Aus dem Institut für Medizinische Informatik und Biometrie - IMB
Direktor: Herr Prof. Dr. rer. med. Ingo Röder

CLASSIFICATION OF GLIOBLASTOMA MULTIFORME PATIENTS BASED ON AN INTEGRATIVE MULTI-LAYER FINITE MIXTURE MODEL SYSTEM

Dissertationsschrift

zur Erlangung des akademischen Grades

Doktor der Medizinischen Biometrie und Bioinformatik

Doctor rerum medicinalium (Dr. rer. medic.)

vorgelegt

der Medizinischen Fakultät Carl Gustav Carus

der Technischen Universität Dresden

von

Jaime Alberto Campos Valenzuela, Ing.

geboren am 09.03.1985 in Santiago de Chile

Dresden 2016

1. Gutachter: Prof. Dr. Lars Kaderali

2. Gutachter: Prof. Dr. med. Evelin Schröck

Tag der Einreichung: 05. Oktober 2016
Tag der Verteidigung: 23. August 2018

Dedicated to my family, who supported me from the distance.

Dedicated to my new family, who accompanied me all this time.

CONTENTS

I INTRODUCTION 1

1 MOTIVATION 3

2 INTRODUCTION 7

2.1 State of the art in Glioblastoma Multiforme research and integrative analysis . . . 7

2.1.1 Traditional classification of gliomas . . . 7

2.1.2 Glioblastoma multiforme profile and classification 8

2.1.3 State of integrative analysis . . . 11

2.1.4 Integrative analysis in Glioblastoma multiforme research . . . 14

2.2 Objectives . . . 17

2.3 Organization and structure of this thesis . . . 18

II METHODOLOGY AND RESULTS 19

3 METHODOLOGY 21

3.1 General methodology . . . 21

3.2 Reproducibility of results . . . 21

3.3 Data obtainment and preprocessing . . . 23

3.3.1 Types of data used . . . 23

3.3.2 Data sources . . . 23

3.3.3 Data Preprocessing . . . 26

3.4 Application of linear regression models . . . 32

3.4.1 Linear regression with cancer data . . . 32

3.4.2 Mixture of finite regression models and EM algorithm . . . 36

3.4.3 Penalization and biological prior knowledge . . . 38

3.4.4 Fitting of the model hyperparameters $\mathbf{K}$ and λ . . . 40

3.4.5 Issues in data integration . . . 41

3.5 Clustering of Glioblastoma patients . . . 41

3.5.1 Co-occurrence Probability . . . 41

3.5.2 Gene-models selection . . . 43

3.5.3 Definition of the co-occurrence probability matrix 43

3.5.4 Co-occurrence probability matrix as a distance matrix . . . 45

3.5.5 Clustering of patients . . . 45

3.6 Patients-wise analysis of the clusters . . . 45

3.6.1 Clinical analysis of the samples . . . 45

3.6.2 Comparison of the clusters to Verhaak's subtypes 47

3.6.3 Genetic signatures in the clusters . . . 47

3.7 Gene-models and features analysis of the clusters of patients . . . 47
3.7.1 Gene-models selection on each cluster of patients 48
3.7.2 Selected gene-models analysis . . . 48
3.7.3 Gene-models' features selection and analysis . . 49
4 MIXTURE OF REGRESSION MODELS: IMPLEMENTATION AND RESULTS 53
4.1 TCGA data analysis and construction of the gene-models 53
4.1.1 Number of models and covariates . . . 53
4.1.2 Source of the covariates . . . 55
4.2 Implementation of the MFLRMP algorithm and performance . . . 56
4.3 MFLRMP application, general results and analysis . . . 57
4.3.1 Set up and execution of the MFLRMP algorithm 58
4.3.2 General results obtained by the execution of the MFLRMP algorithm . . . 58
4.4 Summary of the implementation and execution results of the Mixture regression models algorithm . . . 66
5 BIOLOGICAL RESULTS AND ONCOLOGICAL IMPLICATIONS OF THE TRAINED MODELS 69
5.1 Analysis of co-occurrence and clustering of patients . . 69
5.1.1 Calculation of the co-occurrence value between samples . . . 69
5.1.2 Hierarchical clustering of the co-occurrence matrix 70
5.2 Patients-wise analysis of the clusters . . . 71
5.2.1 Covariates source for each cluster . . . 71
5.2.2 Clinical analysis of the clusters . . . 73
5.2.3 Comparison to Verhaak subtypes . . . 78
5.2.4 Genetic signatures in the clusters . . . 80
5.3 Gene-models and features analysis of the clusters of patients . . . 85
5.3.1 Gene-models selection for each cluster of patients 85
5.3.2 Analysis of shared gene-models . . . 87
5.3.3 Features analysis of the gene-models for each cluster of patients . . . 91
5.4 Summary of results . . . 97

III DISCUSSION AND CONCLUSION 101
6 DISCUSSION 103
7 CONCLUSION 111

IV SUMMARY 115

V APPENDIX 121
A DATA 123
A.1 Shared patients samples between dataset . . . 123

A.2 Features with the highest appearances ratios on each cluster 127
B IMAGES 129
B.1 Analysis of the silhouette values of the co-occurrence matrix 129
B.2 Heatmaps of co-occurrence distances for the gene-models grouping for the patients' clusters 129
B.3 Venn diagram of the selected gene-models for the patients clusters 131

BIBLIOGRAPHY 133

LIST OF FIGURES

Figure 1 General diagram of the origin and classification of glioblastomas . . . 10
Figure 2 General methodology of the project . . . 22
Figure 3 Main preprocessing steps for the TCGA datasets 31
Figure 4 Co-occurrence example . . . 42
Figure 5 Probability co-occurence curve . . . 44
Figure 6 Histogram of the number of covariates for the models . . . 54
Figure 7 Source of the models' covariates . . . 55
Figure 8 Performance comparison between different implementations of the FMLR algorithm . . . 57
Figure 9 Distribution of covariates by number of subpopulations . . . 59
Figure 10 Distribution of covariates by λ values . . . 60
Figure 11 Distribution of λ values for each number of subpopulations . . . 61
Figure 12 Distribution of fraction of covariates with non-zero coefficients by λ values . . . 62
Figure 13 Source of the models' covariates after penalization . . . 63
Figure 14 Fraction of the features with non-zero coefficients for each layer . . . 64
Figure 15 Comparison of number of features: original and penalized . . . 65
Figure 16 Boxplot of the coeffients α . . . 66
Figure 17 Histograms of co-occurrence values . . . 70
Figure 18 Heatmap of the co-occurrence matrix . . . 72
Figure 19 Distribution of features with non-zero coefficients by layer and cluster . . . 72
Figure 20 Distribution of the patients' age for each cluster 74
Figure 21 Male and female ratios for each cluster . . . 75
Figure 22 Karnofsky scores for each cluster . . . 77
Figure 23 Survival curves of the patients grouped by clusters . . . 78
Figure 24 Mutations' distribution for Cluster 5 and others . 83
Figure 25 Heatmap of the co-occurrence distances for Cluster 1 . . . 86
Figure 26 Heatmap of the co-occurrence distances for Cluster 5 . . . 86
Figure 27 Distribution of gene-models by patients' clusters 88

Figure 28 Number of appearances of non-zero covariants over the clusters 92
Figure 29 Distribution of silhouette values of the co-occurrence matrix 129
Figure 30 Heatmaps of co-occurrence distances for the gene-models 130
Figure 31 Venn diagram of the 6 clusters over their selected gene-models 131

LIST OF TABLES

Table 1 Main characteristics of the TCGA data 25
Table 2 Size of the raw interaction datasets 26
Table 3 Datasets after processing 30
Table 4 Interaction datasets after processing 30
Table 5 Number of elements for each cluster 71
Table 6 Missing samples in the clinical data 73
Table 7 Distribution of non-de novo and de novo samples for each cluster 75
Table 8 History of neoadjuvant treatment 77
Table 9 Number of samples for the Verhaak and Campos studies 79
Table 10 Contingency table shared samples 79
Table 11 Number of samples without mutation signature 81
Table 12 Number of samples with gene mutations 82
Table 13 Number of selected gene-models 87
Table 14 Top gene-models by number of appearances . 89
Table 15 Subpopulations of gene-models for each cluster 90
Table 16 Ratio of shared gene-models with same sub-models for each pair of clusters 91
Table 17 Appearance ratio for top features over clusters and biological layers 93
Table 18 Number of features with non-zero coefficients for each cluster 96
Table 19 Main characteristics of the discovered clusters . 100
Table 20 Set of shared patients between datasets 126

Table 21 Features with a significant number of appearances for each cluster 127

ACRONYMS

SNP Single Nucleotide Polymorphism

CNV Copy-number variations

GBM Glioblastoma Multiforme

PPI Protein-protein interaction

TCGA The Cancer Genome Atlas

BIC Bayesian information criterion

MFLRMP Mixture of Finite Linear Regression Models with Penalization

GO Gene Ontology

Part I

INTRODUCTION

1 MOTIVATION

GLIOBLASTOMA MULTIFORME RESEARCH AND TREATMENT

Glioblastoma Multiforme (GBM), or simply glioblastoma, is a very aggressive and invasive brain cancer. For most of the patients with this disease the prognosis after diagnose is only one year, which makes this disease one of the most frightening cancers. Additionally, because of its origin in the white matter of the brain is difficult to perform any type of intervention, analysis or treatment test without being disruptive in the life of the patient.

Over the years histology has been used as the main tool to classify gliomas. But this capacity is reduced in anaplastic tumors, where the cells are not differentiated. This is the case for GBM where its advance state of undifferentiation makes it difficult to analyze in a deeper manner.

Glioma: General term for tumors in the nervous system that arrise from glial cells

GENERAL PROBLEMS WITH INTEGRATIVE HIGH-THROUGHPUT DATA ANALYSIS IN CANCER

HIGH-THROUGHPUT METHODOLOGIES In the last decades there has been a strong development in the denominated high-throughput methodologies. These methodologies focus on the measurement of biological molecular profiles in a fast and large scale.

Different techniques and platforms have been developed to study and measure a plethora of molecular entities such as mRNA and microRNA expression, mutations and protein identification. This new set of techniques have allowed researches to analyze not only a large number of molecular elements in a fast manner, but also to study large populations of individuals.

Thanks to these methodologies a new era in biological analysis has begun, but with it, the number of issues have risen as well. These issues go from the development and implementation of computational methods used to process and store such amounts of data to the integration of these large sets of heterogeneous data.

The latter issue is the one this work focuses: the integration of highly heterogeneous data. In particular, we want to integrate this data to find information that is not possible to detect by analyzing the data in a stand-alone manner.

EFFECTS IN CANCER RESEARCH The possibility to study larges amounts of diverse data opens a new approach to the study and treatment of complex diseases such as cancer. These diseases are comprised of effects in a large number of different molecules and intermolecular interactions, such as genetic and epigenetic aberrations. While it is possible to study these different molecular elements separately, the results obtained won't necessarily explain the state of the disease as a system. This is a problem that can be solved by applying an integrative strategy and by doing so, a wider and deeper understanding of cancer and its mechanics can be gained.

Hypothesis: Cancer as a system disease

Additionally, thanks to this deeper understanding of the disease it would be possible to move from a general understanding of the disease into a personalized analysis for each patients or group of them. Which in turn could give rise to the development of personalized vaccines and other treatments.

The idea of a personalized understanding of the disease can be applied to the case of complex diseases phenotypes, where even in the case of indistinguishable phenotypes under the microscope, their molecular signatures and relationships are distinct between samples or groups and they can be segregated into different subtypes.

Hypothesis: Patients with complex diseases can be grouped by their molecular interactions

COMPLEXITY OF THE ANALYSIS The proposed analysis suffers from a double complexity. On one side, the search for a statistical model which allows the integration of heterogeneous biological data and on the other, the capability of such a model to segregate the samples used into different groups with unique and distinctive characteristics.

The heterogeneity of the molecular profiles makes it difficult to create a framework were all these different kinds of data can be aggregated and their particular effects analyzed, e. g. to compare the effect of methylation beta values with the effect of the overexpression of a microRNA.

Then, with the trained model and its results analyzed, the question about how to use this complex statistical model to study not only the relationships between the molecules, but how the patients are related to each other remains.

POSSIBLE IMPROVEMENTS IN CANCER RESEARCH The implementation and execution of this novel approach would benefit cancer research tremendously from two different, but complementary angles:

The first one is the focus in the interaction between molecules belonging to different molecular layers, these inter-layer relationships can explain effects that are not present in the analysis of a singular molecular profiles.

Molecular layers such as: gene expression, point mutations, methylation and others

The second angle is to use the discoveries of these new relationships (multi-layer interactions) in the cancer samples to classify patients by them and to find new targets for drugs and future treatments.

Additionally, we hypothesized that only a fraction of elements and relationships are necessarily to perform this classification and analysis. This is based on the idea that only a subset of genetic elements trigger these modifications or are affected by them and for the most elements no aberration is detected.

Hypothesis: Only subsets of molecular elements are relevant

2 INTRODUCTION

In this chapter, the state of the art in integrative cancer and GBM research will be presented. This comprises the use of different molecular and clinical data to study cancer and the application of machine learning algorithms to the analysis and classification of oncology data.

Then, the objectives of this project are presented and finally the organization of this thesis is explain.

2.1 STATE OF THE ART IN GLIOBLASTOMA MULTIFORME RESEARCH AND INTEGRATIVE ANALYSIS

2.1.1 *Traditional classification of gliomas*

Cancer is a major health issue in many parts of the world and is considered one of the most frightening diseases in our time. During the year 2015 in the United States of America it is the second leading cause of death with an expected number of deaths close to 600 000 (Siegel et al., 2015). These statistics have made cancer one of the main research targets, including new and stronger initiatives to deal with this disease, such as the newly presented National Cancer Moonshot Initiative[1] in the United States.

Even when malignant brain tumors represent a small portion of the total yearly number of deaths by cancer ($\sim 2{,}5\,\%$) (Siegel et al., 2015), their prognosis, invasiveness and negative repercussions in the patients' life made them a focus in cancer research. Around 80 % of malignant brain tumors correspond to malignant gliomas. Gliomas refer to a tumor that originated from the neoplastic glial cells located anywhere in the central nervous system, most generally in the brain (Q. T. Ostrom et al., 2014; Cloughesy et al., 2014).

Classification of gliomas has historically been a very difficult task, and it was not until 1920 when the so called *histological period* started (Scherer, 1940). In this period, the analysis of the anatomy of the cells using microscopy is used for the study and classification of gliomas.

Since this period, gliomas have been organized based on their resemblance to their presumed cells of origin, where their histological and immunohistochemical profiles are used for this task. Through this methodology the World Health Organization has segmented malignant gliomas into astrocytomas, oligodendrogliomas, mixed oligoastrocytomas and ependymomas (Goodenberger and Jenkins, 2012).

Histology: Study of the anatomy of cells and other minute structures in animals and plants.

Immunohistochemical: application of histochemical and immunologic methods to chemical analysis of living cells and tissues.

1 http://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative

In addition, the degree of aggressiveness, undifferentiation, anaplasia and necrosis of the malignant glioma is used for the grading of the tumor. The grade goes from I to IV, with grade I being tumors with low proliferative potential with the possibility of cure through surgical resection to grade IV, being the highest grade, tumors associated to fast evolution and fatal outcome, that show malignancy, an active reproduction and are necrosis-prone (Louis et al., 2007; Goodenberger and Jenkins, 2012).

As an example, oligodendrogliomas can be classified into grades II and III (low grade and anaplastic), while astrocytomas can be classified into grade II (low grade), III (anaplastic) and IV (primary and secondary gliomastomas) (Goodenberger and Jenkins, 2012).

In this work we will focus in particular in GBM. This is due to their already shown aggressiveness, but also because of their fast development and the significant portion of the total number of malignant gliomas that they represent. This and other characteristics are shown in detail in the next section.

2.1.2 *Glioblastoma multiforme profile and classification*

In order to analyze GBM, it is necessarily to present the main characteristics of these tumors. In particular, their development path, survival rate, genetic profiles and patients profile are detailed in this section.

DEVELOPMENT PATH: Glioblastoma Multiformes are classified as grade IV astrocytomas, but the path of their development is not unique. GBM development can follow two paths (Ohgaki, 2005; Goodenberger and Jenkins, 2012):

The first one is a de novo development, where there is small (even none) clinical or histological evidence of development from a lower grade tumor, and a very fast development (< 6 months) which are present at the time of diagnosis as full-blown tumors.

In the second path, GBMs can be developed in slow fashion from previously diagnosed lower grade astrocytomas (low grade and anaplastic).

The former variety is called Primary Glioblastoma Multiforme, while the latter is denominated Secondary Glioblastoma Multiforme. These two types of GBMs are the main subdivision of GBM tumors and any new classification method should be compared to them.

SURVIVAL RATE: Different types of gliomas have distinctive survival rates, which is related to their grade. This is possible to observe in the 5-year survival rate for each astrocytoma, where for low grade (diffusive) astrocytomas, anaplastic astrocytomas and GBMs (including

primary and secundary GBMs) it is $\sim 48\,\%$, $\sim 28\,\%$ and $\sim 5\,\%$ respectively (Quinn T Ostrom et al., 2015).

In addition, the median survival for secondary GBMs is 7,8 months, which has been shown to be significant different compared to the median survival for primary GBMs which is 4,7 months (Ohgaki, 2005).

GENETIC ABERRATIONS: During the last decades, the use of genetic studies have allowed the analysis of the different GBMs by their genetic profiles. Thanks to this analysis is it possible to see the contrast between primary and secondary GBMs, and the accumulation of aberrations during the development of secondary GBMs.

For the comparison between primary and secondy GBMs several differences between their genetic profiles have been discovered. The most significant ones, considering difference in incidence ratio, are shown here. In the first place, the LOH in chromosome 10 is the most frequent genetic aberration in GBMs. This alteration occurs in 60-80 % of the cases and is found in both GBMs (Ohgaki, 2005). By analyzing this LOH with more detail it is found that the LOH of the chromosomal arm 10q has a similar frequency of incidence in both primary and secondary GBMs, but the LOH for 10p is found mostly in primary GBMs. Additionally, mutations in the PTEN gene and amplifications of EGFR have been found mostly exclusively in primary glioblastomas, with rates of 25 % and 36 % respectively (Ohgaki, 2005; Ohgaki and Kleihues, 2007).

LOH: Loss of heterozygosity. Loss of one of the allele in heterozygous somatic cell.

On the other side, mutations in the gene TP53 are mostly found in secondary GBMs (65 % versus 28 % (Ohgaki and Kleihues, 2007)). Furthermore, this aberration is found in the precursor cells of secondary glioblastomas (low grade and anaplastic astrocytomas) in a rate $> 50\,\%$ and are the first aberration detected in them. LOH in the arm 19q also appears in a much larger rate in secondary GBMs ($> 50\,\%$) than in primary ($< 5\,\%$).

In a similar way, mutation in the gene IDH1 have been observed in around 70 % of the grade II and III astrocytomas (low grade and anaplastic) and other gliomas such as oligodendrogliomas. This high incidence is also present in secondary glioblastomas ($> 70\,\%$), but has a smaller presence in patients with primary glioblastomas ($\sim 5\,\%$), and because of this, this aberration has the potential of being a very specific marker for secondary GBMs (Ohgaki and Kleihues, 2013; Agnihotri et al., 2013).

In the top of Figure 1, the different development paths of GBMs are shown. The 3 different paths and the main genetic difference between them cam be seen.

AGE DISTRIBUTION: The age of the patients diagnosed with GBM differs between the two types. For patients with primary GBMs the

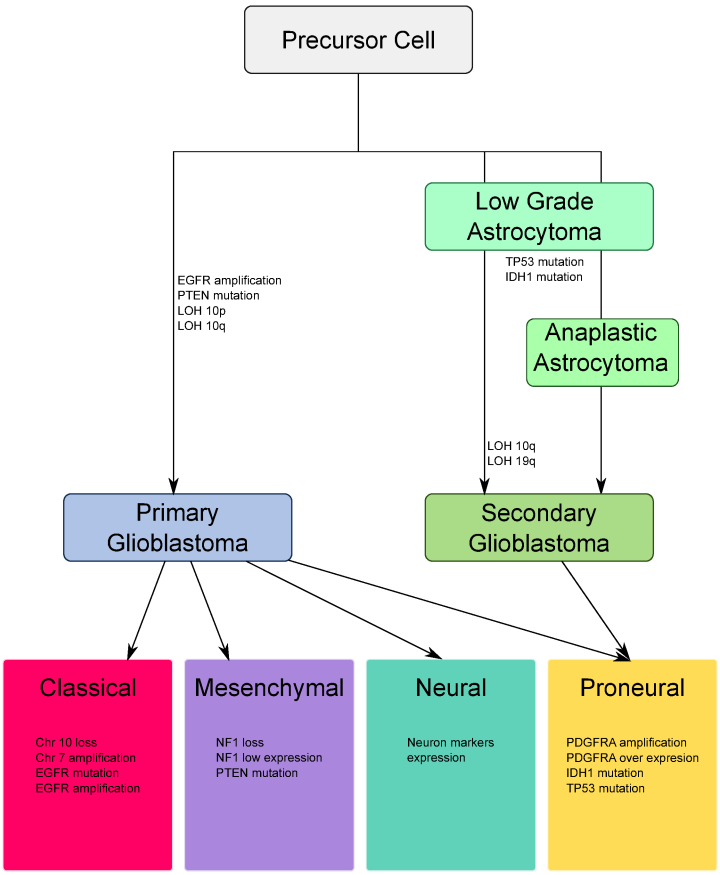


Figure 1: General diagram of the origin and classification of glioblastomas and their characteristics.
Top: Relationship between astrocytomas and GBMs and the main genetic variations found.
Bottom: Subtypes found by Verhaak et al. and their main characteristics.

mean age is 62 years old, while for secondary GBMs is 45 years old (Ohgaki and Kleihues, 2007).

CNS: Central Nervous System

INCIDENCE: Considering primary brain and CNS gliomas GBM accounts for over 55 % of the total number of cases, while low grade diffuse astrocytoma accounts for over 8,6 % and anaplastic astrocytoma over 6 % of the cases. Considering all the primary brain and tumors in the CNS, GBM accounts for 27 % of all of these tumors and 80 % of the malignant tumors (Quinn T Ostrom et al., 2015).

The fraction of tumors being cataloged as primary GBM is around 90 % of the total number of GBM cases. Making de novo GBMs much more common than secondary GBMs (Ohgaki and Kleihues, 2013; X Zhang et al., 2012).

2.1.3 *State of integrative analysis*

The genetic alterations presented before were obtained from studies that analyze one type of data at the time. The data can be gene expression signatures measured with microarrays, copy number variation measured with aCGH, networks such as pathways and others. In the last decade, an integrative approach has been applied to the study of complex diseases. In this type of studies, several types of data, also called 'omics' or layers, are integrated into large-scale multidimensional dataset and analyzed altogether.

aCGH: Array Comparative Genomic Hybridization

A novel and advance example of this approach is the PanCancer initiative, where since 2013 it has analyzed cancer samples in a double integrative approach. On one side, the integration of multiple data layers and on the other, the use of samples from multiple cancer types (The Cancer Genome Atlas Research Network et al., 2013), in order to look for molecular signatures across different tissues.

Integrative analysis has the following three main objectives (Kristensen et al., 2014). Each method can be focused to one or more of these:

RELATIONSHIP ANALYSIS: Study of the mechanisms and relationships between the different elements of a molecular dataset or between elements of different sets, e. g. mRNA-mRNA or mRNA-microRNA interactions.

PATIENTS SEGMENTATION: Use of the integration methods to separate the different samples into subtypes which are profiled by their high-dimensional molecular signatures.

CLINICAL OUTCOME: Application of the integrative methods to predict the clinical outcome (survival or efficacy of therapy) to patients.

