, Zhi-Ming ; Clemmons, David R. ; Leroith, Derek ; Roberts, Charles T. ; Fontana, Joseph A.: Identification of the insulin-like growth factor binding proteins 5 and 6 (IGFBP-5 and 6) in human breast cancer cells. In: *Biochemical and biophysical research communications* 183 (1992), Nr. 3, S. 1003–1010

[SSD+95] Schena, Mark ; Shalon, Dari ; Davis, Ronald W. ; Brown, Patrick O. u. a.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. In: *SCIENCE-NEW YORK THEN WASHINGTON-* (1995), S. 467–467

[STV06] Sha, Naijun ; Tadesse, Mahlet G. ; Vannucci, Marina: Bayesian variable selection for the analysis of microarray data with censored outcomes. In: *Bioinformatics* 22 (2006), Nr. 18, S. 2262–2268

[SWL+14] Shi, Yan ; Wang, Yingyi ; Luan, Wenkang ; Wang, Ping ; Tao, Tao ; Zhang, Junxia ; Qian, Jin ; Liu, Ning ; You, Yongping: Long non-coding RNA H19 promotes glioma cell invasion by deriving miR-675. In: *PLoS one* 9 (2014), Nr. 1, S. e86295

[SXF+12] Su, Dan ; Xu, Haiyan ; Feng, Jianguo ; Gao, Yun ; Gu, Linhui ; Ying, Lisha ; Katsaros, Dionyssios ; Yu, Herbert ; Xu, Shenhua ; Qi, Ming: PDCD6 is an independent predictor of progression free survival in epithelial ovarian cancer. In: *Journal of translational medicine* 10 (2012), Nr. 1, S. 1

[TB99] Tipping, Michael E. ; Bishop, Christopher M.: Probabilistic principal component analysis. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61 (1999), Nr. 3, S. 611–622. http://dx.doi.org/10.1162/089976699300016728. – DOI 10.1162/089976699300016728

[TBSM14] Thomas, Minta ; Brabanter, Kris D. ; Suykens, Johan A. ; Moor, Bart D.: Predicting breast cancer using an expression values weighted clinical classifier. In: *BMC Bioinformatics* 15 (2014), S. 411. http://dx.doi.org/10.1186/s12859-014-0411-1. – DOI 10.1186/s12859–014–0411–1. – ISSN 1471–2105

[Teh11] Teh, Yee W.: Dirichlet process. In: *Encyclopedia of machine learning*. Springer, 2011, S. 280–287

[TL15] Therneau, Terry M. ; Lumley, Thomas: Package 'survival'. In: *R Top Doc* 128 (2015)

[UJPK09] Ulrich, Theresa A. ; Juan Pardo, Elena M. ; Kumar, Sanjay: The mechanical rigidity of the extracellular matrix regulates the structure, motility, and proliferation of glioma cells. In: *Cancer research* 69 (2009), Nr. 10, S. 4167–4174

[vDv+02] van 't Veer, Laura J. ; Dai, Hongyue ; van de Vijver, Marc J. ; He, Yudong D. ; Hart, Augustinus A. M. ; Mao, Mao ; Peterse, Hans L. ; van der Kooy, Karin ; Marton, Matthew J. ; Witteveen, Anke T. ; Schreiber, George J. ; Kerkhoven, Ron M. ; Roberts, Chris ; Linsley, Peter S. ; Bernards, René ; Friend, Stephen H.: Gene expression profiling predicts clinical outcome of breast cancer. In: *Nature* 415 (2002), Januar, Nr. 6871, S. 530–536. http://dx.doi.org/10.1038/415530a. – DOI 10.1038/415530a

[VDVHV+02] Van De Vijver, Marc J. ; He, Yudong D. ; Veer, Laura J. ; Dai, Hongyue ; Hart, Augustinus A. ; Voskuil, Dorien W. ; Schreiber, George J. ; Peterse, Johannes L. ; Roberts, Chris ; Marton, Matthew J. u. a.: A gene-expression signature as a predictor of survival in breast cancer. In: *New England Journal of Medicine* 347 (2002), Nr. 25, S. 1999–2009