Here, we present several kinds of methods for integration analysis. We focus on the methodology and not on the biological results of these implementations, which are going to be discussed in the following section. These methods can be grouped into different categories, depending on the core methodology used to integrate the multidimensional data (Kristensen et al., 2014):

SEQUENCIAL ANALYSIS This type of studies takes several different types of datasets and analyzed them in sequence. This means that the statistical method used takes as input only one of the molecular sets and after the analysis the results are integrated.

One common type of sequential analysis is the integration of Copy-number variations (CNV) and gene expression data. This is due to the importance of somatic CNVs in the alteration of the expression of key

genes, such as oncogenes and tumor suppressor genes (Huang et al., 2012). Additionally, this methodology allows to separate possible genes that have a fundamental role in the disease (driver genes) and those that may not be involved but present some sort of alteration (passenger genes) (Louhimo et al., 2012).

One step further in this kind of analysis can be seen in the study by Sun et al. (Sun et al., 2011), where in addition to the gene expression profiles and the CNV data, it has added methylation data to the analysis. In this work, an extensive analysis of each molecular profile was applied independently, which included clinical validation. After this extensive process, the correlation between the expression fold change and the methylation was obtained. Using a similar approach, the relationship between expression and CNV was calculated.

GENE SET ANALYSIS One integration scheme that is widely used in genomic studies as complementary analysis is the gene set enrichment analysis. Generally, this method is applied in the latter stages of the analysis in order to obtain broader results from the initial study.

The general approach for this family of methods is to integrate a set of genes with a calculated statistic (e. g. p-value) with a known set of genes, such as biological pathways or gene ontologies. Then the enrichment of the known gene set is obtained using the statistical of the genes. This is performed using different methodologies, which vary in the sophistication degree of the integration scheme. Here we present the two most used methods (Abatangelo et al., 2009) and one novel approach that integrates topological information:

- The simplest method is the use of a contingency table between the studied genes against the known set of genes. A threshold is defined for the calculated statistic to separate the set of studied genes and fill the table with those in the known gene set and those that are not. Finally, a Fisher's exact test is applied to the table to study the statistical significance of it (Abatangelo et al., 2009).

GSEA: Gene Set Enrichment Analysis

- Instead of defining a threshold for the statistic, the GSEA algorithm (Subramanian et al., 2005) uses all the values to implement a ranked list of the studied genes based on their statistic. This rank is compared to the known gene set in order to see if the genes belonging to the known set are at the beginning or end of the ranked list. A score is calculated based on the position in the list of genes of interest. Additional analyses are performed in order to obtain a random distribution of the score, to test its significance, and to take into account multiple hypothesis testing.

SPIA: Signaling Pathway Impact Analysis

- The topological information of the gene set is used in algorithms such as SPIA (Tarca et al., 2009) and PathOlogist (Greenblum

et al., 2011), where the differential expression of genes are used to calculate the enrichment of the pathway (perturbation and activity for each respective tool) and considering the different relationships in the structure. This analysis can be applied along clinical and phenotypical information.

PENALIZED REGRESSION ANALYSIS The integration of different types of data is possible in a regression model, where the target variable is an measured outcome (phenotypical (Zhu and Hastie, 2004; Shen and Tan, 2005), genomic (Peng et al., 2010) or other) and the covariates of the model is a mixture of different molecular signatures. To perform this analysis the variables must be standardized (Kristensen et al., 2014), and due to the large number of element on each 'omic' layer there is a need to shrink the number of variables. This is possible to achieve using penalization terms, such as *lasso* or elastic net. This approach allows the model to select the set of covariates that explains the target variable.

Elastic net: penalization method that uses two penalization terms: L_1 *and* L_2

This method is used as a base in this work and a detailed description is presented in the METHODOLOGY chapter.

NETWORK-BASED ANALYSIS Studies based on networks utilize graphs to model the interaction of the different elements of the multidimensional dataset. In general, each node in the network is an element and every edge is a link between elements. The different algorithms varies greatly in how to model the network, calculate the connections and analyze the resulting graph (Cho et al., 2012).

In Chuang et al. (Chuang et al., 2007), a network is constructed based on Protein-protein interactions (PPIs) and gene expression profiles are used to compare possible subnetworks in the graph based on the differential expression of the nodes. In this case, the edges and nodes were given by previous knowledge (PPIs) but the methodology to analyze the network was novel and allowed to find possible subnetworks as biomarkers of breast cancer.

Instead of using differential data over a known structure, another approach is to calculate the correlation between nodes and use this value to select those interactions that are more significant. Xue et al. (Xue et al., 2007) used this approach to study the aging process, where the correlation between the expression profiles of genes is calculated and then used to select the edges of the PPI network that are most significant. Finally utilizing hierarchical clustering it is able to select subnetworks that are most related to aging.

Thanks to its ability to model multiple types of data the use of graphs is widely used in integrative analyses and new approaches appear constantly, presenting novel methods to model the data and to study it.

2.1.4 *Integrative analysis in Glioblastoma multiforme research*

The use of integrative schemes to analyze complex diseases and phenomenas such as cancer or aging has allowed for the discovery of putative treatment targets, which can be single genes or modules of interacting elements.

As with other complex diseases, there has been an interest in analyzing the molecular complexity of GBM using integrative approaches. The impact of these analyses have been very significant and have changed the way GBMs are studied. The main projects and results are going to be shown here, focusing on their methodology as well as in their biological findings.

2.1.4.1 *First large multi-omic study by The Cancer Genome Atlas (TCGA)*

The first integrative high-impact project was published in 2008 by the The Cancer Genome Atlas Research Network (The Cancer Genome Atlas Research Network, 2008). In this paper, the CNV, the DNA methylation and the gene expression profiles for 206 primary GBM samples, as well as Single Nucleotide Polymorphism (SNP) data for a subset of 91 samples, were integrated in a sequential manner. From the samples with SNP data, a subset of them presented hypermutated profiles, which had a history of treatment for the disease.

With the combination of CNV, expression and SNP, it was found that the NF1 gene presents an erratic profile, with a correlation between genetic aberrations and the expression level of the gene, but this was not consistent in all the samples.

Using the same data, genetic alterations and variation in their expression were found in two members of the ErbB family (EGFR and ERBB2). While genetic aberrations in EGFR have been found previously in primary GBMs, according to this work mutations in ERBB2 had been reported only in one previous case. Members of the PI3K family were also found to have mutations: PIK3CA mutations have been already been linked to GBM, and PIK3R1 mutations, which are not common particularly in GBM cases.

Additionally, the CpG islands methylation analysis showed a methylation of the MGMT promoter, which codes for a DNA repair enzyme. A positive correlation was found in samples with the hypermutated phenotype and this methylation pattern.

Finally, the validated CNVs and SNPs were mapped to biological pathway, finding an enrichment in the p53 and RB tumor suppressor pathway, and the RTK signaling. The number of core elements in the pathways with aberrations was over 59 % for CNV and over 78 % for mutations in the all three pathways.

2.1.4.2 *Glioblastomas subtypes from Verhaak et al.*

The second work to be shown is the study by Verhaak et al. (Verhaak et al., 2010) in collaboration with the TCGA. This paper uses the profiles of gene expression, CpG island methylation and genomic aberrations (CNVs and SNPs) of 200 GBM samples and 2 normal brain samples, and it follows a sequential analysis of the multidimensional information.

The main result of this work is the definition of 4 subtypes of GBM patients, based on the gene expression profiles. These subtypes were obtained using the hierarchical clustering algorithm over the expression data and through additional analyses 210 genes for each subtypes were selected that were considered most representative.

The subtypes were named after their signature genes and resemblance to previous coined terms. In addition to the gene expression analysis, a genomic aberration study was performed for each subtype:

CLASSICAL This subtype is characterized by amplifications in chromosome 7 and loss in chromosome 10, which are common in GBM samples, but it was found in all the classical samples. EGFR showed a statistical significant amplification and over-expression compared to the other subtypes, considering that these characteristics are common in GBM. In addition, mutations in the TP53 gene were not found in this subtype. Finally, in samples of the classical subtype there is a correlation between the amplification of the EGFR gene and the deletion of CDKN2A, which is part of the RB pathway.

MESENCHYMAL Deletion and low expression of the NF1 gene are hallmarks of this subtype. Additionally, 70 % of the samples with mutations in the NF1 gene belong to this subtype. Comutations of the NF1 and PTEN genes were observed in this subtype, affecting the AKT pathway. This subtype is named after the expression of Mesenchymal markers (e. g. CHI3L1 and MET). Genes in the necrosis factor super family pathway and NF-κB pathway present over-expression.

PRONEURAL The alterations of the gene PDGFRA are one of the major features of this subtype. Focal amplifications of this gene in conjunction with the over-expression of it, were found almost exclusively in this subtype. Another major feature is the mutation of the IDH1 gene, where 11 of the 12 samples with these mutations belong to this subtype. LOH and mutations of TP53 were frequent, but the common amplification of the chromosome 7 and loss of chromosome 10 were less prevalent in this subtype. This subtype is named after the presence of genes related to the process of development such as DCX, DLL3, ASCL1 and TCF4.

NEURAL This subtypes is typified by the expression of the genes NEFL, GABRA1, SYT1 and SLC12A5, which are neuron markers. It was found that in the majority of samples belonging to this subtype, normal cells were present.

In addition to these genetic characteristics, clinical correlations with the subtypes were found. The majority of samples of secondary GBM (3 of 4) were classified as Proneural. While recurrent tumors were allocated in all subtypes, it was also found that the most consistent association was between the patients' age and the subtype, where younger patients were allocated in the Proneural subtype.

2.1.4.3 *Methylatator phenotype in Gliomas*

In 2010 a paper by Noushmehr et al. and the TCGA Research Network focused in the CpG island methylation profile of 272 GBM samples (Noushmehr et al., 2010). By using clustering algorithms, a set of 24 samples were found with a distinctive profile of methylation at a subset of loci. Due to this profile, the phenotype was named CpG island methylator phenotype (G-CIMP).

Due to the use of the same samples as in Verhaak et al., it was possible to correlate the G-CIMP with the previously discovered subtypes. Almost 90 % of the G-CIMP samples were found to belong to the Proneural subtype and the G-CIMP samples represented 30 % of the total samples allocated previously to this subtype. Finally, a relationship was found between the G-CIMP and the mutation of the IDH1 gene, considering not only secondary but all the tumors where IDH1 mutations are more common.

2.1.4.4 *Increase in number of samples and data for GBM analysis*

The fourth work by the TCGA Research Network for GBM analysis was published by Brennan et al. in 2013 (Brennan et al., 2013). In this study over 500 GBM samples were used for a sequential analysis. While the same 'omics' layers were used (gene expression, genomic variations (CNV and SNP) and CpG sites methilation), this work had paired samples for all the different layers and used array-based and NGS platforms.

NGS: Next-Generation Sequencing

It must be noted that the TCGA selects mostly primary GBM and due to that, genomic features that are present most commonly in secondary GBMs are not recurrent in this set of samples, such as IDH1 mutations. EGFR was found to be one of the most mutated genes with most of the cases found along an genomic amplification, additionally a plethora of altered transcripts were found for this gene. Mutual exclusivity alterations affecting the p53, Rb and PI3K pathways were confirmed.

2.1.4.5 *Additional analyses*

The use of the subtypes has allowed the analysis of GBM from a new angle. In Setty et al. (Setty et al., 2012), a linear model is trained for each gene in a subtype using the CNV of the gene, the expression of related microRNAs and the methylation of the promoter of that gene as covariates. With this methodology, the gene expression can be predicted for genes inside a subtype and by analyzing the most significant features of the linear models on each subtype, it was possible to select certain feature as main regulators.

In Savage et al. (Savage et al., 2013), the combination of gene and microRNA expression, CNV and methylation data were integrated using an extension of a Dirichlet Process mixture model.

Through this methodology it was possible to cluster each dataset of 'omic' information and integrate them into a consensus cluster. Its main result was the difficulty to integrate the different layers through consensus, due to the small overlap between layer-wise clusters, which they concluded was due to the complex nature of the disease and the relationship between the biological layers, which makes it difficult to define the straightforward subtypes.

.

2.2 OBJECTIVES

This project has as its main objective the development of a novel framework for the integration and analysis of heterogeneous molecular data of cancer patients and the detection of subgroups of patients in the data.

The main objective is based on the two main hypothesis of this work. The first one is that the aggregation of diverse biological data allows to study complex diseases as systems and the second one is that in complex diseases, the patients can be grouped by their intermolecular interactions, which is not possible by traditional histological methods.

Additionally, the integration of the multidimensional data must result in a interpretable model and that the model used must be able to cope with a dataset with a much larger number of variables than data points.

These requirements add several layers of complexity to our framework, with particular objectives for each one them:

DATA OBTAINMENT AND PREPROCESSING A model as complex as the one proposed here requires a large number of samples and the capacity to integrate them into the model. This necessity implies the selection of experimental platforms that maximized the number of samples and data with the specific characteristics (e. g. non sparse).

MODEL DEFINITION AND IMPLEMENTATION A base model that allows to implement the proposed framework must be found and modified in order to integrate the heterogeneous and high-dimensional biological data, along with the capability to cluster the samples used.

MODEL PERFORMANCE OPTIMIZATION The complexity and size of our data means that the time needed to apply our framework can take very long periods of time, in the magnitude of months, thus the performance optimization is critical in this work

CLUSTERING OF PATIENTS After the application of the algorithm, a methodology must be defined to use the resulting models to cluster the patients.

BIOLOGICAL IMPLICATIONS The final particular objective is the analysis of the groups of patients defined previously. The study comprises the analysis of the patients in each cluster, and the models and features that are significant in that cluster.

2.3 ORGANIZATION AND STRUCTURE OF THIS THESIS

This document is structured in 3 Parts and 7 chapters, each one with a specific goal.

Part i, entitled Introduction, contains the first two chapters MOTIVATION and INTRODUCTION.

In the first chapter the main motivations for this work are presented from a medical and statistical point of view, detailing the needs and voids in the current state of research. In Chapter 2, we find the general introduction to our work, presenting the state of the art in cancer research and integrative analysis, the main and particular objectives of this work and finally, the organization is located at the end of this first chapter.

The second part, Methodology and Results, contains three chapters, which can be considered the core of this project.

Chapter 3 refers to the methodology used and defined by this project, which goes from the theoretical background and implementation of the algorithm to the method of analysis of its results. Furthermore, Chapter 4 and Chapter 5 are comprised of the results obtained in this work. In the first chapter, the results concerning the implementation and execution of the algorithm, along with the general characteristics of the models obtained are shown. The second chapter groups the biological implication and analysis of the resulting models.

The last part, Discussion and Conclusion, is comprised by the discussions and conclusions of this work.

Finally, an Appendix is included. In this part, the data and images not included in the main text are shown.

Part II

METHODOLOGY AND RESULTS

3 METHODOLOGY

In this chapter, the methodology developed and used in this project is presented: firstly, a general representation of the different processes involved in this work, and later, all the particular methods and their implementation are described.

The main corpus of this chapter, the detailed description of the methods, is divided into 3 different sections, one for each of the main phases in the pipeline.

These steps comprise the whole process of the project and are intended to be completely reproducible. Because of this, several elements can be found in open repositories under open licenses as explained in Section 3.2.

3.1 GENERAL METHODOLOGY

This project is divided into 3 distinctive but connected phases, these are:

DATA OBTAINMENT AND PREPROCESSING: In this step GBM data from the TCGA repository was downloaded and preprocessed in order to ensure the correct application of the downstream processes.

APPLICATION OF THE MIXTURE REGRESSION MODEL: The algorithm is implemented and applied over the data previously preprocessed.

RESULTS ANALYSIS AND BIOLOGICAL IMPLICATIONS: The resulting models are analyzed and the results are put in a biological context.

In Figure 2 the general process is presented where is possible to observe the 3 main sections of the methodology and how they interact to shape the project's pipeline.

3.2 REPRODUCIBILITY OF RESULTS

For this project, the main scripts and the C++ package developed are available in the online repository for this project: Thesis repository.

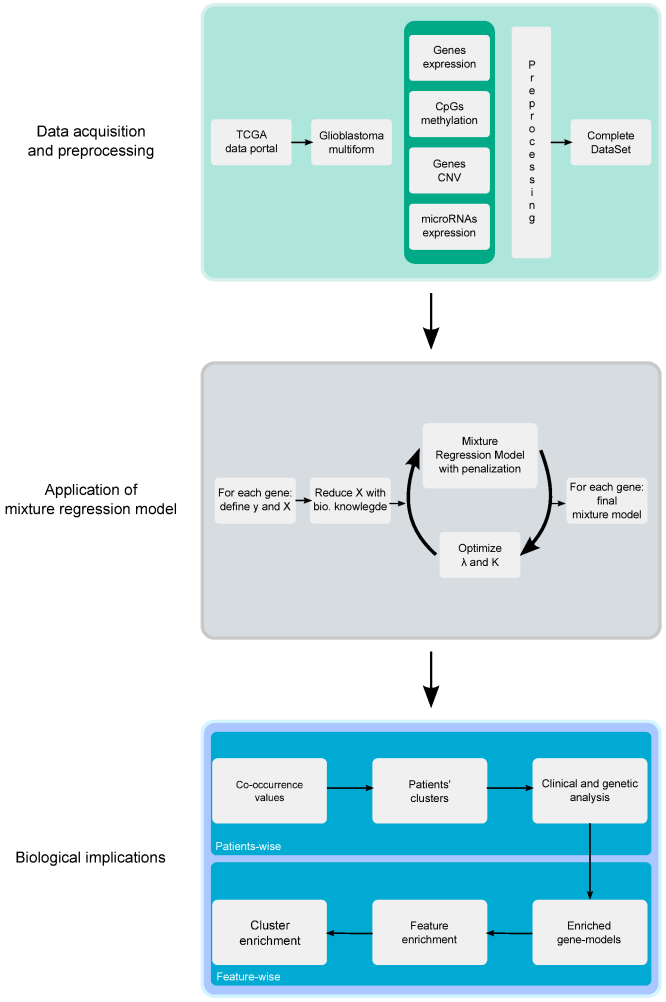


Figure 2: General organization of the methodology of the project, with the three main stages represented.

3.3 DATA OBTAINMENT AND PREPROCESSING

The first phase of this project is the obtainment of the data and its preprocessing. Here a detailed account of the data and the methods is presented.

3.3.1 *Types of data used*

Several directives were defined in order to choose the type of data for this project. In the first place, GBM was selected from a number of possible cancer types with publicly available datasets. This decision was taken based on our experience from analyzing brain-related tumor samples in previous projects (Eisenreich et al., 2013; Klink et al., 2013), which allow us to discuss and compare any novel discovery with our collaborators.

In the second place, for the molecular profiles, two critical conditions should be met: a large number of available samples and a low complexity in the integration of the data into a linear model. The first condition refers to a critical point in the use of learning algorithms: a high number of data points is needed to train successfully any complex model. For example, transcriptomic sequencing data is available for analysis, but the number of samples is still very low compared to array-based data for this type of cancer [1]. The second condition can be seen in the integration of SNP data, where the high dimensionality (e. g. one feature for each base in the genome/exome) and sparsity (e. g. low number of somatic mutation in comparison with the whole genome/exom (Guichard et al., 2012; Pleasance et al., 2010)) makes it very costly to integrate into a linear model.

Because of these restrictions, the data used in our study is comprised by expression (genes and microRNA), methylation (CpG sites methylation) and CNV, which are referred in this document as the molecular layers of our models.

In addition, several datasets of additional information were used in this project, such as human pathways, PPI, somatic mutations and microRNA interactions. These databases were used in the construction of the models to connect the 4 molecular layers (gene and microRNA expression, CpGs methylation and CNV) and to apply somatic mutation analysis in our findings. The list of these databases can be seen in the following section.

3.3.2 *Data sources*

The data used in this work comes from several sources, which are:

1 In February 2015 there were publicly available 170 RNA-seq and 558 microarray samples of GBM data in the TCGA repository.

TCGA DATA PORTAL: Repository from The Cancer Genome Atlas which contains molecular profiles for a plethora of cancer types, including GBM. The following datasets were obtained here: gene expression, CNV, methylation and microRNA expression (The Cancer Genome Atlas Research Network, 2014). Data downloaded in February 2015.

REACTOME HUMAN PATHWAYS: An open and curated human pathway database (Croft et al., 2014; Milacic et al., 2012). Data downloaded in November 2014.

NCI PATHWAY INTERACTION DATABASE: Pathway database curated by the National Cancer Institute (Schaefer et al., 2009). Data downloaded in November 2014.

BIOGRID: Interaction database named as The Biological General Repository for Interaction Datasets. Is a curated and public database with over 800 000 interactions (Chatr-Aryamontri et al., 2015). The data was downloaded in November 2014.

DIP: The Database of Interacting Proteins is an experimentally validated PPI database (Salwinski et al., 2004). The data was downloaded in October 2014.

ILLUMINA HUMAN METHYLATION BEADCHIP: Manifest file for the Illumina HumanMethylation450K platform (Illumina Inc., 2014). Contains genes related to the CpG sites in the chip. The data was downloaded in October 2014.

MIRTARBASE: The miRNA target database is a experimentally validated database for microRNA interactions (Hsu et al., 2014). The data was downloaded in October 2014.

MIRDB: Database for miRNA target prediction and functional annotations, which uses the tool MirTarget to predict targets of miRNAs (Wong and Wang, 2015). The data was downloaded in October 2014.

COSMIC: The Catalogue of somatic mutations in cancer is a free and public database of the mutations found in the TCGA and ICGC projects (Forbes et al., 2015). The data was downloaded in January 2016.

ICGC: International Cancer Genome Consortium.

In Table 1 detailed information about the datasets obtained from the TCGA repository are presented. For the interaction databases, the original raw values are presented in Table 2.

Table 1: Details on the data obtained from the TCGA data repository.

Data type	Platform	Number of Samples	Number of Features
gene expression	Affymetrix GeneChip Human Genome HT U133A Array	558	12 000
microRNA expression	Human miRNA Microarray 8x15K	584	1 500
CpG sites Methylation	Infinium HumanMethylation27 & HumanMethylation450 BeadChip Kit	289 & 144	27 000 & 450 000
CNV	Affymetrix Genome-Wide Human SNP Array 6.0	279	Depends on processing

Table 2: Size of the raw interaction datasets.

Dataset	Number of Interaction	Type of database
Reactome	152 266	Pathways
NCI	8 420	Pathways
Biogrid	245 355	Protein-protein interactions
DIP	6 294	Protein-protein interactions
Illumina Human Methylation BeadChip	485 461	Gene-CpG sites
miRTarBase	39 110	Gene-microRNA
miRDB	3 855 248	Gene-microRNA
COSMIC[a]	24 293	SNPs

a Number of SNPs found in our samples only. Original file was discarded due to size.

3.3.3 *Data Preprocessing*

The preprocessing of the data is a crucial step in the correct performance of learning algorithms (Kotsiantis et al., 2006). For this work, the focus is set on the normalization and preparation of the heterogeneous data for its integration and not in the preprocessing of the raw data.

For the data allocated in the TCGA data repository there are different preprocessing levels, which indicate the stage in the preprocessing pipeline. A short summary of these levels for array-based experiments is:

LEVEL 1: Raw data, as it comes from the platform software.

LEVEL 2: Background corrected data at the probe level, sometimes normalized.

LEVEL 3: Background corrected data at the gene level, most commonly normalized.

The method for normalization and mapping differs greatly on the data type and the institution behind the process. Below is shown a description of this process for each type of data used.

3.3.3.1 *Gene expression data*

Gene expression data was obtained as Level 3 from the TCGA portal, which means that the data had been quantile normalized and the probes had been mapped to their respective genes. The first step was to aggregate duplicated samples, on this level 29 of the 558 samples were duplicates. The duplicates shown a high correlation ($\rho > 0.92$) for all the samples. Because of this the duplicated samples were mean-aggregated and 539 final patient-level samples were obtained.

Next, the data was log-transformed and the features were centered. This process allowed us to get centered and normal-distributed values, which are necessary to integrate and compare the different expression profiles and find subsets of patients with particular expression values.

Because of issues in the integration of the heterogeneous data in a linear model the expression data was quantile normalized and min-max scaled to ensure a common range with the other data types, and to eliminate the possibility that outliers would take the whole range. For further discussion see Section 3.4.5.

3.3.3.2 *CpG sites methylation data*

The methylation data was obtained from two different platforms in its Level 3 form, where the probes had been mapped to Illumina's CpG sites id. The platforms were: Infinium HumanMethylation27 & HumanMethylation450 BeadChip Array.

The HumanMethylation450 array can be considered an upgrade of the HumanMethylation27, where additional probes have been added. Because of this characteristic it is possible to merge the information of this two platforms into one dataset, where only the shared probes (and CpG sites in Level 3) between platforms are conserved. Between these platforms over 26 000 CpG sites were shared. For some probes there was a high number of missing data points ($> 50\%$) in one of the platform. These probes were eliminated from both platform, leaving a final number of 21 000 CpG sites.

In the case of patient sample duplications it was found that only one sample appeared in both platforms, while 4 had duplicates in one of the platforms. In a similar fashion, as in the gene expression data process, the correlation between the duplicates was found to be high ($\rho > 0.95$) and the same mean-aggregation was applied. For the patient in both platform the correlation was even higher in the shared CpG sites ($\rho \sim 0.97$) and the aggregation was applied to it as well.

In this case the reported values for the CpG sites are the beta-values, which are define as fractions so their natural range is between 0 and 1 and no scaling was necessary.

3.3.3.3 *miRNA expression data*

In the case of microRNA expression, samples of Level 1 data were obtained and preprocessed. This difference with the other data types is due to several inconsistencies found in the Level 3 data. These inconsistencies appeared when comparing the number of samples between the raw data and the preprocessed data. Furthermore, difficulties were found to replicate the methodology used to create the Level 3 data.

The Bioconductor package *AgiMicroRNA* (Lopez-Romero, 2016) was used for the preprocessing, which allows a simple processing for Agilent array-based experiments and additional probes and genes filtering.

In a similar way to other datasets, the data was mapped to microRNAs, log-transformed, centered, quantile-normalized and finally min-max scaled. The control probes were filtered out, as well as genes with less than 75 % of non-`NA` values, which left us with 244 features.

Eight samples with duplicates were found, all of them showing a high correlation ($\rho > 0.80$). A mean-aggregation was applied leaving 576 unique samples.

3.3.3.4 *Copy-Number Variation data*

A different approach was used in the preprocessing of the CNV data. In this instance the Level 3 data obtained only showed copy number variations in genomic segments. In order to map these segments into genes and process the files the package *org.Hs.eg.db* (Carlson, 2016) was used to retrieve all human genes and non-coding elements, and their position. Additionally, the package *CNTools* (J Zhang, 2016) was used to map the genomic segments to these genomic elements (genes and non-coding elements) and created a matrix object from the data, obtaining around 22 000 features.

After this procedure, it was necessary to aggregate 28 duplicated samples using the mean, the majority shown a high correlation ($\rho > 0.85$), but for some of them the correlation was low ($\rho < 0.30$). It was decided to apply the aggregation nevertheless because there are no technical reasons to choose one replicate over the other and having multiple replicates would forbid us to apply the mixture of linear models.

One particularity of these samples is the presence of the CNV for normal tissue. This was used to normalized the CNV of the tumor samples, following the relationship:

$$\mathrm{CNV}_{\mathrm{Final}} = \log_2 \left(\frac{\mathrm{CNV}_{\mathrm{Tumor}}}{\mathrm{CNV}_{\mathrm{Normal}}} \right)$$

Additionally, a filter was applied to the features, where genes and other genomic elements were eliminated if their variance was under a certain threshold for all the samples($\mathrm{MAD} \leqslant 0.1$).

MAD: Median absolute deviation

Finally, the features were centered and the min-max scale was applied in order to ensure a range between $[-1, 1]$.

3.3.3.5 *Interaction datasets*

Several interaction databases were processed, these interactions allowed us to find relationships between the elements of each layer or between elements in different layers. This information was used to restrict possible covariants in our model, as explained in Section 3.4.3.1.

The databases were downloaded and separated into 2 groups: PPI databases and inter-molecular interaction databases. The first one refers to databases with binary interaction between 2 proteins, this includes the databases Reactome, NCI, Biogrid and DIP. The second group contains binary interactions between molecules of different kind, e. g. gene-microRNA interaction, these databases are Illumina Human Methylation BeadChip, miRDB and MTI.

The processing consisted of sorting the binary interactions, for example, the interaction $\mathrm{B} \longrightarrow \mathrm{A}$ becomes $\mathrm{A} \longrightarrow \mathrm{B}$, followed by an elimination of self and duplicated interactions, therefore removing directed interactions and leaving only undirected ones.

The PPI databases were merged into one unique binary interaction set without duplicates. This procedure was also done for the 2 gene-microRNA databases (miRDB and MTI), with the particularity that for the miRDB set contains only those putative interactions with over 80 % confidence were used.

3.3.3.6 *Final dataset*

After the preprocessing step, the number of samples and features for each dataset have been modified, as well as some of the values of the sets due to the normalization and aggregation. Similar variation can be seen in the interaction sets. In Table 3, the final dimensions and data profiles for each dataset are shown, and in Table 4, the size of the binary interaction sets are presented. After the preprocessing step, the patients shared by all the datasets were selected summing up a total of 324 paired samples.

The general methodology for each one of the data types is displayed in Figure 3.

Table 3: Datasets after processing.

Data type	Number of Samples	Number of Features	Type of Distribution	Range of Data
Gene expression	539	12 042	Normal	[-1,1]
miRNA expression	576	244	Normal	[-1,1]
CpG sites Methylation	289 & 144[a]	21 048	Bimodal	[0,1]
CNV	503	2 559	Normal	[-1,1]

a For the 27K and 450K HumanMethylation Platforms respectively

Table 4: Interaction datasets after processing.

Dataset	Number of Interaction	Type of Interaction
Merged PPI	413 212	Protein-protein
Illumina Human Methylation Bead-Chip	31 375	Gene-CpG site
microRNA databases	4 921	Gene-microRNA

Figure 3: Main preprocessing steps for the TCGA datasets.

3.4 APPLICATION OF A MIXTURE OF LINEAR REGRESSION MODELS ON HETEROGENEOUS CANCER DATA

The main model used in this work corresponds to a penalized mixture of finite linear regression models. Because of its complexity, it has been decided to present this model in an agglomerative way. The simple core of the model, the linear regression model, is explained first and layers of complexity are added into it in the following sections.

3.4.1 *Linear regression with cancer data*

3.4.1.1 *Definition of linear regression model*

Given a dataset comprised of N samples with $P+1$ variables:

$$\{y_n, x_{n,1}, \ldots, x_{n,p}\}_{n\in[1,N]}$$

For one sample, n, a linear regression model is defined as linear relationship between a response or output variable, y_n, and the other P variables, called explanatory variables or covariates, plus an error term ϵ_n:

$$y_n = \beta_{n,0} + \beta_{n,1}x_{n,1} + \ldots + \beta_{n,p}x_{n,p} + \epsilon_n$$

Where the β values represent the effect or coefficients of each covariate over the output variable and $\beta_{n,0}$ is defined as the intercept, a constant effect independent of the covariates.

In a linear regression model it is assumed that all these coefficients are shared between the samples, i. e. $\beta_{n,0} = \beta_0 \ \forall n \in [1, N]$. These coefficients are also unknown and their learning is the main objective of fitting a regression model.

This set of models can be written in matrix form, which is preferred in this document. The different elements for the matrix form are defined below; in first place, the vector of output variables **y**:

$$\mathbf{y} := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

In second place the design matrix **X**, which comprises the covariates of the system:

$$\mathbf{X} := \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

Here, each row represents a sample and each column an explanatory variable.

Finally two additional vectors are defined: $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$:

$$\boldsymbol{\beta} := \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \boldsymbol{\epsilon} := \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The first vector groups the p different coefficients, one for each covariate, while $\boldsymbol{\epsilon}$ is comprised by the n random errors.

An additional step is performed to include the intercept coefficient into the matrix form. For this, the intercept term is added to the $\boldsymbol{\beta}$ vector:

$$\boldsymbol{\beta} := \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

And an extra column is added to the design matrix:

$$\mathbf{X} := \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

This column allow us to represent the intercept while maintaining its independence from the covariates.

Finally, putting all together we get the matrix form of the linear regression model:

$$\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

3.4.1.2 *Application of linear regression model with biological data*

In order to apply the linear regression model to biological data, the datasets were created and the different elements of the model (covariates, output variable and random error) defined.

The datasets obtained from the TCGA repository were comprised by 4 different molecular profiles or layers: gene and microRNA expression, CpG island methylation and CNV. This data was paired, which means that for each patient there was information available at all layers. This characteristic came from the design of the experiments by the TCGA consortium, where a plethora of analysis were performed to each patient.

Due to this characteristic, the datasets could sorted by patient and the different layers combined.

The different datasets were:

$$\text{Gene Expression} = \begin{bmatrix} \text{gene exp}_{1,1} & \cdots & \text{gene exp}_{1,\text{p\_gene}} \\ \vdots & \ddots & \vdots \\ \text{gene exp}_{n,1} & \cdots & \text{gene exp}_{n,\text{p\_gene}} \end{bmatrix}$$

$$\text{miRNA Expression} = \begin{bmatrix} \text{miRNA exp}_{1,1} & \cdots & \text{miRNA exp}_{1,\text{p\_miRNA}} \\ \vdots & \ddots & \vdots \\ \text{miRNA exp}_{n,1} & \cdots & \text{miRNA exp}_{n,\text{p\_miRNA}} \end{bmatrix}$$

$$\text{CpG Methylation} = \begin{bmatrix} \text{CpG meth}_{1,1} & \cdots & \text{CpG meth}_{1,\text{p\_meth}} \\ \vdots & \ddots & \vdots \\ \text{CpG meth}_{n,1} & \cdots & \text{CpG meth}_{n,\text{p\_meth}} \end{bmatrix}$$

$$\text{CNV} = \begin{bmatrix} \text{CNV}_{1,1} & \cdots & \text{CNV}_{1,\text{p\_cnv}} \\ \vdots & \ddots & \vdots \\ \text{CNV}_{n,1} & \cdots & \text{CNV}_{n,\text{p\_cnv}} \end{bmatrix}$$

Where $\text{p\_gene}, \text{p\_miRNA}, \text{p\_meth}, \text{p\_cnv}$ are the number of different elements (features) for each set and we defined p as the total number of the elements. Giving that, the complete dataset used in this project was:

$$\text{Complete Dataset} =$$

$$\left[\ \text{Gene Exp} \ \middle|\ \text{miRNA Exp} \ \middle|\ \text{CpG Methylation} \ \middle|\ \text{CNV}\ \right]$$

For this project, each element of the Gene Expression Dataset was considered as a possible output variable (**y**). This implied that there were p_gene models (or gene-models) with $p - 1$ covariates. For the error element it was considered to be an independent variable following a centered normal distribution, i. e. $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Which gave us the following linear regression model for a particular gene:

$$\begin{aligned} \mathbf{y}_j = & \mathbf{X}^{\text{gene exp}}_{\forall gene \neq j} \cdot \boldsymbol{\beta}^{\text{gene exp}} + \mathbf{X}^{methylation} \cdot \boldsymbol{\beta}^{methylation} + \\ & \mathbf{X}^{\text{miRNA exp}} \cdot \boldsymbol{\beta}^{\text{miRNA exp}} + \mathbf{X}^{CNV} \cdot \boldsymbol{\beta}^{CNV} + \boldsymbol{\epsilon}_n \end{aligned}$$

This type of model assumed a lack of correlation between the independent variables, which cannot be sustained in a co-regulated biological system. The use of a penalization term (see Section 3.4.3) allowed the system to eliminate interactive variables. The implications of this procedure in our results are examined in the discussion chapter.

3.4.1.3 *Integration of multiple CpG sites methylation plastforms in the linear model*

The preprocessing of the methylation datasets presented an unique characteristic: the use of multiple platforms for its measurement (Infinium HumanMethylation27 and HumanMethylation450 BeadChip Arrays). This opened an issue in the integration of these datasets in a mixture regression model, where the effect from the different platforms could create platform-driven subpopulations.

A method to eliminate this kind of bias is the use of dummy variables (Faraway, 2014). Where a dummy variable is a binary variable that separates the samples based on a categorical effect. In general:

$$y = \beta_0 + \beta_1 x_1 + \alpha\delta$$

Here we have δ as our dummy variable, which for some samples will be 0, and thus the model will be a normal linear regression model: $y = \beta_0 + \beta_1 x_1$. While for some other samples it will take the value of 1 and the linear model will become: $y = (\beta_0 + \alpha) + \beta_1 x_1$. Any difference that comes from the categorical effect will be taken by the extra intercept α. In our model, δ will differentiate the patients' platform.

A problem with this approach occurs when it is used in conjunction with other data types. For example, considering a CpG site methylation Meth_1 and a gene expression Gene exp_1 our model becomes:

$$y = \beta_0 + \beta_1 \text{Meth}_1 + \beta_2 \text{Gene exp}_1 + \alpha\delta$$

For the samples where $\delta = 1$ we have:

$$y = (\beta_0 + \alpha) + \beta_1 \text{Meth}_1 + \beta_2 \text{Gene exp}_1$$

and thus the parameter α will explain not only the categorical effect on Meth_1, but also a possible biological effect from Gene exp_1. To solve this issue an interaction dummy variable was used instead of an additive dummy variable:

$$y = \beta_0 + \beta_1 \text{Meth}_1 + \beta_2 \text{Gene exp}_1 + \alpha\delta\text{Meth}_1$$

So when we have $\delta = 1$ the model becomes:

$$y = \beta_0 + \beta_2 \text{Gene exp}_1 + (\beta_1 + \alpha)\text{Meth}_1$$

With this method only the categorical effect is taken by α and instead of modifying the intercept it modifies the slope of the variable.

This approach can be extended to the case with p_{meth} different CpG sites.

$$y = \beta_0 + (\beta_1 + \alpha_1\delta)\text{Meth}_1 + (\beta_2 + \alpha_2\delta)\text{Meth}_2 + \cdots$$
$$+(\beta_{p\_meth} + \alpha_{p\_meth}\delta)\text{Meth}_{p\_meth}$$
$$+\beta_{p\_meth+1}\text{Gene exp}_1$$

In that case for $\delta = 1$:

$$y = \beta_0 + (\beta_1 + \alpha_1)\text{Meth}_1 + (\beta_2 + \alpha_2)\text{Meth}_2 + \cdots$$
$$+(\beta_{p\_meth} + \alpha_{p\_meth})\text{Meth}_{p\_meth}$$
$$+\beta_{p\_meth+1}\text{Gene exp}_1$$

We want to set all the α with the same value, i. e. same effect for each CpG site:

$$y = \beta_0 + \beta_1\text{Meth}_1 + \beta_2\text{Meth}_2 + \cdots$$
$$+\beta_{p\_meth}\text{Meth}_{p\_meth} + \alpha \sum_{i=1}^{p\_meth} \text{Meth}_i$$
$$+\beta_{p\_meth+1}\text{Gene exp}_1$$

Which means we can add an extra covariant that will be 0 for all the samples from platform 1 ($\delta = 0$) and will be $\sum_{i=1}^{p\_meth} \text{Meth}_{ij}$ for each sample j from platform 2 ($\delta = 1$), and an extra coefficient α which will take the categorical effect of the platforms.

3.4.2 *Mixture of finite regression models and EM algorithm*

The main hypothesis of this work is that in different patients subgroups there are, for certain genes, different linear models that explain the data. This hypothesis is motivated by the biological assumption that complex diseases, such as cancer, have multiple molecular subpopulations even when the disease phenotype is the same. In other words, under the microscope the samples of tissue cannot be differentiated, but based on the molecular signatures the samples correspond to different subtypes of the same disease.

To study this, an extension of regression models can be use, which consists of a probabilistic model comprised by a mixture of a finite number of linear regression models by which the data will be explained. These linear models are called submodels or subpopulations, because they define a model for each possible subpopulation in the data.

For the basic regression model it is possible to present it from a probabilistic perspective (Bishop, 2006):

$$y_n = \boldsymbol{\beta}^T x_n + \epsilon$$
$$\Longleftrightarrow$$
$$p(y_n|x_n, \boldsymbol{\beta}, \sigma^2) = \mathcal{N}(y_n|\boldsymbol{\beta}^T x_n, \sigma^2)$$

Now we define the mixture of regression models using the probabilistic perspective:

$$p(y_n|x_n, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(y_n|\boldsymbol{\beta}_k^T x_n, \sigma^2)$$
$$\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\beta}, \sigma^2\}$$

Where **K** independent linear models are considered, each one with its own set of coefficients $\boldsymbol{\beta}_k$ and same noise variance σ^2. This model is a mixture of the **K** subpopulations, where each one of them is weighted by a mixing coefficients π_k. With:

$$\sum_{k=1}^{K} \pi_k = 1$$

Giving all the data points $\{\mathbf{y}, \mathbf{X}\}$ the log likelihood of this density is:

$$\ln p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(y_n|\boldsymbol{\beta}_k^T x_n, \sigma^2) \right)$$

Normally, after obtaining the log likelihood framework, it is possible to find the parameters of the model that maximize the likelihood given the data. In this case, due to the presence of the summation inside the logarithm, it is not possible to apply the logarithm function over the Gaussian as in the single Gaussian method, and so, the derivative of the log likelihood function no longer allows us to obtain a close form solution (Bishop, 2006). To solve this problem a common approach is to use the EM algorithm.

3.4.2.1 *EM algorithm*

The EM algorithm (**E**xpectation-**M**aximization algorithm) is a method that finds the maximum likelihood estimates of parameters in a statistical model which contains unobserved variables (latent variables) (Bishop, 2006).

The latent variables for the mixture of regression models are the variables that determine the subpopulation from which a data point originates.

The EM algorithm is a 2-steps iterative method, comprised by the Expectation step (E-step) and the Maximization step (M-step).

In the E-step, the expectation of the likelihood function of the complete dataset (with the latent variables) is calculated. For this, the initial values of the parameters are necessary. This step calculates the posterior probability of the latent variables.

For the M-step, the expectation of the complete dataset likelihood is maximized by choosing a new set of parameters. These new parameters are used as initial values in the new iteration.

Through this iterative process, the optimal values for the parameters of the models are chosen.

For a detail description of the algorithm and its application in mixture of regression models, please see Chapter 14 in (Bishop, 2006)

3.4.3 *Penalization and biological prior knowledge*

A recurrent problem in regression models is the presence of a large number of features, P, when compared to the number of samples, N. This characteristic can affect in multiple ways our methodology. The main one is denominated the curse of dimensionality, where in a space with a very large number of dimensions ($\mathrm{P} \gg \mathrm{N}$) the distance between points loses its significance, due to the sparsity of the points. And thus, many models will fit the data in a similar fashion, making them indistinguishable.

Additionally, the execution time needed to learn the models grows with the number of features, making it unfeasible to train this models when there are thousands of features.

For this project two approaches were used to deal with this issue: biological prior knowledge and *lasso* penalization. The first one is applied before the execution of the Mixture of Finite Linear Regression Models with Penalization (MFLRMP) and modifies the design matrix used as input, while the second method is applied during the fitting of the mixture model.

3.4.3.1 *Biological prior knowledge*

The biological knowledge that has been gatherer by previous experiments can by used to reduce the number of covariants of a model. The objective of this procedure is not to find the space of true interactions, but to shrink the space by eliminating interactions not supported by the prior knowledge.

The methodology used to filter features is the following:

- Binary interactions obtained from multiple sources, see Section 3.3.2, were sorted and filtered, to eliminate duplicates and self interactions.

- The interactions are expanded to all the molecular profiles. For example, $\text{Protein}_A \longleftrightarrow \text{Protein}_B$, a binary protein-protein interaction was expanded to $\text{Gene}_A \longleftrightarrow \text{Gene}_B$ and $\text{Gene}_A \longleftrightarrow$ CpGs related to Gene_B.
- The covariants in the models are filtered by the expanded interactions, i. e. if a gene, CpG or microRNA does not appear in the expanded interactions, it is eliminated from the model.

With this filtering the possible number of covariants for each gene with expression data is reduced from the possible $p - 1$ covariants to a smaller set and thus, the design matrix of each model was reduce before the fitting of the mixture model.

3.4.3.2 *Lasso penalization*

The *lasso* or L^1 penalization adds a regularization term to an error or likelihood function[2], where the regularization term is the L^1 norm of the model coefficients. The *lasso* penalization term is defined as:

$$\lambda \|\beta\|_1$$

with:

$$\|\beta\|_1 = \sum_j |\beta_j|$$

When this term is added to the negative log-likelihood function it will try to shrink the summation of coefficients during the minimization process. A particular characteristic of the *lasso* penalization is that the coefficients will be driven to 0 and not to small non-zero values as with other norms (Tibshirani, 1996). The number of non-zero coefficients is governed by the hyperparameter λ, where a larger value results in a stronger penalization.

An additional term was added following the implementation of a penalized mixture regression model by Städlet et al. (Städler et al., 2010):

$$\lambda \sum_{k=1}^{K} \pi_k^{\gamma} \|\beta_k\|_1$$

where the penalization term considers all the models and each norm is weighted by the mixing coefficient of that subpopulation. This approach allowed us to relate the penalty to the relative size of the subpopulation, which is a common practice as stated in Khalili and Chen (Khalili and Chen, 2007). The γ coefficient can take values $[0, 1]$ and controls the weighting of the mixing coefficient. In this implementation a $\gamma = 1$ is used.

2 In the case of linear regression these are equivalent.

The *lasso* component gives the term penalization in the name of our main algorithm: Mixture of Finite Linear Regression Models with Penalization.

Finally, the *lasso* behaves in a particular way in cases where highly correlated covariates are present: it selects one of the variables from the group randomly and gives a zero-coefficient to the others (Bühlmann et al., 2013). This behavior permits the use of biological variables, with a possible co-regulated interactions, in a regression framework, but with the cost of selecting only one of the variables. The effect of this action in our results are discussed in the discussion chapter.

3.4.4 *Fitting of the model hyperparameters* $\mathbf{K}$ *and* λ

Two parameters have been presented in different equations over this chapter, but their fitting has not been discussed yet. These parameters are : $\mathbf{K}$ and λ, which govern two different characteristics of the mixture model.

The total number of subpopulations in the mixture, $\mathbf{K}$, is not known *a priori*, and in a similar way with other clustering algorithms (e. g. K-means), it must be selected by an iteration over different possible values.

The penalization parameter λ governs the portion of the total coefficients that will be driven to zero, where with a larger lambda the penalization is stronger, therefore a larger portion of the coefficients will become zero. On the contrary, when lambda is small the penalization term will be negligible and fewer coefficients will be driven to zero.

The method to fit these two parameters and find the optimal value for them ($\mathbf{K}^*$ and λ^*) follows the implementation by Städlet et al. (Städler et al., 2010). In the first place, a grid of possible values for $\mathbf{K}$ and λ is defined: $\mathcal{K} := \{\mathbf{K}_0, \ldots, \mathbf{K}_l\}$ and $\Lambda := \{\lambda_0, \ldots, \lambda_m\}$. With l, m and the sets defined in Section 4.3.1.

Then an iterative process is applied to select the optimal parameters: for each $\mathbf{K} \in \mathcal{K}$ the log-likelihood loss is calculated for each $\lambda \in \Lambda$ following a cross-validation approach. With these cross-validation errors calculated, the different λs are compared and the one with the smallest error is selected for each $\mathbf{K}$, denominated λ_K, with $\mathbf{K} \in \{\mathbf{K}_0, \ldots, \mathbf{K}_l\}$.

BIC: Criterion to compare and select models. Defined as a penalized likelihood function, which takes into account the number of parameters in the model

Finally, for each $\mathbf{K} \in \mathcal{K}$ and the paired λ_K a model is fitted. To compared these models the Bayesian information criterion (BIC) is calculated and the model with the lowest BIC is selected. Thus, the optimal parameters are those that minimize the BIC.