[VHP+10a] Verhaak, Roel G W. ; Hoadley, Katherine A. ; Purdom, Elizabeth ; Wang, Victoria ; Qi, Yuan ; Wilkerson, Matthew D. ; Miller, C R. ; Ding, Li ; Golub, Todd ; Mesirov, Jill P. ; Alexe, Gabriele ; Lawrence, Michael ; O'Kelly, Michael ; Tamayo, Pablo ; Weir, Barbara A. ; Gabriel, Stacey ; Winckler, Wendy ; Gupta, Supriya ; Jakkula, Lakshmi ; Feiler, Heidi S. ; Hodgson, J G. ; James, C D. ; Sarkaria, Jann N. ; Brennan, Cameron ; Kahn, Ari ; Spellman, Paul T. ; Wilson, Richard K. ; Speed, Terence P. ; Gray, Joe W. ; Meyerson, Matthew ; Getz, Gad ; Perou, Charles M. ; Hayes, D N. ; , Cancer Genome Atlas Research N.: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. In: *Cancer Cell* 17 (2010), Jan, Nr. 1, S. 98–110

[VHP+10b] Verhaak, Roel G. ; Hoadley, Katherine A. ; Purdom, Elizabeth ; Wang, Victoria ; Qi, Yuan ; Wilkerson, Matthew D. ; Miller, C R. ; Ding, Li ; Golub, Todd ; Mesirov, Jill P. u. a.: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. In: *Cancer cell* 17 (2010), Nr. 1, S. 98–110

[vHv+12] van Vliet, Martin H. ; Horlings, Hugo M. ; van de Vijver, Marc J. ; Reinders, Marcel J. T. ; Wessels, Lodewyk F. A.: Integration of Clinical and Gene Expression Data Has a Synergetic Effect on Predicting Breast Cancer Outcome. In:

PLoS ONE 7 (2012), Juli, Nr. 7, S. e40358. http://dx.doi.org/10.1371/journal.pone.0040358. – DOI 10.1371/journal.pone.0040358

[VVDVDV+02] Van't Veer, Laura J. ; Dai, Hongyue ; Van De Vijver, Marc J. ; He, Yudong D. ; Hart, Augustinus A. ; Mao, Mao ; Peterse, Hans L. ; Kooy, Karin van d. ; Marton, Matthew J. ; Witteveen, Anke T. u. a.: Gene expression profiling predicts clinical outcome of breast cancer. In: *nature* 415 (2002), Nr. 6871, S. 530–536

[VWC+15] Vassilakopoulou, Maria ; Won, Minhee ; Curran, Walter ; Souhami, Luis ; Prados, Michael ; Langer, Corey ; Rimm, David ; Hanna, Jason ; Neumeister, Veronique ; Smart, William ; Diaz, Aidnag ; Atkins, James ; Komarnicky, Lydia ; Schultz, Christopher ; Howard, Steven ; Dicker, Adam ; Knisely, Jonathan: GENO-21BRCA1 PROTEIN EXPRESSION PREDICTS SURVIVAL IN GLIOBLASTOMA PATIENTS FROM A NRG ONCOLOGY/RTOG COHORT. In: *Neuro-Oncology* 17 (2015), Nr. suppl 5, v96. http://dx.doi.org/10.1093/neuonc/nov215.21. – DOI 10.1093/neuonc/nov215.21

[WBM+13] Wang, Wenting ; Baladandayuthapani, Veerabhadran ; Morris, Jeffrey S. ; Broom, Bradley M. ; Manyam, Ganiraju ; Do, Kim-Anh: iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. In: *Bioinformatics* 29 (2013), Nr. 2, S. 149–159. http://dx.doi.org/10.1093/bioinformatics/bts655. – DOI 10.1093/bioinformatics/bts655

[WDPT14] Wade, Sara ; Dunson, David B. ; Petrone, Sonia ; Trippa, Lorenzo: Improving prediction from dirichlet process mixtures via enrichment. In: *Journal of Machine Learning Research* 15 (2014), Nr. 1, S. 1041–1071

[Wei92] Wei, LJ: The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. In: *Statistics in medicine* 11 (1992), Nr. 14-15, S. 1871–1879

[Wel05] Welling, Max: Fisher linear discriminant analysis. In: *Department of Computer Science, University of Toronto* 3 (2005), Nr. 1

[WFS+11] Waha, Anke ; Felsberg, Jörg ; Simon, Matthias ; Hartmann, Wolfgang ; Pietsch, Torsten ; Waha, Andreas: *A member of the SLC25 family is epigenetically inactivated in human gliomas and suppresses cell growth in vitro*. 2011