3.4.5 *Issues in data integration*

In statistical algorithms, the problem of feature dominance can arise when one of the features, or a set of them, have a much larger variance than the rest. This can be seen in algorithms such as PCA and other learning algorithms, where the variance plays a main role (Berg et al., 2006; Borgognone et al., 2001).

PCA: Principal component analisis

With the implementation of the mixture regression model with penalization in this study, an issue arose when covariates with very different variances were used. In this case, the feature with the highest variance dominated the model, even in the case when all the features have the same mean (results not shown).

This means that the *lasso* penalization will give the covariates with larger variances most of the non-zero coefficients and it will heavily penalized those with smaller variances. This happens due to the minimization scheme, where it is "cheaper" (smaller L^1 norm) to minimize the coefficients of the features with small variance, than it is to minimize those with large variance.

In order to avoid this, a min-max scaling approach was applied, which aims to define a common range for all the data types.

Min-Max scaling: procedure to ensure that the range of a variable is fixed between a minimum and maximum value

Another problem triggered by the min-max scaling is the dominance of outliers. A min-max scale in a sample with outliers will scale-down all the non-outliers, therefore eliminating this information. Because of this issue, a quantile normalization was applied previous to the min-max scaling, where the package *preprocessCore* (Bolstad, 2015) was used for the quantile normalization.

Quantile normalization: technique to force variables to follow the same distribution

3.5 CLUSTERING OF GLIOBLASTOMA PATIENTS

One of the objectives of this project is to cluster the samples (patients) into distinct groups, and to analyze and profile the resulting groups.

In this section, the methodology for the first part of this objective is presented. In particular, the definition and use of the co-occurrence value is shown. This value is the central element that allows us to use the resulting gene-models for the clustering of the samples.

3.5.1 *Co-occurrence Probability*

The first step to perform the clustering was the definition of the co-occurrence probability ($\mathrm{P}_{\mathrm{CoO}}$) for a pair of samples. The co-occurrence represents the chance that 2 patients are in the same subpopulations for a single gene-model.

In Figure 4, the case of 3 samples in 3 gene-models can be seen. Considering only the first gene-model (A) the red and blue samples are in the same subpopulation, while the green sample is by itself.

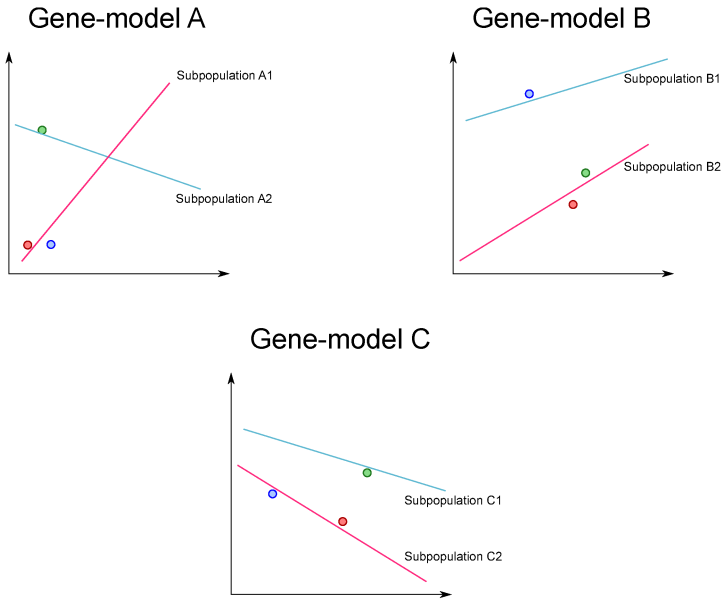


Figure 4: Co-occurrence example: Considering 2 linear models, the co-occurence for 3 samples is shown. For the pair of samples red-blue the co-occurrence happens in 2 of the 3 models, while for the pair red-green they share only 1 subpopulation (B2).

Thus, the pair red-blue has a larger co-occurrence probability than the pair red-green.

In order to calculate this value the posterior probabilities for each patient over each gene-model subpopulation are used. These probabilities came from the application of the MFLRMP algorithm, where the chance that a patient belongs to a particular subpopulation after the gene-models fitting was obtained.

Because the posterior probabilities for a pair of patients are independent from each other, it was possible to combine these values as following:

$$\mathrm{P}_{\mathrm{post}}(\mathrm{i},\mathrm{j}) = \mathrm{P}_{\mathrm{post}}(\mathrm{i})\mathrm{P}_{\mathrm{post}}(\mathrm{j})$$

The co-occurrence probability for the pair of patients i,j for a gene-model with 2 subpopulations was defined using the posterior probabilities as:

$$\mathrm{P}_{\mathrm{CoO}}(\mathrm{i},\mathrm{j}) = \underbrace{\mathrm{P}_{\mathrm{post}}^{\text{model 1}}(\mathrm{i})\mathrm{P}_{\mathrm{post}}^{\text{model 1}}(\mathrm{j})}_{\mathrm{P}_{\mathrm{post}}^{\text{model 1}}(\mathrm{i},\mathrm{j})} + \underbrace{\mathrm{P}_{\mathrm{post}}^{\text{model 2}}(\mathrm{i})\mathrm{P}_{\mathrm{post}}^{\text{model 2}}(\mathrm{j})}_{\mathrm{P}_{\mathrm{post}}^{\text{model 2}}(\mathrm{i},\mathrm{j})}$$

The value $\mathrm{P}_{\mathrm{CoO}}(\mathrm{i},\mathrm{j})$ represents the probability that the patients are together for that gene-model. It is possible to extend it to any gene-model with different numbers of subpopulations.

Finally, the total co-occurrence probability, also denominated co-occurrence value, considering all the gene-models trained, can be calculated by the mean over all of them:

$$P^{\text{Total}}_{\text{CoO}}(i,j) = \frac{1}{L}\sum_{l}^{L} P^{l}_{\text{CoO}}(i,j)$$

where L is the total number of gene-models used in the analysis.

As an example of the total co-occurrence value, in Figure 4 3 gene-models are shown (A, B and C) with 3 samples (red, blue and green). Each gene-model presents 2 subpopulations, named 1 and 2 for each one of them. It can be seen that the red and blue samples are present in the same subpopulations for the models A and C, while the pair red-green of samples appears together only in subpopulation B2.

In this example, the co-occurrence of samples red and blue is $2/3$, while for red and green is $1/3$.

3.5.2 *Gene-models selection*

For the co-occurrence analysis, the gene-models were filtered based on two criteria: Firstly, the necessity to have at least 2 subpopulations, which comes as an obvious requirement considering that this is an analysis aiming to calculate the co-occurrence in multiple subpopulations.

Secondly, a maximum threshold was defined for the posterior probability of each subpopulation. This means that the total posterior probability of a subpopulation (mean of all the individual probabilities) can not be over 0,75. This threshold was defined to eliminate the chance use of gene-models where a subpopulation dominates over the others. For those cases, it was found that the results were non-informative and in practice, are considered as gene-models with 1 subpopulation.

For the data presented in this work, the number of gene-models with more than 1 subpopulation comprises about 3 200 gene-models (~32 % of the total). While the number of gene-models without a dominating subpopulation is around 2 400 models (~23 % of the total).

3.5.3 *Definition of the co-occurrence probability matrix*

With the calculation of the $P^{\text{Total}}_{\text{CoO}}$ for each pair of patients, the co-occurrence probability matrix (CoM) was created. This matrix has dimensions $N \times N$ and its cell composition is such that the cell (i,j) represents the total co-occurrence value for the pair of patients i and j. This matrix is symmetrical, but one critical characteristic is that the diagonal of the matrix is not necessarily 1, which is expected of distance matrices.

N: *total number of samples*

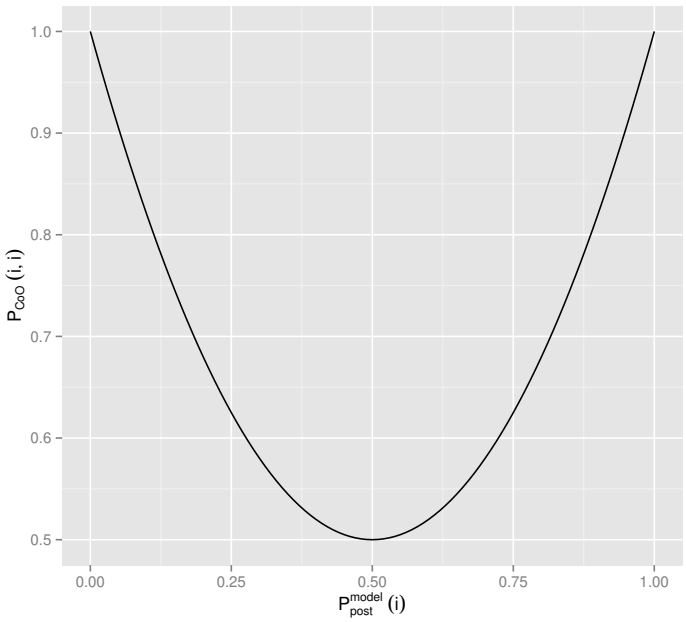


Figure 5: Curve of the co-occurence probability for the case of self-occurrence.

This happens due to the definition of the co-occurrence probability, where in the case of co-occurrence of the patient with itself it becomes:

$$\begin{aligned} P_{CoO}(i,i) &= P_{post}^{model\ 1}(i)P_{post}^{model\ 1}(i) + P_{post}^{model\ 2}(i)P_{post}^{model\ 2}(i) \\ &= P_{post}^{model\ 1}(i)^2 + P_{post}^{model\ 2}(i)^2 \end{aligned}$$

Considering: $P_{post}^{model\ 2}(x) = 1 - P_{post}^{model\ 1}(x)$ we get:

$$P_{CoO}(i,i) = 1 - 2P_{post}^{model\ 2}(i) + 2P_{post}^{model\ 2}(i)^2$$

The profile of the co-occurrence probability for this case is shown in Figure 5. An inspection of the curve reveals that the only possibility to have $P_{CoO}(i,i) = 1$ is when the posterior probability is either 0 or 1, which explains why it is common to find values other than 1 in the diagonal of the CoM. In addition, due to the parabolic profile of the curve the probabilities are skewed towards 0,5, making values close to 1 uncommon.

3.5.4 *Co-occurrence probability matrix as a distance matrix*

It is possible to use the CoM as a distance matrix in order to analyze the clustering of the patients based on their co-occurrence probabilities. This process was performed by converting to values of the matrix so that a high co-occurrence became a value close to 0 and a low co-occurrence was close to 1. This was done with the following linear transformation:

$$\mathrm{D}^{\text{Total}}_{\mathrm{CoO}}(i,j) = 1 - \mathrm{P}^{\text{Total}}_{\mathrm{CoO}}(i,j)$$

It is important to notice that $\mathrm{D}^{\text{Total}}_{\mathrm{CoO}}(i,j)$ is not an actual distance, in particular the coincidence axiom is not respected as shown in the previous section.

Coincidence axiom:
The distance of an object with itself is zero
$d(x,x) = 0$

One approach that can be use to overcome this issue is to normalize the matrix, making every element to the diagonal equal to 1. This procedure was tried but the cost of normalizing the matrix, e. g. min-max row-wise normalization, is the loss of the symmetry. Because of this drawback, this solution was not implemented and only a min-max normalization was applied to the whole matrix in order to make the values further sparse from each other.

3.5.5 *Clustering of patients*

The pseudo-distance matrix defined using the co-occurrence values is used as input for the hierarchical clustering algorithm. This method is used with our data to find subgroups of patients from the fitted gene-models. These subgroups are denominated in this work as clusters of patients.

The algorithm is run using complete agglomeration and the resulting dendrogram is cut based on visual inspection as well as silhouette analysis, which are shown in the results.

3.6 PATIENTS-WISE ANALYSIS OF THE CLUSTERS

The methods used to analyze the samples on each cluster are presented here, they comprise the clinical analysis, the comparison to previous presented clusters and the samples' mutations enrichment.

3.6.1 *Clinical analysis of the samples*

The impact of each clinical variable over each cluster of patients was analyzed.

The clinical data was obtained from the TCGA database. For 3 patients there was no clinical data available.

From the TCGA database, the following sets were used to perform the clinical analysis: age, gender, tumor type, Karnofsky performance scores, treatment history and survival. In general, the utilization of these sets was straightforward and in some cases, the datasets were incomplete. The information for each set is shown in the Section 5.2.2.

The particularities of the processing step for each set are:

AGE: The age of the patients at the time of diagnosis was segmented by the clusters found previously. To compare if the age in the groups originated from the same population a Kruskal-Wallis rank sum test was performed over the groups. This test is a non-parametric version of the one-way ANOVA test and it was selected due to the non-normal distribution of the age found through visual inspection. It was calculated using the `kruskal.test` function with default options in R.

GENDER: In the same fashion as with the age analysis, the gender of the patients was analyzed cluster-wise. To study any possible significant relationship between gender and clusters, a Pearson's Chi-squared test was apply to the data. The test was applied using the `chisq.test` function in R with default parameters for a contingency table.

TUMOR TYPE: The samples in the TCGA data were annotated with the histological types of the tumors found. The possible classes are: *Glioblastoma Multiforme (GBM)*, *Treated primary GBM* and *Untreated primary (de novo) GBM*.

A comparison with the supplementary files in Verhaak et al. (Verhaak et al., 2010) showed that the first two classes could represent secondary GBMs or recurrent tumors, and because there was no additional information regarding the histological type of the tumor, those classes were aggregated as non-de novo tumors and its distribution compared to the de novo patients.

KARNOFSKY PERFORMANCE SCORES: In a similar fashion as the gender analysis, a Pearson's Chi-squared test was applied to analyze the discrete and qualitative Karnofsky scores.

Neoadjuvant treatment: Treatment performed before surgery, e. g. chemo or hormonal therapy.

TREATMENT HISTORY: The history of neoadjuvant treatment was analyzed between cluster using a contingency table in a similar way as in the Tumor Type analysis.

SURVIVAL: For the survival analysis, the vital status and days to death were obtained from the clinical dataset. With this data the survival was modeled using Cox proportional-hazards regression with the R package `surv`. A log-rank test was performed over the survival curves in order to compare them altogether.

3.6.2 *Comparison of the clusters to Verhaak's subtypes*

The clusters obtained in this project were compared to the subtypes reported by Verhaak et al. (Verhaak et al., 2010). This was possible due to the use of the same database for the samples. The obtainment of additional files from the TCGA portal[3] allowed us to annotate the samples with the subtype found in Verhaak et al.

From the original 202 samples used in the Verhaak et al., only 50 are used in this project as well. The rest of the samples appear in the early stages of the data preprocessing, but were eliminated due to the lack of information in the 4 molecular layers (no-paired data).

The analysis is carried out by a direct comparison of the distribution of the shared samples in a contingency table.

3.6.3 *Genetic signatures in the clusters*

The somatic mutations used in this analysis were obtained from the COSMIC database and it contained information for 220 of the samples used in this study. The samples without this information were not taken into account and the impact of the missing information can be seen in the result section.

Following the analysis performed in Verhaak et al. (Verhaak et al., 2010), the mutations were aggregated by patients and genes, allowing us to analyze the number of patients on each cluster with at least one mutation in a particular gene locus.

Then a contingency table was created for each case considering the patients in a cluster and the number of them with a mutation in a gene locus. This procedure allowed us to test the significance of the relationship between a cluster and the presence of a high number of mutated samples for a particular gene. To test this relationship, the Fisher's exact test was chosen over the Pearson's Chi-squared test due to the small number of samples in some cases.

The test was run only for the cases were at least 6 samples presented a mutation for a particular gene and the p-values obtained from the test were adjusted using the Benjamini and Hochberg method.

3.7 GENE-MODELS AND FEATURES ANALYSIS OF THE CLUSTERS OF PATIENTS

After the definition of the clusters of patients in Section 3.5.5, the next step was to study the gene-models that enrich each cluster and its features.

3 https://tcga-data.nci.nih.gov/docs/publications/gbm_exp/

3.7.1 *Gene-models selection on each cluster of patients*

The aim of the first method of this section was to select a subset of significant gene-models for each cluster. By significant models we referred to the gene-models where the patients of a cluster had the highest co-occurrence and thus are the ones that defined that cluster.

There were different ways to do this. One method to select those defining gene-models was to calculate the mean co-occurrence probability for each gene-model over all the pairs of patients in a cluster and selected those models that had a mean value over a certain threshold. This method was similar to the one used to find the patients' clusters, but instead of calculating the mean values for each pair of patients over all the gene-models, in this case, the value was calculated over all the pairs of patients for each model.

A second methodology was to apply a hierarchical clustering over the co-occurrence distances (1 - co-occurrence value), calculating the distance between the gene-models over all the pairs on each cluster and to use visual inspection to separate the gene-models.

The former method was tested, but the definition of the cut-off value was not easily found and a method that utilized the distance between the models was preferred.

The results of the clustering of the gene-models are shown in Section 5.3 and the heatmaps for each cluster is shown in Section B.2 of the Appendix. The number of selected groups was determined by visual inspection due to the agglomeration of gene-models with high co-occurrence values in a small and well defined cluster.

3.7.2 *Selected gene-models analysis*

After the definition of the subsets of gene-models for each cluster the analysis over the gene-models was performed, we aimed to compare the distribution of these models between the clusters and to analyze their significance.

The first study was centered in the shared gene-models, in particular, gene-models that are present in several clusters. This analysis was performed in a straightforward manner: by analyzing the number of shared gene-models and their distribution.

In addition, an enrichment analysis was performed over the target genes of the shared gene-models with more than 4 appearances, in order to study if the target genes of these models belong to a specific pathway or other set. The enrichment was performed through the ConsensusPathDB website[4] (Kamburov et al., 2012). This allowed us to perform gene set enrichment analysis with an updated database (Release 31 - September 2015). The enrichment analysis was performed

4 http://cpdb.molgen.mpg.de/

using the KEGG[5] and Reactome pathway databases and the Gene Ontology sets[6].

KEGG: Kyoto Encyclopedia of Genes and Genomes

After the analysis of the shared gene-models, a study considering the different subpopulations of each cluster was performed. This additional characteristic in the analysis is vital considering that for each gene-model a cluster of patients is located in one of its subpopulations. This means that in the case of a gene-model that is shared by 2 clusters, if the samples of each cluster are not in the same subpopulation then the gene-model is actually not shared. This was done for gene-models with a presence in 4 or more clusters and for the gene-models shared in every pair of clusters.

3.7.3 *Gene-models' features selection and analysis*

Another approach to analyze the selected gene-models for each cluster of patients was to study the features that comprised these models for each cluster.

The analysis was performed over the features with non-zero coefficients for each selected gene-model subpopulation, taking into account that the patients of each cluster are located in one specific subpopulation.

Three different methodologies were proposed for the enrichment analysis of the features:

- An intra-cluster analysis, where shared features between the different gene-models on each cluster are studied.
- Ranked coefficients of each feature based on a significance test for linear regression and application of a ranked gene set enrichment analysis.
- Gene-set enrichment analysis of the features with non-zero coefficients.

The results for the first and third methods are presented in Section 5.3.3. The second method was not implemented and the discussion for this decision is presented in DISCUSSION.

3.7.3.1 *Intra-cluster feature aggregation*

In this approach, for each layer, the features with non-zero coefficients were aggregated for all the selected gene-models' subpopulations in a cluster. The analysis focused on the number of appearances of features over all the significant subpopulations, and included the 4 layers in the initial approximation. However, the study of the distribution of

5 http://www.genome.jp/kegg/
6 http://geneontology.org/

appearances of the features for each layer showed that the microRNA layers appeared with a much larger frequency than the other layers. Considering the small set of elements in that layer it was difficult to analyze and discriminate them.

After the exclusion of the microRNA layer, a coefficient denominated *appearance ratio* was defined. This value represents the total number of appearances of a covariant in the selected gene-models and subpopulations in a cluster divided by the total number of selected gene-models.

For a feature i, the appearance ratio (AR) is defined as:

$$\mathrm{AR_i} = \frac{\text{Number of Apperances}_\mathrm{i}}{\text{Number of Gene-models}} \times 100$$

This value was defined in order to compare the number of appearances between clusters, considering that the number of selected gene-models for each cluster is different and thus, a feature in a cluster with a large number of selected gene-models could appear more often than one in a cluster with small number of selected gene-models.

The features considered in the analysis must have at least 2 appearances and an appearance ratio equal or over 2 %. These values were considered as minimum thresholds and allowed us to discriminate features that might have a high appearance ratio in clusters with a small number of selected gene-models.

To study the significant features, a contingency table was used to compare their distribution by layer and cluster. Then, an analysis of the features was performed by cluster, based on the literature presented in the introduction.

Finally, for the case of significant features that appear in several clusters, an analysis of their subpopulation was performed, in order to study if not only the features, but also their subpopulations were shared between clusters.

3.7.3.2 *Gene-set enrichtment analysis*

The second methodology for the feature analysis of the selected gene-models was the use of gene-set enrichment. As presented in Section 2.1.3, the enrichment analysis is based on the comparison of two sets, one obtained from our results and the other from biological prior knowledge databases, such as pathways. In our project this was performed by taking into account all the features on a cluster with non-zero coefficients and computing with a previously known set (e. g. gene ontology, biological pathway) the statistical significance whether those features are overrepresented in that set.

A problem faced by this method was the composition of the gene sets to test, which, as the name suggest, are comprised by genes, which makes the use of the additional layers difficult.

Because of this characteristic, the CpG sites methylation layer was mapped into the gene layer using the manifest file for the Illumina Human Methylation Beadchip. On the other side, the microRNA layer was not taken into account given the previous results where several features of this layer are present in most models making it difficult to discriminate significant appearances.

After the mapping of the methylation features, we proceeded to apply the enrichment analysis and as in Section 3.7.2, the ConsensusPathDB interaction database was used. However, in this case, only the KEGG pathway database was used, and all the genes in the design matrix were defined as the background set.

The decision to work exclusively with the KEGG pathways database was done based on preliminary results, where the inclusion of additional databases delivered an extremely large number of significant enriched sets, making the discrimination of interesting sets difficult.

The obtained significant pathways were grouped based on their function and analyzed by their presence in the different clusters.

4

MIXTURE OF REGRESSION MODELS: IMPLEMENTATION AND RESULTS

The results obtained by this work have been separated into those concerning the technicalities of the implementation and execution of the algorithm of mixture of regression models, and the implication of the results obtained from it in the biological context of cancer research.

In particular, this chapter contains the analysis of the preprocessed data obtained from the TCGA data portal, the definition of the gene-models and their characteristics. Then it shows the implementation and run of the MFLRMP, and finally the results concerning the execution and performance of the algorithm.

4.1 TCGA DATA ANALYSIS AND CONSTRUCTION OF THE GENE-MODELS

The data obtained from the TCGA repository was preprocessed to create the datasets needed for this project as shown in the METHODOLOGY. Preprocessing was not only done for aggregation and normalization of the samples, but also to ensure a common range between the different data types, also called layers (Section 3.4.5).

One critical characteristic of the samples to be used in this work is that they have paired data for all the 'omic' layers. This subset of patients is referred to as the *shared patients* set and consists in 324 patients. The patients' id can be found in Section A.1.

4.1.1 *Number of models and covariates*

One we had the preprocessed datasets comprised of paired samples, the next step was the creation of the models. These models would follow this linear relationship for each $gene_j$:

$$\begin{aligned}\mathbf{y}_j = &\mathbf{X}^{\text{gene exp}}_{\forall gene \neq j} \cdot \boldsymbol{\beta}^{\text{gene exp}} + \mathbf{X}^{methylation} \cdot \boldsymbol{\beta}^{methylation} + \\ &\mathbf{X}^{\text{miRNA exp}} \cdot \boldsymbol{\beta}^{\text{miRNA exp}} + \mathbf{X}^{CNV} \cdot \boldsymbol{\beta}^{CNV} + \boldsymbol{\epsilon}_n\end{aligned}$$

Where $\mathbf{y}_j$ is the expression of the $gene_j$, $\mathbf{X}$ are the covariates, composed by the different layers, $\boldsymbol{\beta}$ are the coefficients of the covariates in the linear model and $\boldsymbol{\epsilon}$ is the error term. The model is also called 'gene-model' in this study due to its definition from a gene with expression data.

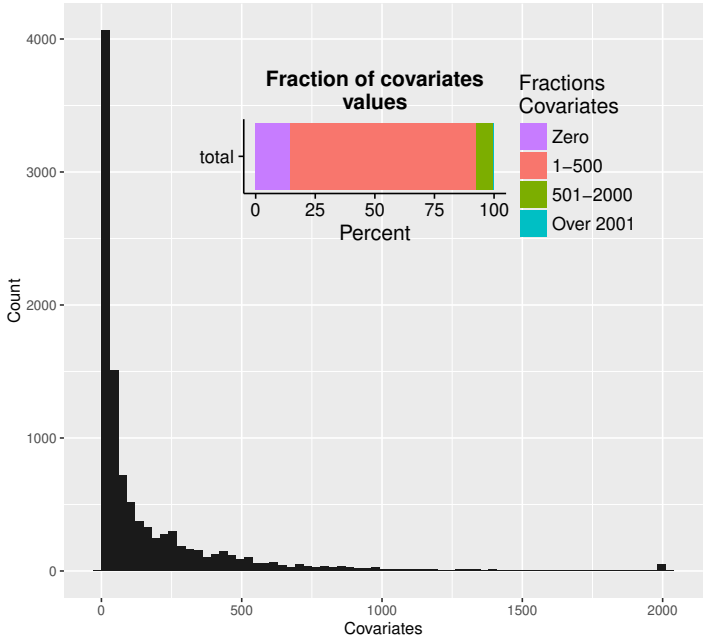


Figure 6: Main plot: Histogram of the number of covariates for the gene-models after the application of the biological prior knowledge. Models with zero covariates not shown and values over 2 000 are aggregated to 2 000.
Top plot: Grouped number of covariates by range of values.

Initially the number of models was equal to the total amount of genes with expression data and all the features in the sets were considered as covariates. This gave an initial value of 12 042 models, each one with $\sim$ 35 000 covariates. This would mean that the MFLRMP algorithm would be run over 12 000 times for models with thousands of covariates.

As explained in Section 3.4.3.1, the number of covariates was reduced through the use of biological prior knowledge. This resulted in gene-models with different numbers of covariates. In Figure 6, the distribution of the number of covariates for each gene-model, as well as the fraction of models with different number of covariates, are shown.

The distribution of the number of covariates follows an exponential distribution, with most values between 1 and 500. This can be easily seen in the fraction plot, where the portion of gene-models with more than 2 000 features is almost non-existent and the fraction of models with zero covariates is around 15 %. Excluding the models with zero or over 2 000 covariates the total number of models to train is 10 247.

The exclusion of gene-models with over 2 000 features was due to their low number and high cost. Models with a high number of covariates are difficult to train and time consuming, and because of the curse of dimensionality the trained models can be meaningless (see Section 3.4.3).

The threshold of 2 000 covariates was selected arbitrarily. This value was chosen to minimize the number of lost models (~ 50 models) and it was found by trial and error that the time expended in the last 50 models was similar to the time used to train all the gene-models with less than 2 001 covariates. Because of this characteristic 2 000 was selected as threshold.

Finally, the number of gene-models to run was reduced to over 10 000 but the most important element is that the number of covariates are mostly under 500. This is crucial considering that every gene-model is trained independently with the MFLRMP algorithm.

4.1.2 *Source of the covariates*

Another statistic considered was the source of the covariates of the gene-models after the application of the prior biological knowledge. In Figure 7, the 4 sources (biological layers) of the covariates for the models are shown. For each one, the distribution of the fraction of covariates that comes from that source is shown for every model.

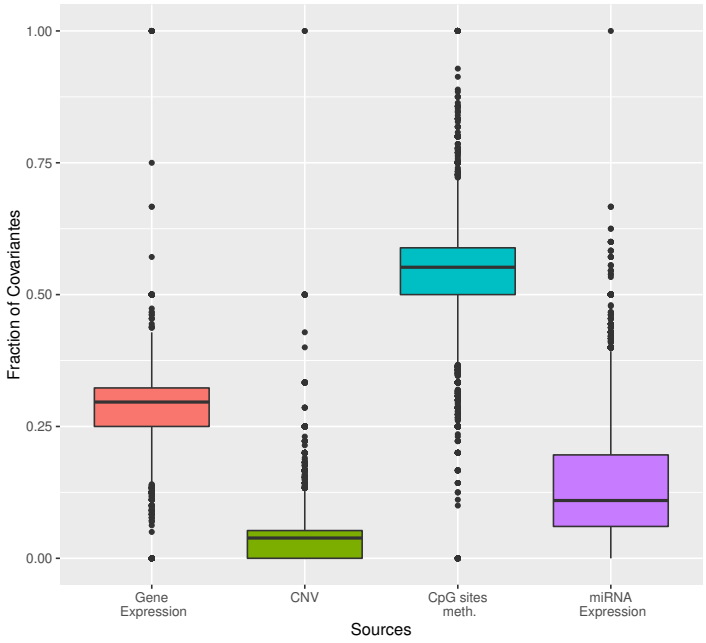


Figure 7: Distribution of the covariates sources of the gene-models. Considering biological prior knowledge and before the application of the penalization.

We can see that there is a significant difference in the distribution for each layer. On once side over 50 % of the gene-models' covariates come from the CpG sites methylation layer, while on the other side the CNV layer has a median fraction of 5 %. This difference is explained by a combination of the number of elements on each dataset and the

number of interactions between that layer and the gene set. In the case of CpG sites methylation layer there are over 20 000 elements and 30 000 interactions, while for CNV there are around 2 500 features and 400 000 interactions. With such a small number of elements in the CNV set, the high number of interactions are not significant and this explains the low presence of the CNV in the design matrix. Something similar happens with the microRNA variables, where there is around 244 elements and 5 000 interactions.

4.2 IMPLEMENTATION OF THE MFLRMP ALGORITHM AND PERFORMANCE

The original implementation of the MFLRMP algorithm can be found in Städler et al. (Städler et al., 2010). As part of this project, a new implementation of this algorithm was performed, aiming to produce a faster implementation. To accomplish this, the code was ported to `C++` and a wrapping R package was created for it. To link the `C++` code to the R package the libraries `Rcpp` (Eddelbuettel and François, 2011; Eddelbuettel, 2013) and `RcppArmadillo` (Eddelbuettel and Sanderson, 2014) were used. `RcppArmadillo` was especially useful to code the linear algebra needed for this algorithm.

The new implementation was compared to the original one, which was written in pure R code. This comparison was performed using the `microbenchmark` package[1] with 100 iterations for each comparison. A ratio was calculated from these results, which reflects how many *times faster* the new implementation is compared to the original one.

$$\text{Ratio} = \frac{\text{Median Time original}}{\text{Median Time new}}$$

The comparison was run over 3 model parameters, modifying on each comparison 1 parameter and leaving the other 2 constant. The parameters are: **K**, number of submodels; **N**, number of samples; **P**, number of features. The value of the parameters when not considered as variables are: $\mathrm{K} = 3$, $\mathrm{N} = 200$ and $\mathrm{P} = 100$.

It can be seen in Figure 8 that there is always a gain in running speed when using the new implementation - for all the conditions the ratio is over 1. This gain diminishes for more complex models, this characteristic is valid for all three parameters, but the difference is more drastic with the variation of the total number of samples.

For conditions with simple models (i. e. $\mathrm{N} \leqslant 100$ & $\mathrm{P} \leqslant 50$), the gain was over 10 times (results not shown). These results point out an overhead in the implementation, possible due to memory movement between the R data structures and `C++`. Additionally, larger models were considered and tested to study the possible convergence of the

1 https://cran.r-project.org/web/packages/microbenchmark/index.html

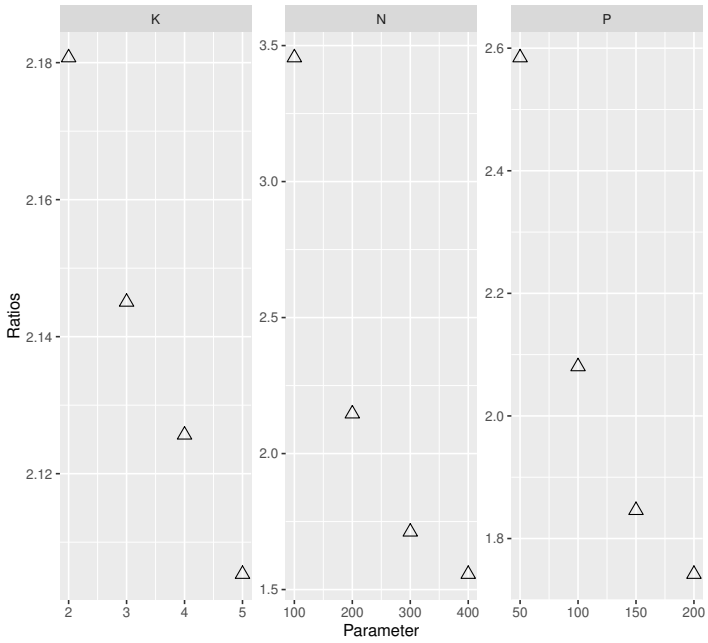


Figure 8: Performance comparison between the R and the C++ implementations of the FMLR algorithm. The ratio is defined between the median run time of the R version and the the C++ version. The comparison is performed over **K**: number of submodels, **N**: number of samples and **P**: number of features.

ratios (results not shown) for high values of the parameters. Firstly, the number of submodels were tested with values between 50 and 100. It was found that the ratio stays stable at 2,00. In the case of the number of samples, sample sizes of 800 and 1 000 were tested and for both cases the ratio was $> 1{,}2$. Finally, the number of features were tested for several values between 200 and 4 000. This was done due to significant differences for each tested case. A stable ratio value was found for $P > 1\,000$ which is around 1,00. Suggested solutions and discussion can be found in the DISCUSSION.

Our implementation can be found in the project's GitHub repository (Campos-Valenzuela, 2015). Which is open and publicly available and any improvement to the code is welcome.

4.3 MFLRMP APPLICATION, GENERAL RESULTS AND ANALYSIS

The MFLRMP algorithm was applied over all the models of the genes with expression data, as explained in Section 3.4.1.2. In total 10 247 different gene-models were trained independently, which comprises around of 85 % of the total number of gene-models.

4.3.1 *Set up and execution of the MFLRMP algorithm*

***K:** number of subpopulations or submodels*

For each of the models to train a set of parameters must be set beforehand. In particular, for mixture models the hyperparameter **K** is unknown and several values are used and compared. In a similar fashion the hyperparameter λ, which governs the strength of the penalization term, is also unknown and must be selected over a set of possible values through a comparison of cross-validation errors. Other parameters must be set as initialization values for the EM algorithm. These parameters are: the initial posterior probabilities for each data point on each model (also known as responsibilities) and, the variance of the error for each one of the **K** subpopulations.

The values chosen for the parameters are:

K: The values range from 1 to 5, considering previous results in the literature (Verhaak et al., 2010).

λ: Six values were considered and they range from 1 to 20 following an exponential profile. The values are: 1,00, 1,82, 3,31, 6,02, 10,96 and 19,95.

INITIAL POSTERIOR PROBABILITIES: The heuristic method presented in Städler et al. (Städler et al., 2010) was followed, where for each sample one of the K submodels was randomly given high probability ($\sim 80\%$).

VARIANCE OF ERROR: The variance was set as 0,5 for all the models, following Städler et al. (Städler et al., 2010) as well.

4.3.2 *General results obtained by the execution of the MFLRMP algorithm*

In this section, the results obtained from the application of the MFLRMP algorithm over the gene-models are shown. This section focuses on the results of the algorithm and not on their biological implications. The latter are shown and analyzed in the following chapter.

The algorithm was executed 10 427 times, once for each model. The selection of the hyperparameters **K**, number of subpopulations, and λ, strength of the penalization term, was performed for each run.

4.3.2.1 *Number of subpopulations for each gene-model*

The number of optimal subpopulations was found following the methodology presented in Section 3.4.4. The distribution of the number of subpopulations against the number of covariates is presented in Figure 9. The main hypothesis, which is based on the curse of dimensionality, is that for models with a large number of covariates the number

of subpopulations is going to be one. This hypothesis is due to the fact that in high dimensions any model can explain the data with a similar error, because with this configuration distances and errors lose their significance.

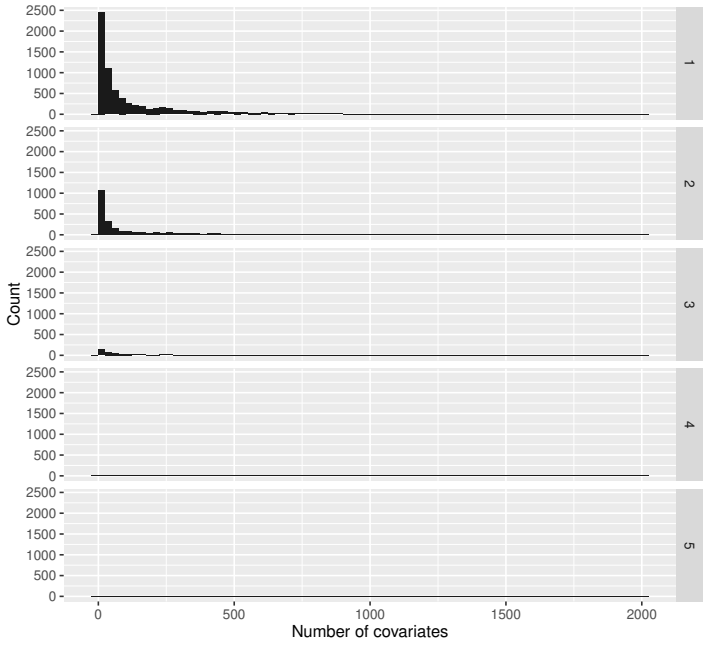


Figure 9: Distribution of the number of gene-models' covariates grouped by the number of subpopulations. Number of covariates obtained from the biological prior knowledge.

Most of the models present no subpopulations, which is concordant with one of our main biological hypothesis, where only a subset of genes present aberration in their regulation and other interactions, but for the most of them a single subpopulation close to their normal interactions should be found.

From the histograms, can be seen a negative correlation between the number of covariates and the number of subpopulations. This effect can be observed in the gene-models with 2 or more subpopulations, where the number of covariates decreases for gene-models with larger number of subpopulations.

The most frequent case of multiple subpopulations are the models with 2 subpopulations, which corresponds to ~ 25 % of the total number of models trained and over 80 % of them have less than 100 covariates. This confirms that the models with multiple subpopulations are heavily skewed to the lower range of their number of covariates.

4.3.2.2 λ *and penalization strength*

Following the iterative processes defined to find the optimal λ for each possible number of subpopulations (Section 3.4.3), several values of

λ were used. Here the optimal λs found are analyzed and the most significant result presented.

COVARIATES DISTRIBUTION FOR λ VALUES The distribution of the number of covariates for the selected λs is shown in Figure 10.

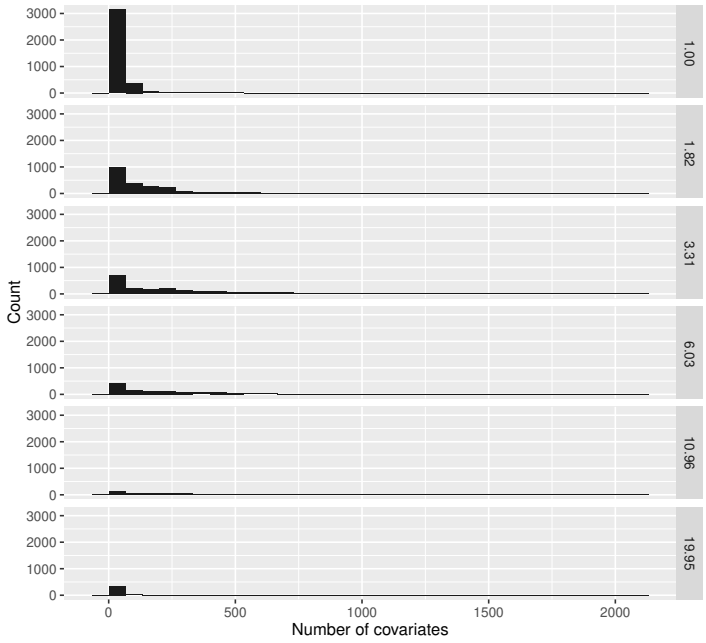


Figure 10: Distribution of gene-models' covariates segmentated by the selected λ. Number of covariates obtained from the biological prior knowledge.

Two results stand out from their distribution: the first one is the range of features with the smallest λ, this range is short in comparison with the other distributions and is concentrated under the 200 covariates. The second one is the preference for models with a large number of features for λs over 3.

These two results cannot be analyzed by themselves, because we are not applying a penalization to a simple linear model, we are applying it to a mixture of regression models. Due to that characteristic, the number of subpopulations are critical in the analysis, as shown in the following section.

One important element to highlight is the frequency of the highest λ term, which had a small preference overall, and how the number of models with a large number of covariates decreases after $\lambda = 6{,}03$ for larger λs.

These results suggest that the set of possible λs is adequate and there is no need of larger λs to consider.

DISTRIBUTION OF λS FOR DIFFERENT NUMBER OF SUBPOPULATIONS The frequency of the selected λ for the number of subpopulations (**K**) is shown in Figure 11.

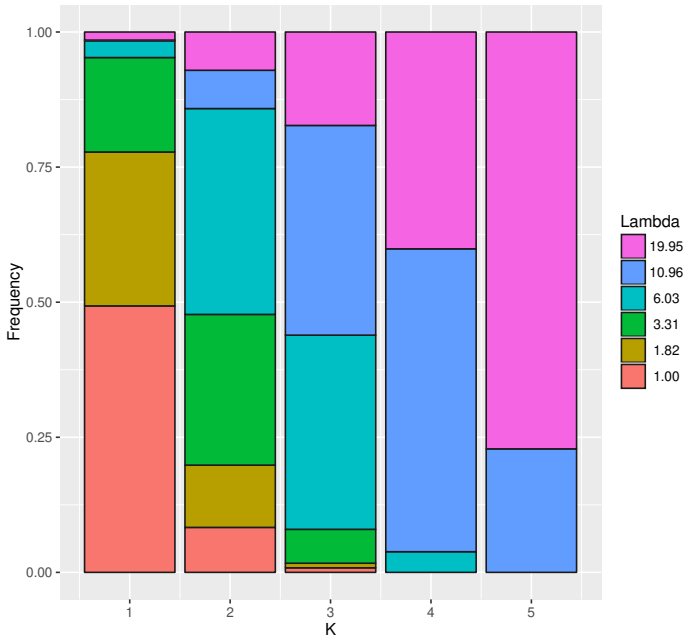


Figure 11: Mosaic plot of the distribution of the selected λs for each number of subpopulations (**K**) after the application of the MFLRMP algorithm to the gene-models.

The main results arise from the comparison of the distribution at the highest possible λ and the lowest. For $\mathbf{K} = 1$, the smallest λ is selected with the highest frequency, which is the opposite for $\mathbf{K} = 5$, where the highest λ is selected. The rest of the λ values follow a similar distribution, small λs appear more frequently for models with smaller number of subpopulations and the reverse is observed for the models with many subpopulations.

A larger λ is related to a higher penalization and thus a smaller set of non-zero coefficients. Models with many subpopulations could have a smaller cross-validation error with larger penalizations due to the unique profiles of the subpopulations, which would be drastically different between each other by the features with non-zero coefficients. In the case of models with a single population, the selection of the smallest λ can be interpreted as a small variation in the error due to the number of features selected, and thus, an stable unique population model.

INTEGRATION OF THE DISTRIBUTION ANALYSES Considering all the previous results one could see a strange disagreement between them. On one hand, the gene-models with many covariants tend to have one subpopulation, but at the same time most of the

gene-models with a large number of covariates have a large λ selected. Which is not in concordance with the mosaic plot just presented. Where for gene-models with small number of subpopulations, the strength of penalization is low.

This disagreement is due to how we analyze the distribution, we tend to focus in the small but interesting results, such as the distribution of the subpopulations for gene-models with high number of covariates. While our general analysis is correct and large gene-models tend to have small subpopulations, we must not forget that gene-models with less than 250 covariates are the most common by a large margin, and are highly present in the distribution of gene-models with one subpopulation.

Thus, an analysis focused in the distribution of gene-models with a large number of covariates can be unclear by itself due to the important presence of gene-models with few covariates, and the study taking into account only the number of subpopulation and penalty strength should be favored.

PENALIZATION STRENGHT ANALYSIS An interesting set of results is the relationship between the selected λs and the impact of them in the models.

As a first approach to this analysis, the number of features (covariates) with non-zero coefficients (not penalized) for the gene-models segmented by the selected λ is shown in Figure 12.

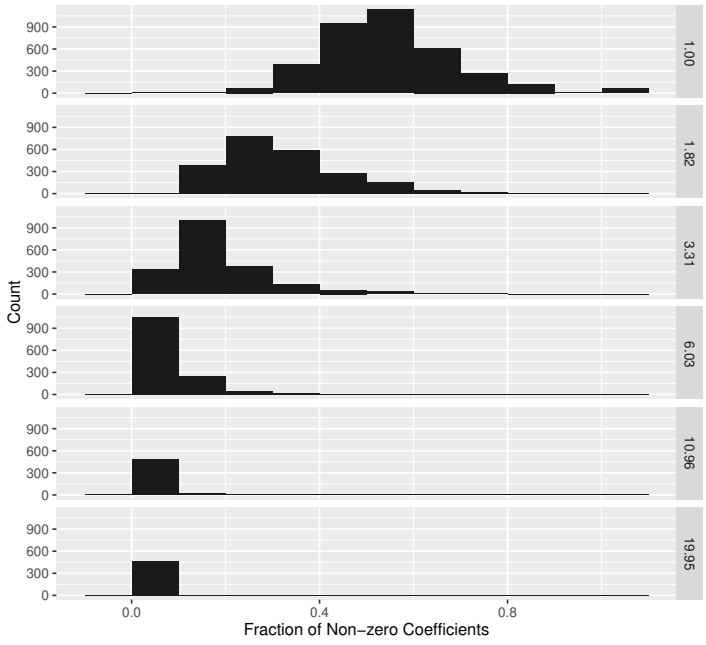


Figure 12: Distribution of the fraction of covariates with non-zero coefficients for all gene-models against the penalization term λ.

The immediately observable result is the shift to lower values of the fraction of features with non-zero coefficients when the selected value of λ grows, which can be taken as a proof of concept of the impact of the penalization term in our model.

One interesting case is the distribution of fractions of non-zero coefficients for the smallest λ, where their mean is around 0,5 and the distribution is symmetrical. This means that with a small penalization term we are able to eliminate 50 % of the features for most models, and can be interpreted as half of the covariates have a minimal significance in the definition of the subpopulations.

An additional analysis of the penalization is the study of the impact in the biological layers.

Previously, in Figure 7, the source of the covariates in the design matrix was shown. One approach to study the penalization is to see how it impacted the different sources (biological layers) of the design matrix and to compare the fraction of covariates from a source before and after the penalization. In Figure 13, the source of the covariates after the penalization is shown.