[WMD+14] Wang, Bo ; Mezlini, Aziz M. ; Demir, Feyyaz ; Fiume, Marc ; Tu, Zhuowen ; Brudno, Michael ; Haibe-Kains, Benjamin ; Goldenberg, Anna:

Similarity network fusion for aggregating data types on a genomic scale. In: *Nature methods* 11 (2014), Nr. 3, S. 333–337. http://dx.doi.org/10.1038/nmeth.2810. – DOI 10.1038/nmeth.2810

[Wol92] Wolpert, David H.: Stacked Generalization. In: *Neural Networks* 5 (1992), S. 241 – 259. http://dx.doi.org/10.1016/s0893-6080(05)80023-1. – DOI 10.1016/s0893–6080(05)80023–1

[WT12] Witten, Daniela M. ; Tibshirani, Robert: A framework for feature selection in clustering. In: *Journal of the American Statistical Association* (2012)

[WTH09] Witten, Daniela M. ; Tibshirani, Robert ; Hastie, Trevor: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. In: *Biostatistics* (2009), S. kxp008

[WYX+17] Wang, Xiuxing ; Yang, Kailin ; Xie, Qi ; Wu, Qiulian ; Mack, Stephen C. ; Shi, Yu ; Kim, Leo J. ; Prager, Briana C. ; Flavahan, William A. ; Liu, Xiaojing u. a.: Purine synthesis promotes maintenance of brain tumor initiating cells in glioma. In: *Nature neuroscience* 20 (2017), Nr. 5, S. 661

[XHIY17] Xue, Song ; Hu, Man ; Iyer, Veena ; Yu, Jinming: Blocking the PD-1/PD-L1 pathway in glioma: a potential new treatment strategy. In: *Journal of hematology & oncology* 10 (2017), Nr. 1, S. 81

[YFM+01] Yeung, Ka Y. ; Fraley, Chris ; Murua, Alejandro ; Raftery, Adrian E. ; Ruzzo, Walter L.: Model-based clustering and data transformations for gene expression data. In: *Bioinformatics* 17 (2001), Nr. 10, S. 977–987

[YH11] Yau, Christopher ; Holmes, Chris: Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. In: *Bayesian analysis (Online)* 6 (2011), Nr. 2, S. 329

[YSM11] Yuan, Yinyin ; Savage, Richard S. ; Markowetz, Florian: Patient-specific data fusion defines prognostic cancer subtypes. In: *PLoS Comput Biol* 7 (2011), Nr. 10, S. e1002227

[ZH05] Zou, Hui ; Hastie, Trevor: Regularization and variable selection via the elastic net. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2005), Nr. 2, S. 301–320

[ZHF+11] Zhang, Jin-fang ; He, Ming-liang ; Fu, Wei-ming ; Wang, Hua ; Chen, Lianzhou ; Zhu, Xiao ; Chen, Ying ; Xie, Dan ; Lai, Paul ; Chen, Gong u. a.: Primate-specific microRNA-637 inhibits tumorigenesis in hepatocellular carcinoma by disrupting signal transducer and activator of transcription 3 signaling. In: *Hepatology* 54 (2011), Nr. 6, S. 2137–2148

[ZLL+12] Zhang, Shihua ; Liu, Chun-Chi ; Li, Wenyuan ; Shen, Hui ; Laird, Peter W. ; Zhou, Xianghong J.: Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. In: *Nucleic acids research* (2012), S. gks725. http://dx.doi.org/10.1093/nar/gks725. – DOI 10.1093/nar/gks725

[ZZ14] Zitnik, Marinka ; Zupan, Blaz: Survival regression by data fusion. In: *Systems Biomedicine* 2 (2014), Nr. 3, S. 49–55. http://dx.doi.org/10.1080/21628130.2015.1016702. – DOI 10.1080/21628130.2015.1016702

[ZZ15] Zitnik, Marinka ; Zupan, Blaz: Data Fusion by Matrix Factorization. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (2015), Januar, Nr. 1, S. 41–53. http://dx.doi.org/10.1109/TPAMI.2014.2343973. – DOI 10.1109/TPAMI.2014.2343973. – ISSN 0162–8828, 2160–9292