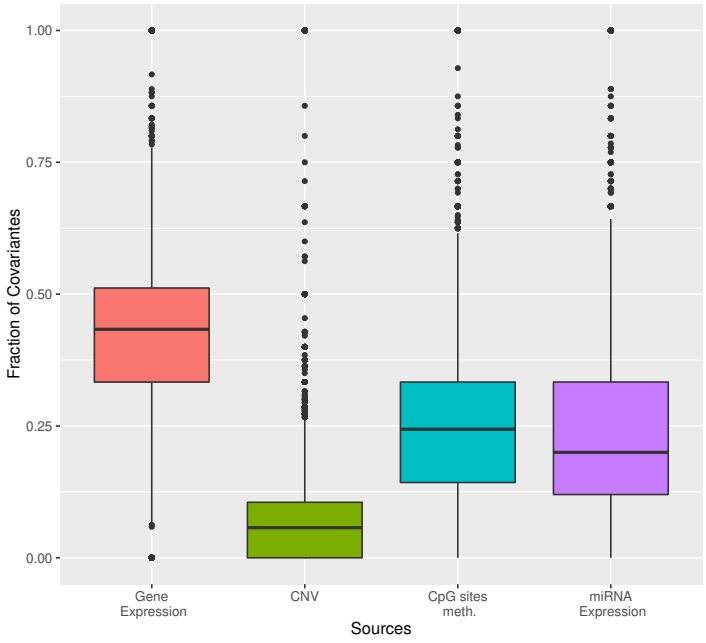


Figure 13: Distribution of the covariates sources for the gene-models. Considering biological prior knowledge and after the application of the penalization. Only features with non-zero coefficients are shown.

The most drastic change is seen with the CpG sites methylation layer, where it changes from being the most important source of covariates for most gene-models (over 50 %) to the second most important source of covariates, tied with the microRNA expression layer. This effect could have been induced by the binary profile of the methylation or

the high correlation of the CpG sites, and thus, the elimination through the *lasso* method.

For the other layers the change was less noticeable. For the expression layers, genes and microRNA, an increase in the proportion of covariates was found, which could be induced by the diminishing number of features from the methylation layer. A particular result, which was presented previously but with less impact, was the microRNA layer as a source of covariates. The number of covariates from the microRNA layer after the penalization is smaller, but comparable to the one from the gene expression layer even though the number of possible covariates is much smaller. The reason for this result is the number of interactions between the small microRNA set and the gene set. With almost 5 000 microRNA-gene interactions each element of the microRNA has a high chance to appear in the design matrix of any gene-model. The CNV layer shows no large variation, its proportion is still small, but it has been benefited by the hard penalization of the methylation layer.

Genes: 12 000 covariates
microRNA: 244 covariates

Finally, in Figure 14, another angle to study the penalization is presented. Here the fraction of features with non-zero coefficientes in a gene-model for each layer is shown. A fraction close to 0 shows that most of the features of a layer in a gene-model have a zero coefficient, while a fraction close to 1 means that most of the features of that layer have a non-zero coefficient. This is applied over all the gene-models and grouped by layer. The goal of this plot is to connect Figure 7 and Figure 13.

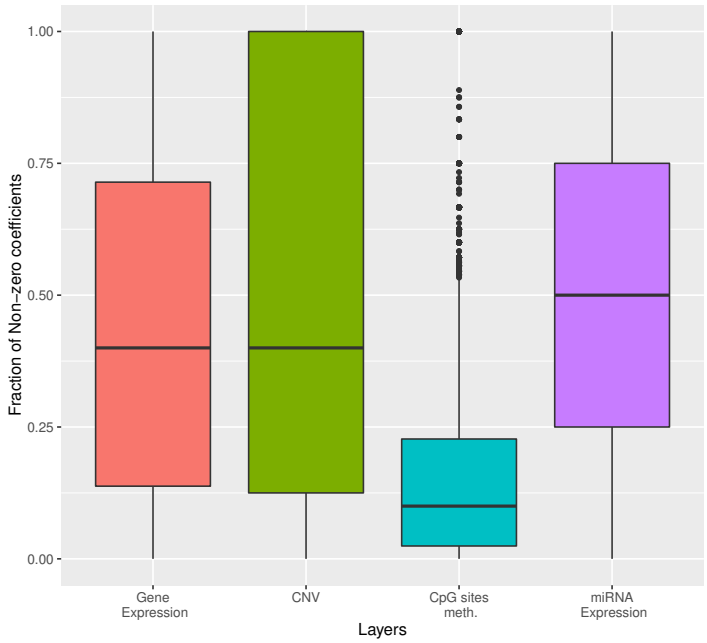


Figure 14: Fraction of the features with non-zero coefficients for each layer after the penalization considering all the gene-models.

It can be seen for the methylation layer that the fraction of non-zero coefficients is close to 0, which means that for the most gene-models the majority of features in this layer have a zero coefficient. This explains its drop in the previously presented plots.

For the other layers there is a large variance but the median of the distribution is around 50 % for all of them. This plot also shows that the penalization is heavily model-dependent and there is no preset favored layer.

COMPARISON NUMBER OF COVARIATES BEFORE AND AFTER PENALIZATION The distribution of the amount of features against the number of subpopulations is shown again in Figure 15, with the addition of the distribution of the number of features with non-zero coefficients. This comparison helps to visualize the strength of the penalization on the models. It can be seen how the number of features is drastically reduced, especially for models with a large number of features.

In all the distributions there is no model with over 200 features with non-zero covariates (dashed line) in Figure 15.

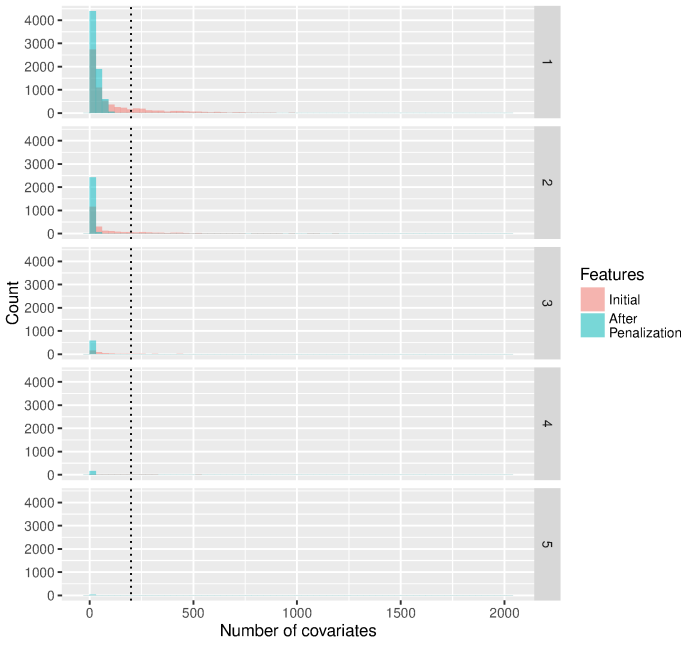


Figure 15: Comparison of distribution of original number of features against features with non-zero coefficients. Dashed line marking 200 features.

4.3.2.3 *Human Methylation Platforms and dummy variable*

In Section 3.4.1.3, the dummy variable to measure the difference in the CpG sites methylation data due to the different platforms was de-

scribed. Along with this variable its linear coefficient (α), which would represent the difference in the slope between samples measured with different platforms was presented. In Figure 16, the distribution of the α values is shown. It can be seen that the distribution is highly concentrated around 0, the 10 % and 90 % quantiles are $-0{,}022$ and 0,015 respectively. These values show that for most of the models there is no major difference between the platforms.

There are 25 outliers with an absolute value over 1, which represents a minimal proportion of the total number of coefficients α (10 212[2]). These outliers where analyzed and they corresponded to models where some samples presented an outlier subpopulation. However, the difference was not significant enough to derive another subpopulation and its difference was explained by a large coefficient α for the platform dummy variable.

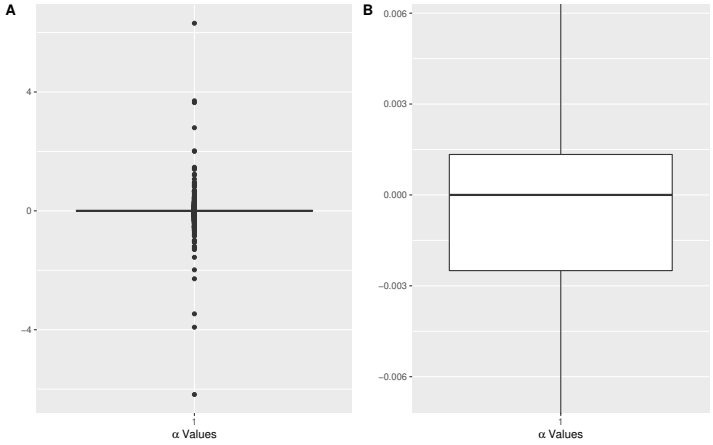


Figure 16: Boxplot of the coefficient α of the methylation platform dummy variable for all the gene-models with CpG sites methylation data. **A**: Boxplot of the total α values. **B**: Zoom into the values close to 0.

4.4 SUMMARY OF THE IMPLEMENTATION AND EXECUTION RESULTS OF THE MIXTURE OF FINITE LINEAR REGRESSION MODELS WITH PENALIZATION ALGORITHM.

Here the main results obtained from the implementation and run of the MFLRMP algorithm are shown and briefly discussed. These results are comprised by the generation of the datasets used to train the MFLRMP and the technical results from the execution of the MFLRMP algorithm.

Because the obtainment and preprocessing of the datasets were focused in the METHODOLOGY chapter, the attention in this chapter was put in the conformation of the gene-models to be used with the

2 Not all the gene-models comprise CpGs methylation data

MFLRMP. After considerations based on the number of covariates, the number of gene-models to train was set in 10 247. By utilizing the biological prior knowledge, the number of covariates for the model was modified and for most of them the number of covariates was less than 500.

These datasets were used as input for the MFLRMP algorithm, where the models were trained independently. One important characteristic of the MFLRMP algorithm is the inclusion of the *lasso* penalization term.

The effect of this penalization can be seen in a plethora of results. The most significant one is the further reduction in the number of covariates for the gene-models. Before the training of the models and the use of the *lassso* penalization, most design matrix were comprised of less than 500 variables. This number was reduced drastically after the penalization, where no models had more than 200 variables with non-zero coefficients.

Another effect of the penalization was the distribution of the source of covariates, where for the models before the penalization there was a significant presence of the CpG sites methylation layer. This layer was heavily penalized by the *lasso* penalization, while for the expression layers (genes and microRNA) the fraction of the source remained constant, resulting in these 3 layers having similar amount of non-zero coefficients. The CNV layer had a minor presence in the covariates of most of the models and this characteristic did not change after the penalization.

Along with the penalization the amount of subpopulations found with the algorithm was analyzed as well. It was found that the vast majority of gene-models presented only one subpopulation and for the case of multiple subpopulations, the most common number of them was 2.

The study of the penalization strength (λ) and the number of subpopulations (**K**) showed that for gene-models with a single subpopulation the smallest penalization terms was mostly selected, while the models with 4 or 5 subpopulations higher λs were selected. This relationships was considered to reflect the uniqueness and well-defined of multiple subpopulations, given their small number of covariates with non-zero coefficients.

Finally, it was found that the use of different platforms to measure the CpG sites methylation did not show any effect and was considered not significant for the analysis.

In conclusion, the results obtained from the execution of the MFLRMP algorithm showed that there are a defined set of gene-models with multiple subpopulations that have been heavily penalized. In addition, the 'omics' layers presented a similar number of non-penalized number of covariates, with the exception of the CNV layer. With these results we can conclude that the gene-models definition, the application of

the algorithm and the penalization, and subpopulation discovery were successful.

5

BIOLOGICAL RESULTS AND ONCOLOGICAL IMPLICATIONS OF THE TRAINED MODELS

The general results of this project are divided into those focused on the technical implementation and application of the MFLRMP algorithm with GBM data, and those concerning the underlaying cancer biology and are focused on the patients and the different molecular elements.

In this chapter, the latter results and analyses are presented. These are separated into three sections: clustering of the GBM patients, patient-wise analysis of the clusters and gene-models, and features enrichment of the clusters.

The first one refers to the clustering of the patients using the resulting gene-models from the MFLRMP, the second section is comprised by the analysis of the patients in each cluster, their clinical profiles and other analyses. The last section is focused on the significant gene-models for each cluster and their features.

5.1 ANALYSIS OF CO-OCCURRENCE AND CLUSTERING OF PATIENTS

In this section, the implications of the trained gene-models on the patients is analyzed. Firstly, the co-occurrence of the samples in the different gene-models is studied and used to define subgroups of patients. Secondly, the different subgroups obtained are studied based on their clinical characteristics and compared to previously reported results.

5.1.1 *Calculation of the co-occurrence value between samples*

A key set of values defined by this work is the sample co-occurrence probabilities. These values were defined in order to compare the co-occurrence of two samples over all the trained gene-models subpopulations (see Section 3.5.1).

These probabilities permit the comparison of the closeness between each sample with the others samples and thus, generate groups of them. Because of the probabilistic nature of the finite mixture model, these values represent the chance that a pair of patients are together over all the gene-models and are not distances in the mathematical sense.

From the 10 427 gene-models trained, 3 281 have more than 1 subpopulation. From these gene-models an additional selection process

was performed in order to choose only the gene-models without a dominating subpopulation, see Section 3.5.2. In total 2 406 gene-models were selected and used in this analysis.

The co-occurrence values, also called probabilities, for each pair of samples over each gene-model was calculated. This resulted in over 50 000 values for each gene-model. The total co-occurrence probability for each pair was calculated by aggregating the values by the mean over the gene-models, this value represents the probability of co-occurrence over all the models.

Total number of pairs: $\frac{N*(N-1)}{2}$ *with* $N = 324$

Finally, the co-occurrence matrix was created with the total co-occurrence values with dimensions $n \times n$, where N is the number of patients ($N = 324$).

The values in the matrix were linearly transformed as detailed in Section 3.5.2. In Figure 17, a histogram of the data before (**A**) and after (**B**) the transformation is shown.

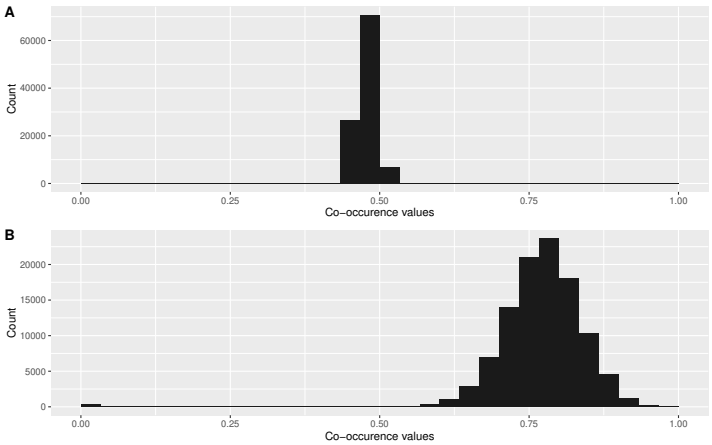


Figure 17: Histogram of the co-occurrence values for each pair of samples before and after transformation for easier discrimination of agglomerated elements.
A: raw values. **B**: Values after transformation.

As can be seen, the transformation allowed us to easily discriminate between the values that were agglomerated before. This process transforms the probabilities to pseudo-distance, where a high certainty has a value close to 1 in the former, while in the latter is close to 0.

5.1.2 *Hierarchical clustering of the co-occurrence matrix*

After the transformation, the co-occurrence matrix was ready to be used as a distance matrix for the agglomerative hierarchical clustering algorithm, which was applied using complete linkage.

Complete linkage: method to combine clusters where the distance between two groups is defined as the farthest distance of elements between groups.

The resulting dendrogram of the hierarchical clustering along with the silhouette values were calculated and analyzed. By visual inspection, it was found that the set of patients could be divided into two.

This clustering was by far the strongest but it left a significant portion of samples agglomerated and without a possible profiling, while the second group was small and with a small distance between members.

Due to this, it was decided to select a larger number of clusters, 6, as the optimal value. By doing this the original well-defined cluster was kept and the samples were divided into sparse groups and the number of samples are equally distributed.

Table 5: Number of elements for each cluster.

Cluster	Size
1	35
2	44
3	52
4	58
5	85
6	50

In Table 5 the size of each cluster are shown. The stable and well defined cluster found previously when $K = 2$ is denominated Cluster 1 for future reference.

The histogram of the silhouette values for different number of clusters can be found in Section B.1, and in Figure 18 the heatmap of the distance matrix is shown along the with the discovered clusters.

5.2 PATIENTS-WISE ANALYSIS OF THE CLUSTERS

After obtaining the clusters for our set of patients, the next step was to profile the samples belonging to each one of them. This process comprised the analysis of the clinical profiles, comparison to previously discovered clusters and the analysis of the mutation profiles.

5.2.1 *Covariates source for each cluster*

With the clusters of patients calculated, it was possible to reanalyze the source of the covariates considering the subpopulations. As shown in the example in Figure 4, each cluster of patients are grouped over one subpopulation for each gene-model. It is then possible to study a cluster of patients by selecting those subpopulations and the features that define them.

Figure 19 presents the fraction of covariates from each source (biological layers) segmented by cluster. This can show if there are dif-

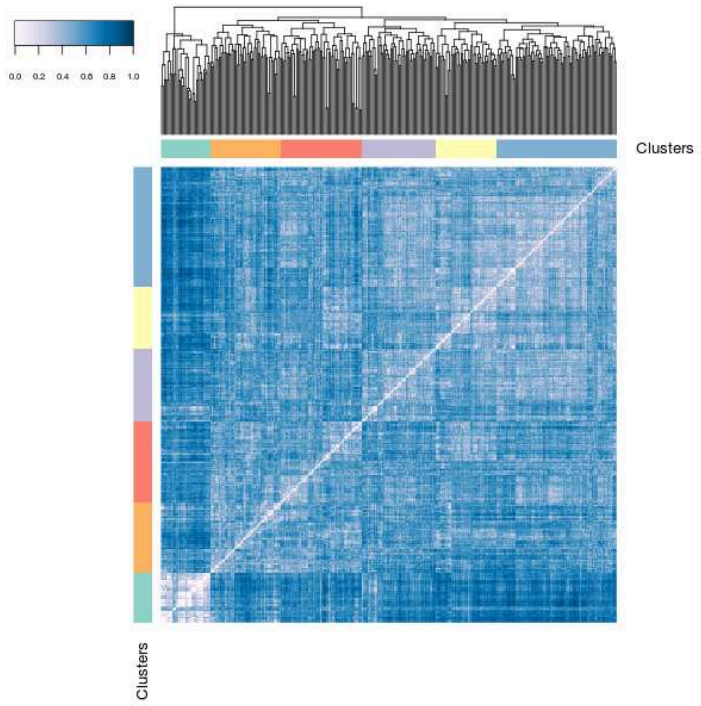


Figure 18: Heatmap of the co-occurrence matrix showing patients clusters.

ferences in the fraction of covariates from a particular source between clusters.

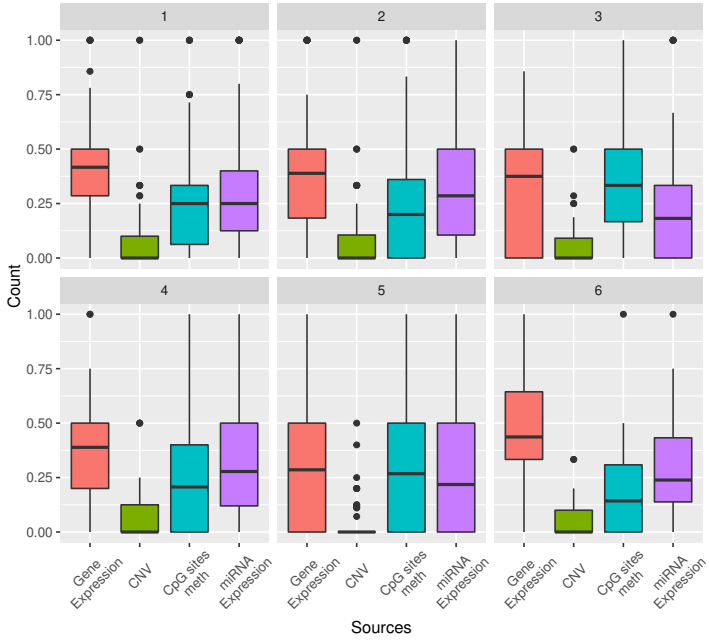


Figure 19: Fraction of the features with non-zero coefficients for each layer after penalization and considering the gene-models that belong to a particular cluster.

Analyzing the boxplots, it is possible to observe that for the CNV layer the distribution is constant over the different clusters. With the sole exception of Cluster 5 which has a very low variance, but with a similar median as the rest of the clusters.

For the other layers a constant distribution was found. In general, the gene expression layer has a larger fraction of covariates for most of the clusters while the methylation and microRNA layers have very close distributions. There is a trade-off between the methylation and microRNA layers; when the methylation value is higher the microRNA fraction is smaller and vice-versa. The only cluster that presented an unique profile is Cluster 5, where the distribution for these layers is almost identical and with very large variance. Making this cluster an interesting case to study for all the layers.

5.2.2 *Clinical analysis of the clusters*

In this section the main clinical parameters are analyzed in function of the defined clusters. The data available in the TCGA repository allowed us to study the distribution of the age, gender, survival, type of tumor, history treatment and Karnofsky score between the groups.

From the data downloaded, 3 samples were not present in the database. Due to this these samples were not considered in the clinical analysis presented here, meaning that 321 samples are present in the following analyses. The list of the missing samples can be found in Table 6.

Table 6: Missing samples in the clinical data.

Patient	Cluster
TCGA-16-1048	3
TCGA-32-2498	3
TCGA-28-2510	5

5.2.2.1 *Distribution of the patients' age over the clusters*

The distribution of the age of the patients at the time of the diagnosis is presented in Figure 20, where the values are grouped by clusters.

The distribution of the age is not homogeneous over all the clusters and it presents a skewed non-normal distribution after visual inspection. Due to this a Kruskal-Wallis rank sum test was run on the data in order to test the hypothesis that the mean is the same for all the clusters. This test reported a Chi-squared value of 14,801 and a p-value $< 0{,}02$.

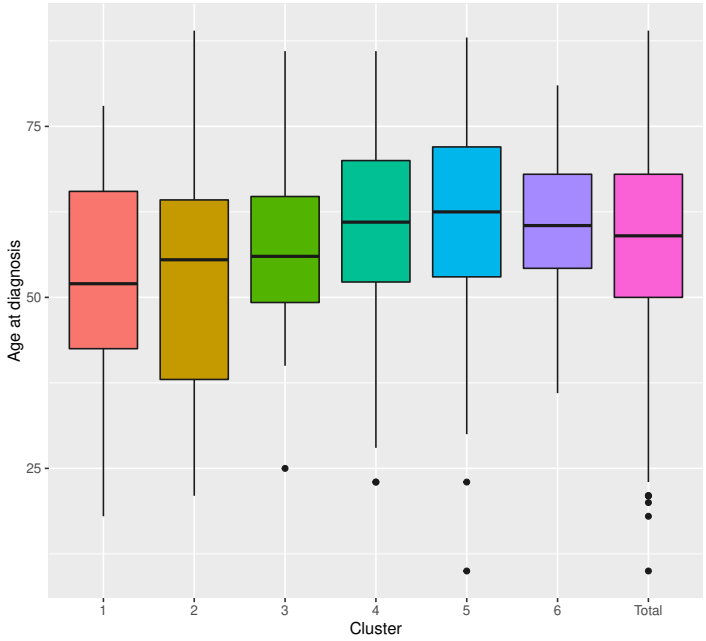


Figure 20: Distribution of the patients' age for each cluster. Clusters 1 and 2 are comprised by younger patients and present a larger variance, while Clusters 4, 5 and 6 have older patients and a smaller variance.

The difference between the means is noticeable when comparing the Clusters 4, 5 and 6, which have mostly patients of old age, with Cluster 1 and 2, which have over 60 % of the patients under 35 years old. Finally, Cluster 3 has a distribution similar to the whole-set distribution and cannot be uniquely profiled.

5.2.2.2 *Gender distribution over the cluster*

The ratio of the patients' gender for the different clusters and for all the samples is illustrated in Figure 21.

For the full set of samples, the ratio of male and female is close to 60/40, which is similar to previously reported ratios (Verhaak et al., 2010).

The gender ratio varies over the different clusters, in particular in Clusters 1 and 5, but the difference appears not to be significantly large. With the intention to test this, a Pearson's Chi-squared Test is applied over the data. The resulting p-value of the test is $> 0{,}06$ and does not allow us to reject the hypothesis that the number of samples for each cluster and gender does not have a significant difference from the expected value.

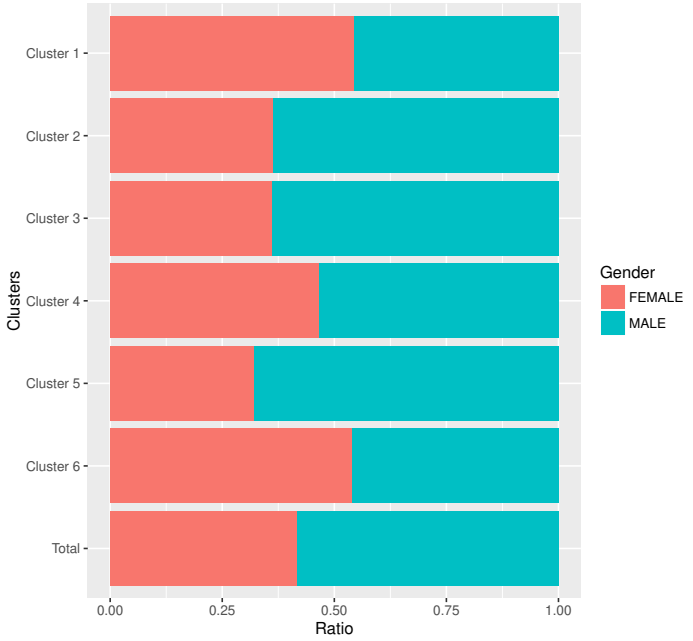


Figure 21: Ratio of males and females for each cluster and for all samples.

5.2.2.3 *Analysis of the number of de novo samples*

Because GBM tumors can arise from lower-grade gliomas or as de novo tumors, the analysis of the type of tumor is in our interest.

From the clinical data is not possible to distinguish between secondary and recurrent tumor samples due to the annotation. Because of this characteristic, these two groups are classified as one, named non-de novo tumors.

Table 7: Distribution of non-de novo and de novo samples for each cluster.

	De novo tumor		
Cluster	False	True	Total
1	8	27	35
2	4	40	44
3	7	43	50
4	0	58	58
5	2	82	84
6	0	50	50
Total	21	300	321

In Table 7, the distribution of de novo and non-de novo samples for each cluster is shown. The de novo samples are over 90 % of the total. This difference can be seen in previous works, for example in Verhaak et al. (2010) there were 19 of 202 samples with this characteristic.

The distribution of the non-de novo samples is concentrated in the Clusters 1, 2 and 3. In particular, Clusters 1 and 3 have the most of the non-de novo samples. In Verhaak et al. (2010) 3 of the 5 secondary tumor samples are in the Proneural subtype and the other 2 in the Classical subtype, while the recurrent tumor samples are distributed fairly over the 4 subtypes.

The distribution of the non-de novo samples shows a stricter enrichment in our study, while considering secondary and recurrent tumors, compared to previous studies. This attribute allows us to characterized Clusters 1 and 3 as the groups where the non-de novo tumors are present and could exhibit similarities with the Proneural and Classical subtypes of Verhaak et al. Because of this, additional comparisons between these groups and subtypes are of interest and are presented in the following sections.

5.2.2.4 *Analysis of the Karnofsky performance score of the tumor samples*

The Karnofsky performance scores were reported for 245 of the samples used in this project. Because of the qualitative nature of the score, its distribution over the clusters is analyzed instead of using it to calculate different statistics.

Figure 22 presents the mosaic plot of the scores over each cluster and all the samples, where the ratio of each score over the total number of elements is shown.

The distribution of scores over the Clusters 1 and 2 presents a large ratio of high-level scores and a lack of any score lower than 60. This profile contrasts with the distribution for the Clusters 4, 5 and 6, where the ratio of higher scores is lower and there is a significance presence of low-level scores. For Cluster 3 it can be seen that the distribution is close to the distribution for all the samples.

Due to the qualitative nature of the Karnofsky values, a Pearson's Chi-squared Test for a contingency table was applied to them in order to test a difference between the observed and expected distribution of the values. The p-value of the test was $< 0{,}04$ and thus, we can reject the hypothesis that the scores follows the expected distribution as described in 3.6.1.

5.2.2.5 *History of neoadjuvant treatment analysis*

The history of administration of treatment before surgery, such as chemotherapy or hormone therapy, is presented in Table 8.

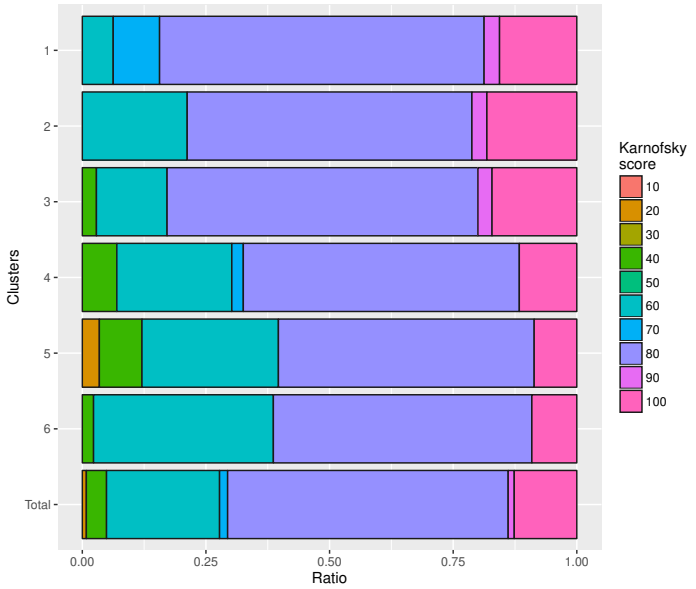


Figure 22: Mosaic plot of the Karnofsky scores for each cluster and the whole set of patients.

Table 8: History of neoadjuvant treatment for each cluster.

	Neoadjuvant treatment		
Cluster	False	True	Total
1	26	9	35
2	43	1	44
3	44	6	50
4	57	1	58
5	82	2	84
6	50	0	50
Total	302	19	321

While the distribution of patients without treatment is similar over all the clusters, the patients with neoadjuvant treatment have a significant presence only in Clusters 1 and 3.

The results presented in Table 8 recall the results previously presented in Table 7 where Clusters 1 and 3 had most of the non-de novo samples. This correlation can be a consequence of the type of the tumor, due to the disposition of secondary tumors for diffusive infiltration (Kleihues and Ohgaki, 1999), where neoadjuvant treatment would be preferable over surgery.

5.2.2.6 *Survival analysis of the patients*

The survival analysis of the samples grouped by their cluster was performed. For 320 patients, there was survival information, including censored data. The 4 missing patients were comprised of the 3 samples without any clinical information identified in Table 6 and in addition, a sample without vital status information, TCGA-16-0861.

The survival curve over the clusters is shown in Figure 23. A red cross is used to represent censored data points.

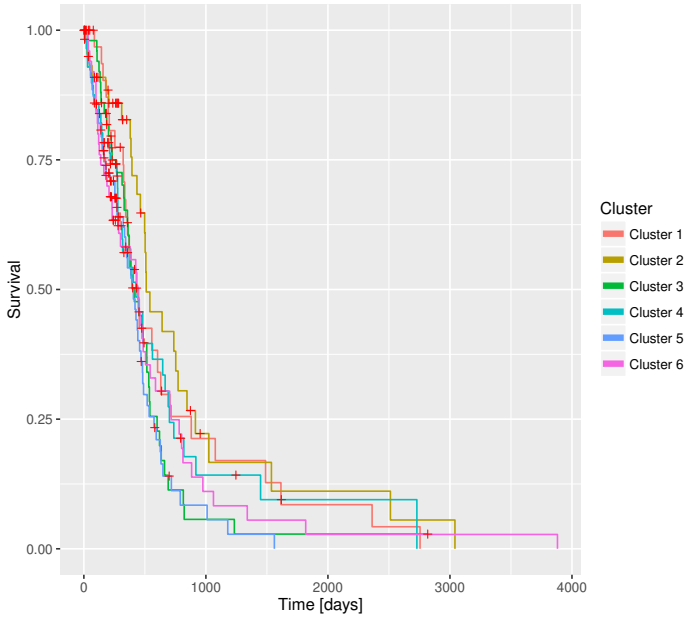


Figure 23: Survival curves of the patients grouped by clusters.

The survival curves were tested using the log-rank test if they have a significance difference. The test returned a p-value $> 0{,}08$ and thus, we cannot reject the hypothesis that they have the same survival function.

An inspection of the curves shows that the curve for Cluster 2 has a higher survival rate overall. After day 1 000, Cluster 1 joins Cluster 2 to form a group with higher survival, which is in accordance with the results showed for their age and Karnofsky scores, making this group of clusters an interesting target of study, especially when contrasted with Clusters 4 and 5, which have a different composition with respect to non-de novo tumors, age distribution and Karnofsky scores.

5.2.3 *Comparison to Verhaak subtypes*

The clusters of patients found using the finite mixture approach can be compared to those reported by Verhaak et al. (Verhaak et al., 2010).

The main difference between them was the reported number of clusters or subtypes, as denominated in Verhaak et al, where as described in the introduction, 4 subtypes were discovered, while our findings found 6 different clusters.

A direct comparison of the clusters was possible due to the use of TCGA data by both studies. A comparison of the size of the datasets is shown in Table 9. It can be seen that this study uses 75 % more samples than the study from Verhaak et al., and when the sets are intersected, only 50 samples are found.

Table 9: Number of samples for the Verhaak and Campos studies.

Study	Number of Samples
Verhaak et al.	202
Campos	324
Intersection	50

This difference comes from the need to use samples in this study. From the original 558 samples in the expression set (Table 1), only 324 have data for the other molecular layers and most of them without this information are found in Verhaak et al.

The comparison between the sets for the 50 shared samples can be seen in the Table 10, where a contingency table shows the distribution of the patients over the subtypes and clusters.

Table 10: Contingency table of the shared samples and their clusters between Campos and Verhaak 2010.

	Verhaak et al. subtypes				
Campos clusters	Classical	Mesenchymal	Neural	Proneural	Total
1	6	3	2	9	20
2	0	0	1	5	6
3	2	11	4	0	17
4	3	0	0	2	5
5	0	1	0	0	1
6	1	0	0	0	1
Total	12	15	7	16	50

An interesting first result was the distribution of the shared samples over the clusters as defined in this work. Clusters 1 and 3 comprise

75 % of the shared samples, while Clusters 2 and 4 have 6 and 5 respectively. Finally, there was an almost complete absence of shared samples for Clusters 5 and 6, where for each one of them there was only 1 shared sample. This result could show that these clusters are new subtypes, not considered in previous studies, but due to the low number of shared samples this hypothesis cannot be corroborated.

The majority of samples for Clusters 1 and 3 were distributed over a few subtypes. The samples in Cluster 1 were allocated mostly in the Classical and Proneural subtypes (15 of 20), while 11 of the 17 samples from the Cluster 3 were present in the Mesenchymal subtype. An additional point of comparison was the survival, age distribution and tumor type found in the Proneural subtype. In this subtype an enrichment of younger patients with better survival was found and 3 of the 5 samples of secondary glioblastoma are grouped here. In comparison with the results present here, Clusters 1 and 2, and in a lesser form Cluster 3, have these characteristics. In summary, through the clinical analysis it was possible to link the Proneural subtype to the Clusters 1 and 2, and the Mesenchymal subtype to Cluster 3.

Most of the subtypes in Verhaak et el. were defined through their molecular signature, in the next section the genetic signatures are presented and analyzed. It also includes the most relevant signatures for the subtypes of Verhaak et al. and other studies.

5.2.4 *Genetic signatures in the clusters*

The analysis of the genetic signatures in the clusters was performed by analyzing the distribution of the somatic mutation in the patients and clusters. For this study, only the SNPs were taken into account, the reason behind this decision is that the mixture model already integrates the signatures of the other molecular layers, such as methylation and CNV.

5.2.4.1 *Data and missing samples*

For this purpose the database COSMIC was used. Here the somatic mutations of the TCGA consortium are stored and cataloged. The data downloaded from the COSMIC database was incomplete and 104 samples were missing. The allocation of the missing samples is shown in Table 11, where the number of missing samples for each cluster is presented along with the percent of missing samples over the total number of patients for that cluster.

From the 104 missing samples, 65 belong to the Clusters 4 and 6, comprising over 55 % of the total number of samples. In a lesser form, Cluster 2 misses information for over 30 % of their samples. This lack of information makes it difficult to analyze these groups and this handicap will be taken into when discussing these results.

Table 11: Number of samples without mutation signature for each cluster.

Cluster	Missing samples	Percent [%]
1	1	2,86
2	14	31,82
3	10	19,23
4	33	56,90
5	14	16,47
6	32	64,00

The genetic analysis was performed over all the clusters, including those with a high number of missing samples, and involved the study of the number of mutated samples on each cluster.

5.2.4.2 *General mutations analysis*

From the over 7 700 genes with somatic mutations, a selection of possible interesting genes was performed based on the minimal number of samples in a cluster with a mutation in that gene. For a gene to be considered it should present mutations in at least 6 samples in one cluster.

In total 34 different genes had mutations in at least 6 samples for any cluster and, in addition, a Fisher's exact test was performed in order to analyze the significance of the mutations. The reported p-values were adjusted using the Bejamini-Hochberg method to take into account the multiple comparison performed.

In Table 12, the 34 genes are shown along with the number of mutated samples for each cluster. The table has been colored to ease its inspection based on the number of mutated samples and the p-value obtained for the cases with more than 6 samples.

STUDY OF CLUSTER 5 An examination of the table reveals that Cluster 5 presented a high number of mutated samples. In fact, only 9 of the selected 34 genes have less than 6 mutated samples in this cluster. This highly mutated number of samples raises the question why this cluster in particular has a higher ratio of mutations per sample than the other clusters. In Figure 24 the distribution of the number of mutated samples for the Cluster 5 and for the other samples can be seen in addition to the density of all the samples.

The density for the number of mutation per sample for the Cluster 5 peaks around 100 mutations per sample, this peak is present in the

Table 12: Number of samples with gene mutations for each cluster. Colored cells represent more than 6 samples mutated in the cluster. Red cells show the level of statistical significance for that cell.

	Clusters					
Genes	1	2	3	4	5	6
ANK2	5	2	9	2	2	0
ASPM	4	2	8	0	5	0
C15orf2	0	0	0	0	7	0
CHEK2	14	5	13	2	2	1
CNTNAP2	0	1	1	0	6	1
DNAH3	1	1	4	0	6	1
EGFR	9	1	6	7	25	7
FBLN2	0	0	0	1	6	0
FLG	1	4	5	1	13	3
HIF1A	6	3	3	0	1	0
HMCN1	2	0	3	1	7	0
IDH1	0	10	1	2	2	0
IRS4	7	1	5	1	1	0
LAMA1	1	0	2	0	7	0
LRP2	0	0	2	1	9	1
MUC16	1	4	7	5	15	4
MUC17	4	0	3	0	10	3
NF1	3	1	2	1	6	0
OBSCN	2	0	3	2	7	0
PCLO	1	2	2	3	7	1
PIK3C3	5	0	6	1	1	0
PIK3CA	3	2	2	0	11	1
POTEC	0	1	2	1	6	2
PTEN	3	7	12	6	19	3
RELN	1	1	2	1	6	2
RYR1	1	1	0	2	6	1
RYR2	1	5	3	3	6	2
RYR3	3	2	3	2	7	1
SLC25A13	4	1	8	2	1	0
SPTA1	1	3	2	3	7	2
SYNE1	1	2	2	1	6	1
TP53	11	17	10	9	12	2
TTN	5	3	10	8	24	6
ZNF429	10	5	12	4	1	0

Value significance: $\geqslant 6$ samples | p-value $< 0{,}05$ | p-value $< 0{,}01$

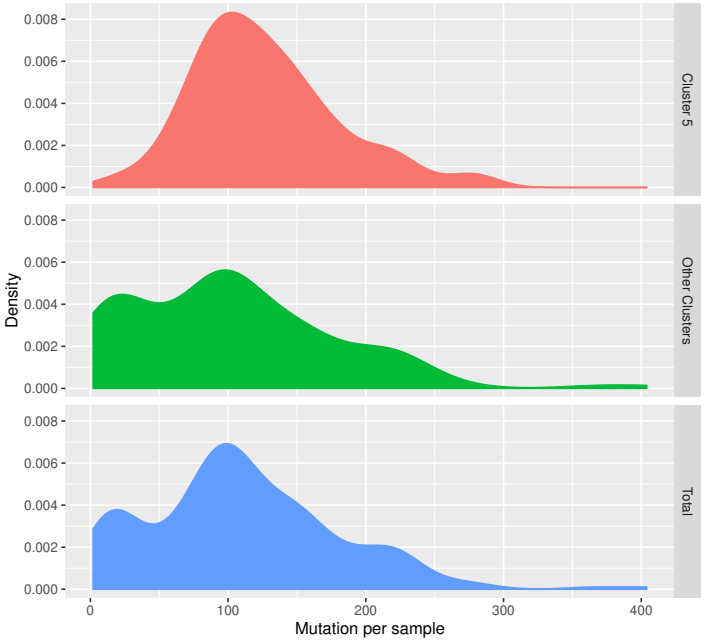


Figure 24: Density of the number of mutation per sample for Cluster 5 and the others.

rest of clusters but has a lesser presence. In fact, the density function for the rest of samples presents a first peak around 20 mutations per sample, which is completely absent in Cluster 5.

A two-sided Kolmogorov-Smirnov test was applied to compare the distribution of the number of mutation per sample between Cluster 5 and the rest of samples. The p-value obtained was $< 0{,}003$, which confirms an unique distribution of mutated samples for Cluster 5, which consist of a larger number of mutation per sample.

An interesting result for this cluster was the low number of significant genes present. Only 5 genes had an adjusted p-value under 0,05 and one under 0,01. This result is due to the large size of the cluster (over 70 samples), which makes the expected value of the test higher and thus closer to large number of mutated samples found.

For most genes with a high number of mutated samples and a low p-value, no association has been reported to gliomas with some important exception that will be discussed in the following section.

5.2.4.3 *Gene mutations in multiple clusters*

The genes EGFR, PTEN, TP53 and TTN were found with an important presence in 4 to 5 clusters.

The genes EGFR, PTEN and TP53 have been involved in several studies regarding glioblastoma (Ohgaki, 2005; Ohgaki and Kleihues, 2007) and have been pointed out as crucial biomarkers in gliomas. Particularly, mutations in EGFR and PTEN have been found mostly in

primary GBMs, while mutations in TP53 are found mostly in secondary GBMs.

On the other hand, TTN has not been directly linked to cancer development and its mutations are considered to be passengers and consequence to the large size of the encoding polypeptide (Greenman et al., 2007).

Due to the presence of mutated genes on several clusters, the significance test of these cases did not result in a small p-value, even more EGFR, PTEN and TTN had no significant cases at all. On the other hand, the importance of these genes cannot be dismissed and their influence in cluster profiling is central. Finally, in the case of TP53, there were two significant cases, Cluster 2 and 5, which could point to a possible grouping of secondary samples.

5.2.4.4 *Gene mutations in individual clusters*

For the case of genes where the mutated samples were concentrated in few clusters, they presented in general a higher significance than those with a high number of mutated samples over several cluster. This is due to the nature of Fisher's test, where the allocation of all mutated samples in one cluster represents a significant correlation between the gene and the cluster, thus a small p-value is calculated. This effect was found for most cases, but for the genes with a high presence in Cluster 5, because of the overall high number of samples in that cluster, it made the correlation weaker and thus less significance.

The genes ANK2, ASPM, C15orf2, CHEK2, HIF1A, IDH1, IRS4, SLC25A13 and ZNF429 presented the former mentioned effect.

From them, the gene IDH1 must be highlighted due to its importance in the genetics of GBMs (Ohgaki and Kleihues, 2013), in particular for its role as a potential specific marker for the secondary GBMs. This mutations was allocated almost exclusively in the Cluster 2, which along the presence of mutation in the gene TP53, opened the door to a characterization of this cluster as secondary GBM.

For the other mutated genes, a gene set analysis of these genes returned a small enrichment in the Central carbon metabolism in cancer pathway[1], with 2 genes on the list belonging to the pathway. This enrichment is expected in tumor samples and refers to the need of cancer cells to support cell growth and survival.

It was searched in literature for information on the mutated genes linked to GBM research, but the only relevant information that was found was for HIF1A (also called HIF-1α). The function of the encoded protein is to operate as a master regulator under hypoxia. This relates it to over 40 genes[2], including cell growth factors. Due to this is has

1 http://www.kegg.jp/pathway/hsa05230
2 http://www.uniprot.org/uniprot/Q16665

been indicated as possible drug target in GBM patients (Van Meir et al., 2010; Agnihotri et al., 2013).

5.2.4.5 *Comparison to subtypes signatures*

The subtypes presented by Verhaak et al. have a very particular genetic signature, due to this the comparison to the mutated profiles of our clusters is important.

The Classical subtype was defined by its CNV profile, aberrations such as EGFR amplification and PTEN loss, along the over-expression of EGFR. But no particular mutation worked as a marker for this subtype. In a similar way, the Neural subtype had no mutation marker and it was overall difficult to define and compare.

In the case of the Mesenchymal subtype, the mutations of the NF1 and PTEN genes were significant markers. In our clusters, the NF1 gene presented a low number of mutated samples in all the cluster and only over 6 mutated samples in Cluster 5, the hypermutated one, but with no statistical significance. PTEN presented a high number of mutated samples in the Clusters 2, 3, 4 and 5. But none of those cases was significant.

Finally, the Proneural subtype has important mutation in the genes IDH1 and TP53, and as discussed before can be related to the Cluster 2, due to the significant number of mutated samples for these genes.

5.3 GENE-MODELS AND FEATURES ANALYSIS OF THE CLUSTERS OF PATIENTS

With the definition of the clusters of patients and the analysis of the samples belonging to each cluster, the next step was to find the gene-models and features that give a particular profile to each cluster. This section is comprised of the selection of the significant gene-models for each cluster, the comparison of them between clusters and finally, the study of the main features of the selected models.

5.3.1 *Gene-models selection for each cluster of patients*

The first analysis performed in this section is the selection of the significant gene-models for each cluster. As discussed in Section 3.7.1, the chosen method to perform this task is the calculation of the distance between gene-models considering the co-occurrence distances for all the pairs of samples in a particular cluster. By this methodology, is possible to use hierarchical clustering to group together the gene-models with large co-occurrence values for a particular cluster. Finally, the group of models with the highest co-occurrence can be selected as significant set.

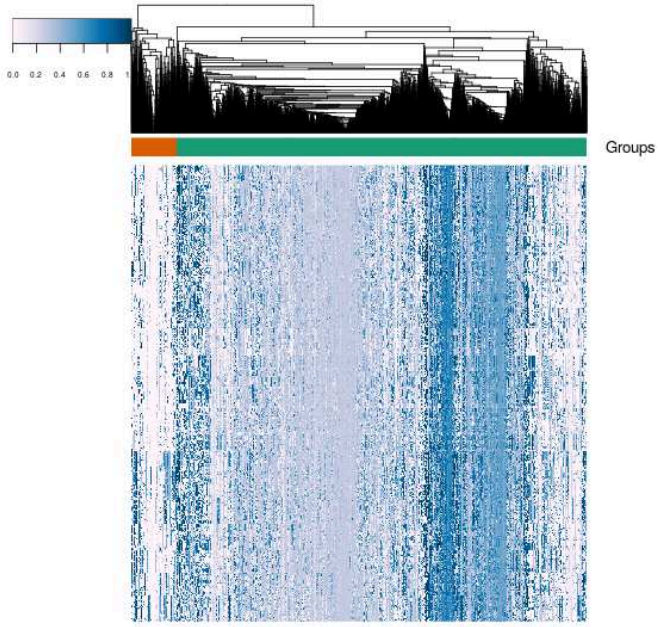


Figure 25: Heatmap of the co-occurrence distances for Cluster 1. Grouping done with K $= 2$ and the selected group is marked in orange.

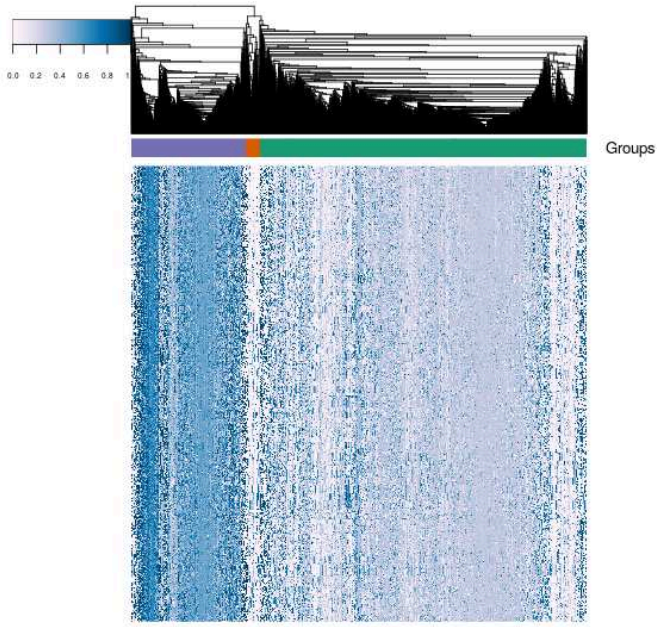


Figure 26: Heatmap of the co-occurrence distances for Cluster 5. Grouping done with K $= 3$ and the selected group is marked in orange.

For most of the clusters this grouping approach resulted into one small and well defined group with high co-occurrence values (or low co-occurrence distances) and a large group with low co-occurrence values. This lead to select **K** (number of groups) equal to 2. For the Clusters 4 and 5, the number of groups was set to $\mathbf{K} = 3$, this was done due to the presence of a cluster with only 2 models in Cluster 4 and the obtainment of a well-defined cluster with very low co-occurrence values that aggregated all the gene-models with high co-occurrence values together in Cluster 5.

In Figure 25 and 26, the dendrograms and heatmaps for the first and fifth clusters are shown. In the figures, the groups, their co-occurrence distances in the heatmap and the distance of their members on the top dendrogram can be seen. Additionally, is possible to spot that for each example, one group is distinctly small and with high co-occurrence values. In both cases this is represented with the color orange.

Table 13: Number of selected gene-models for each cluster of patients

Cluster	Number of gene-models
Cluster 1	242
Cluster 2	267
Cluster 3	54
Cluster 4	68
Cluster 5	70
Cluster 6	41

The number of gene-models selected by this method for each cluster of patients is shown Table 13. For the first two clusters, the number is over 200, which corresponds to ~10 % of the total number of gene-models. For the rest of the clusters, this number decreases to ~3 %. This process permits us to have a unique profile of gene-models for each cluster and, as presented in the next section, to have a low intersection of gene-models between clusters.

Heatmaps for the additional clusters of patients can be found in the appendix Section B.2.

5.3.2 *Analysis of shared gene-models*

With the selection of a set of gene-models that are significant for each cluster, it was possible to study the features and signatures for each set of models.

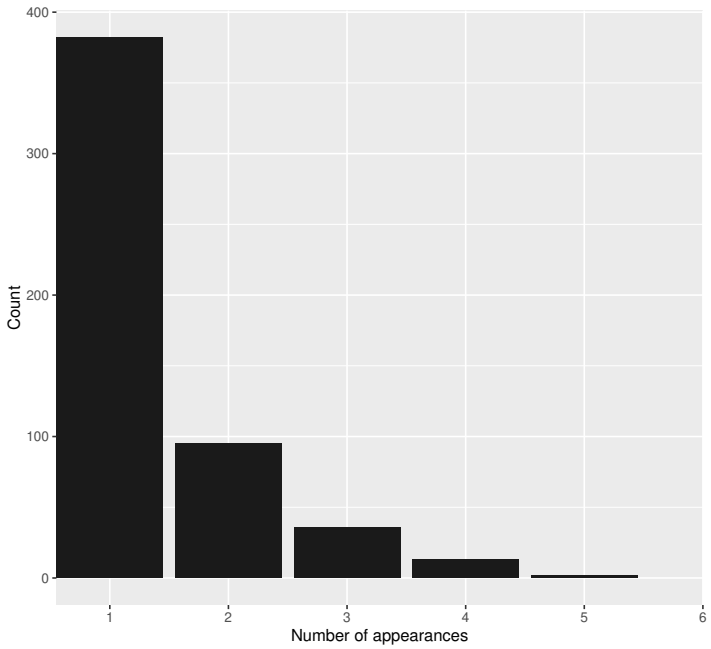


Figure 27: Distribution of the number of appearances of gene-models in the different patients clusters

In the first place, the comparison of the selected gene-models between clusters, also called inter-cluster analysis, was performed. The main question to answer is: are there shared gene-models between clusters and, if so, how different are the selected subpopulations for them on each cluster. It must be noticed that even in the case that two clusters share a gene-model the patients of each cluster might belong to different subpopulations on that gene-model and so the features with non-zero coefficients can have differences between them.

The histogram in Figure 27 shows the distribution of the gene-models by the number of clusters where they appear. In a first look it is possible to observe that the number of shared gene-models is extremely small. There are no gene-models shared by all the clusters, only 2 of them are shared by 5 clusters and 13 for 4 clusters.

In Table 14, the gene-models with 4 or 5 appearances are shown. An enrichment analysis of the target genes of these gene-models showed that there are no pathways enriched, and that only the Gene Ontology (GO) biological process Transmission of nerve impulse (GO:0019226) showed a significant enrichment (adjusted p-value $< 0{,}05$). As the name states this biological process is involved in the transmission of signals through the nervous system, which is expected in samples of brain matter but opens an analysis approach about the effect of the disease in the synapses and potential transmission in patients.

The subpopulations of the gene-models with 4 and 5 appearances for each cluster of patients are shown in Table 15. These values show

Table 14: Gene-models with 4 or 5 appearances in the patients clusters.

Gene-model	Number of Appearances
DPP6	5
STMN4	5
ABAT	4
BAI3	4
CHST1	4
FAM5B	4
FCGBP	4
GAS2	4
GPM6A	4
GPR56	4
KLHL9	4
LGI1	4
NOVA1	4
PGBD5	4
SCN3A	4

on which submodel the patients of the cluster belong to a particular gene-model. For some clusters there are no subpopulations for that gene-model, this is annotated as NA. If different clusters of patients are in the same submodel then they will share the same coefficients for their covariates. The first characteristic observed is the low number of gene-models present in Cluster 6, this hints at the possibility that this cluster has few similarities with the others.

On the other hand, the high number of non-NA values for the Clusters 1 and 2 can be explained due to their large number of gene-models (> 200), making it more probable that they present shared gene-models than the other clusters.

A comparison of the values over the different clusters shows a distinctive difference between Cluster 1 and the other clusters. In 12 of the 15 gene-models, the values for the first cluster were different to the values of the other clusters. Based on this finding, it is possible to hypothesize that Cluster 1 has a very different profile than the other clusters, in a similar fashion as Cluster 6.

While for the Clusters 2, 3, 4 and 5 they presented the same subpopulation for the shared gene-models, which indicates a similar profile between them.

Table 15: Subpopulations of gene-models with 4 or 5 appearances on each cluster of patients.

	Cluster					
Gene-model	1	2	3	4	5	6
ABAT	3	2	NA	2	2	NA
BAI3	1	2	NA	2	2	NA
CHST1	1	2	2	NA	2	NA
DPP6	2	1	1	1	1	NA
FAM5B	1	2	NA	2	2	NA
FCGBP	1	NA	2	2	NA	2
GAS2	2	2	2	2	NA	NA
GPM6A	2	1	NA	1	1	NA
GPR56	2	1	NA	1	1	NA
KLHL9	1	1	1	NA	NA	1
LGI1	1	2	2	NA	2	NA
NOVA1	1	2	NA	2	2	NA
PGBD5	1	2	2	NA	2	NA
SCN3A	1	2	NA	2	2	NA
STMN4	NA	1	1	1	1	1

NA : Cluster of patients has no subpopulations for that gene-model.

Additionally, in the annex Section B.3 the Venn diagram for all the clusters over their gene-models is available. It can be seen that for Cluster 6 (in yellow) 20 of the total 40 gene-models make an appearance only in that cluster.

An additional approach in the study of shared gene-models, is the analysis of the ratio of the pair-wise shared gene-models. The ratio refers to the fraction of gene-models where the patients of the clusters are in the same subpopulation over the total number of shared gene-models This analysis was performed only for pairs of clusters with more than 10 gene-models shared. This is shown in Table 16.

A ratio of 1 means that for all the shared gene-models between the clusters, the patients of both clusters are in the same subpopulations, thus they have the same coefficients for their covariants. While a ratio close to 0 shows that there are none or a very low number of shared gene-models for which the clusters are in the same subpopulation.

From these results it can be seen that onle Cluster 1 has a high ratio with Cluster 6, and in a lower scale with Cluster 3. While the Clusters

Table 16: Ratio of gene-models where the clusters of patients are in the same submodel. Only pairs of clusters with more than 10 shared gene-models are shown.

Cluster A	Cluster B	Ratio
1	2	0,29
1	3	0,75
1	4	0,22
1	5	0,17
1	6	0,93
2	3	0,71
2	4	0,97
2	5	1,00
3	5	1,00
4	5	0,93

2, 3, 4, 5 form a group with high ratios between them, which is in line with the previously shown results.

Considering all the results in this section is possible to observe the formation of 2 groups. The first one formed by Clusters 1 and 6 is an independent group, where the gene-models are not necessarily shared, but the profiles of the clusters are unique. The second group is formed by Clusters 2, 3, 4 and 5, where the clusters are present in the same subpopulation for their shared models and their ratio is over 0,70 for all the pair-wise combinations available. This makes the second group closer and more significant for the profiling of the clusters.

5.3.3 *Features analysis of the gene-models for each cluster of patients*

After the inter-cluster analysis of the shared gene-models and subpopulations, the study of the selected gene-models features is presented here. For this approach, the focus is put on the particularities of the selected gene-models, their subpopulations and the covariants for each of the clusters.

To perform this, two methodologies have been defined: an intra-cluster study of the features, where the features with non-zero coefficients are aggregated and studied for each cluster, and an enrichment analysis of the features, where the features are compared to previously known gene-sets.

5.3.3.1 *Intra-cluster feature aggregation over gene-models*

The first approach for this analysis is to aggregate the different features (gene and microRNA expression, CNV and CpG sites methylation) with non-zero coefficients for all the selected gene-models in a cluster.

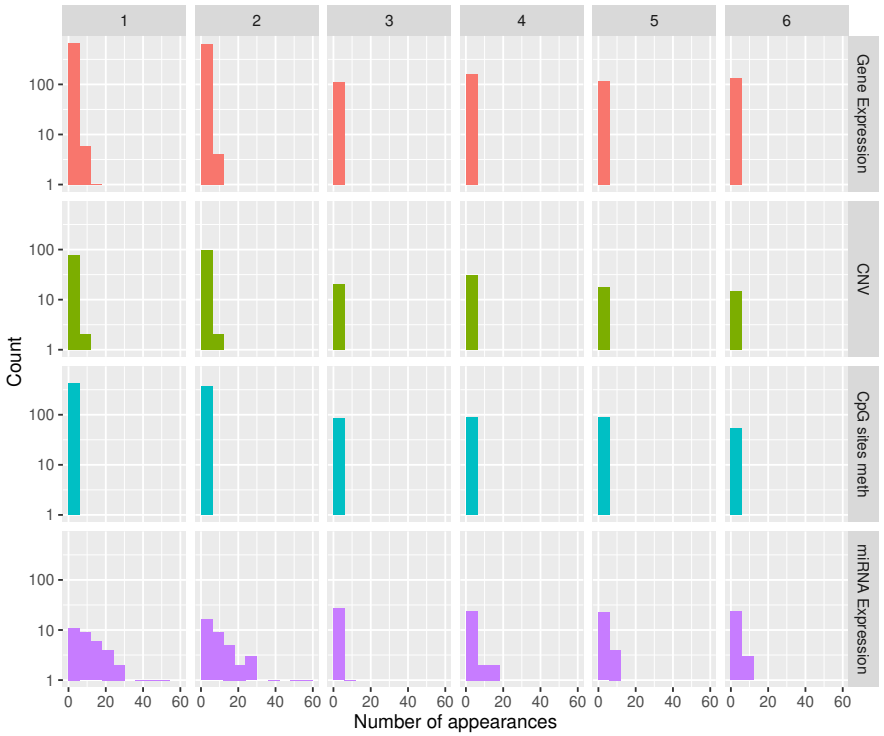


Figure 28: Distribution of the covariants with non-zero coefficients for each layer and cluster. Covariates of all the selected subpopulations were aggregated. Y-axis is in log scale.

In Figure 28, the distribution of the amount of appearances on all the selected subpopulations for the covariants with non-zero coefficients is presented for each layer and cluster.

The distribution of the CpG sites methylation, gene expression and CNV features is almost identical for all clusters. In these cases, most of the elements appears in less than 5 gene-models and in only 15 cases in more than 5 models, with a maximum value of 15 appearances.

A very different set of results is obtained with the microRNA expression layer, where the low number of elements and high connectivity makes them appear in several gene-models for each cluster. Due to this characteristic, the microRNA layer was not used in the following analysis, see Section 3.7.3.1.

Considering the aggregated values for all the covariates, but excluding the microRNA layer, a new value designated ‘appearance ratio’ was calculated, see Section 3.7.3.1.

Table 17: Contingency table for the 24 features with the highest appearances ratios over the clusters and their biological layer.

	Biological layers			
Campos clusters	CNV	Gene Expression	Methy-lation	Total
1	2	9	0	11
2	2	4	0	6
3	0	0	0	0
4	1	1	1	3
5	0	1	0	1
6	0	3	0	3
Total	5	18	1	24

A threshold was implemented to filter the features for analysis, this process was done using the appearance ratio (AR $\geqslant 2$) and the total number of appearances (App. > 2). After its application, 24 features were selected as significant.

In the Table 17, a contingency table of the number of features with significant appearances ratios are shown. The number of elements are presented by the cluster and biological layer.

As expected, based on the source of the covariates presented previously, most of the features with the highest appearances ratios belong to the gene expression layer. One interesting result is the high presence of CNV features, where they appear at a much higher rate than the CpG sites methylation layer. This can be interpreted as a high significance of a few of these features, even when most of them are not significant in the explanation of the dependent variable. This characteristic is supported by the literature, where some genomic aberrations are crucial in the profiling of GBM.

Cluster-wise only Cluster 1 presents over 10 features. For the Cluster 2, 4 and 6 are under 10 and for 2 clusters, 3 and 5, there are almost no selected features. The large number of features for Cluster 1 is surprising, considering that the large number of gene-models was taken into account by the use of the appearance ratio. This could mean a strong relationship between the selected gene-models, as shown in previous results. While for the Clusters 3 and 5, this lack of features could represent a diversity in the gene-models and makes necessary a study of all the features with an enrichment approach.

The table with the features can be found in Section A.2. In this table it is possible to observe the different features and their recurrence over the clusters.

For each cluster the following results were found:

CLUSTER 1 The genes UBC and UBD have an important presence in this cluster. These genes are related to the ubiquitination of proteins, where a ubiquitin-like protein is attached to another protein to signal, in most cases, degradation. Due to its role in regulation it is part of the epidermal growth factor receptor (EGFR) signaling pathway and the NF-κB signaling. In literature no direct role for these genes in cancer was found, only their potential as treatment targets (Vlachostergios et al., 2012; Low et al., 2012).

The gene EGFR has a high number of appearances in two layers: CNV and gene expression. This results is significant, but not unexpected. Its significance comes from the presence of a feature in more than one layer, which shows that this element has an important role in the explanation of gene-models and the cluster itself. This characteristic is not unexpected due to the abundant literature about this gene and its aberrations in the characterization of GBM, as presented in the introduction.

The CNV and gene expression of the gene NRF1 are significant features for several clusters. The role of the encoded protein is to function as a transcription factor for metabolic genes involved in the regulation of cell growth and development required for respiration[3]. The presence of this type of neural-related gene is expected in the study of brain related diseases, but its novelty could mean the discovery of a new research target not involved previously with GBM.

The expression of the APP gene appears in several clusters. This gene encodes a cell surface receptor located in the neurons with several function, such as neurite growth and formation of axons[4]. Its relationship with GBM could be similar as the one for the NRF1 gene.

Finally, several genes known for its involvement in cancer are present, such as TP53 and PIK3R1 (part of the PI3K family), which will be considered for the final profiling of the clusters.

CLUSTER 2 In this cluster the genes NRF1, UBC, EGFR and APP have a significant presence. As these genes were already discussed they will be not characterized again, but their importance will be studied when comparing the different clusters.

In addition to the previously presented gene, the protein encoded by the gene ELAVL1 has the function of binding to mRNAs to increase their stability, putting this gene in the same position as the NRF1 and APP genes.

3 http://www.uniprot.org/uniprot/Q16656
4 http://www.uniprot.org/uniprot/P05067

CLUSTER 3 No features with a significant number of appearances were found for this cluster. This issues comes from the small size of the selected gene-models for this cluster. The set of selected models is the second smallest (54 gene-models) and they present a high dispersion of the features, making it an interesting case of study in the set enrichment analysis.

CLUSTER 4 Besides the already discussed genes EGFR and UBC, the methylation site *cg21053323* appears as significant. This site is related to the gene SUMO3, which encodes the Small ubiquitin-related modifier 3 protein. This ubiquitin related protein has a different function than the UBC and UBD proteins, where it may act as antagonist of ubiquitin in the degradation process[5].

CLUSTER 5 Only the gene expression of the APP gene appears as significant in this cluster. A similar approach as the one presented before for this gene will take place in its study.

CLUSTER 6 This cluster has the smallest set of selected gene-models, but interestingly several features have a high rate of appearance. This differs from the results obtained for Clusters 3 and 5. The gene discussed firstly in Cluster 2, ELAVL1, appears here along the genes UBC and NRF1.

SHARED FEATURES Several features appeared in multiple clusters. To study this set of features an additional analysis was performed, which is the comparison of the subpopulations of the selected models where these features are present. It must be remembered that for each cluster only one of the gene-model's subpopulation is selected. Here the selected subpopulations are compared to see if the shared features belong to the same subpopulation. The covariates to study are: CNV and gene expression of EGFR and NRF1, and gene expression of UBC.

As an additional note, it was thought to perform a comparison of the coefficients (β) of the features between cluster, but because the features belong to different gene-models and, even more, different subpopulations this analysis is not significant.

In the first place, the CNV and expression of the EGFR were compared between the clusters' gene-models, 20 and 14 appearances respectively. It was found that for both layers there is only one case where the same gene-model and subpopulation are selected for more than one cluster. Interestingly, it was the same model for both features, gene-model SPRY2, where the co-appearance occurred.

5 http://www.uniprot.org/uniprot/P55854

A similar situation was discovered for the features of the gene NRF1, where for the CNV covariate (19 appearances) no shared subpopulation was found, but for the expression one (10 appearances) there was one shared subpopulation found.

In the last case, using the expression feature of gene UBC (29 appearances), it was found that for the 4 clusters where this features appeared significantly, 2 subpopulations were shared by 2 different clusters.

These results show that even when the features were significant over several clusters they appear as features of different gene-models and subpopulations and its function as unique markers can be used further.

5.3.3.2 *Gene-set enrichtment analysis*

The second method used is the gene-set enrichment approach. It was applied to features with non-zero coefficients on each cluster against the KEGG pathways and was performed using the ConsensusPathDB interaction database.

Similarly as in the previous analysis, the microRNA layer was not taken into account for this study, because of the high appearance frequency of its member. In addition, the KEGG pathways are composed by genes, due to this the CpG site methylation features were mapped to the gene layer.

Table 18: Number of features with non-zero coefficients for each cluster considering the selected gene-models. Methylation layer mapped to gene and microRNA layer not used.

Cluster	Number of features
1	1 007
2	967
3	194
4	251
5	195
6	193

The number of unique features with non-zero coefficients in the gene space varies greatly between cluster. This is due to the difference in the number of selected gene-models. In Table 18, the number of these features are shown. It is possible to observe the large difference between Cluster 1 and 2, and the rest of the clusters, which is concordant with the number of selected gene-models.

The number of obtained enriched pathway differs between clusters in a similar manner as the number of features. For Cluster 1 and 2 more than 100 pathways were found significant (adjusted p-value $< 0{,}05$), while the rest had around 70 significant pathways with the exception of Cluster 3, where only 42 pathways were obtained.

Given the large number of pathways, the first approach was to group the pathways based on their profile. This resulted into 3 different groups of pathways: disease-related, signaling pathways and diverse pathways. In the first group, gene sets related to diseases and infections such as *Pathways in cancer*, *Type I diabetes mellitus* and *Glioma* are grouped. The signaling related pathway group contains sets that have this function, such as *Rap1 signaling pathway* and *MAPK signaling pathway*. The last group is comprised of pathways that cannot be categorized in the other two, such as *Endocytosis* and *Focal adhesion*.

The second approach was to compare the pathway by their appearances. It was found that 15 pathways are present in all 6 clusters (from a maximum number of 42) and only 34 appear exclusively in only one cluster. This elevated number of shared pathways greatly hinders the characterization of the clusters when using this method. The unique pathways are comprised mostly of disease, structure and metabolic related sets. Cluster 6 has only metabolic-related pathways and is the only one with a unique composition.

Finally, the signaling pathways previously related to GBM such as p53, PI3K-Akt and ErbB signaling pathways were found in all the clusters, except Cluster 3. However, the analysis of the pathways in Cluster 3 found no particular profile for this cluster.

In summary, the use of gene set enrichment is not useful in this setting due to the large and common set of features with non-zero coefficients. The pathways found functioned more as a confirmation of the expected general profiles of the features (signaling functions, neural-related and part of diseases, in particular cancer) but did not allow for the creation of particular profiles for each cluster.

5.4 SUMMARY OF RESULTS

In this chapter, a number of results concerning the clustering of patients and their profiles have been presented. The most interesting and significant results are summarized in this section. Along with them a table, Table 19, has been added with a summary of the characteristics of each cluster. This was done in order to group together the disconnected sequential results obtained through this chapter and to facilitate their study.

In the first place, a clustering of the samples was obtained by using the co-occurrence values of the patients. Through this clustering pro-

cess, 6 clusters were obtained. In particular, Cluster 1 appeared as a well-defined cluster. This was found even when only 2 groups were considered in the clustering. This cohesion between the samples of the cluster made the definition of the other clusters difficult.

The clinical analysis of the clusters, see Table 19, showed to be an important profiling method. In particular, the Clusters 1, 2 and, in a minor manner, 3 formed a group of young samples with high survival rate and Karnofsky scores. While cluster 4, 5 and 6 were composed of older patients and low survival rate and Karnofsky scores.

The comparison to the subtypes of Verhaak et al. was feasible for only the three first clusters. For the rest, no comparison was possible to do.

Somatic mutations in the clusters were studied and the analysis showed that the genes EGFR, PTEN, TTN and TP53 were found to be mutated in over 6 samples in several clusters. These genes, with the exception of TTN, have been pointed out as significant markers of GBM and their presence was expected. From them, only the gene TP53 had a statistical significant allocation in Clusters 2 and 5.

The analysis of the mutated genes that did appeared in a few clusters did not bring any specific result. Most of the genes have no recorded link to GBM and the only ones that presented a link were IDH1, discussed later, and HIF1A, as a tentative drug target in GBM.

Gene-models were studied in order to select a set of significant gene-models for each cluster. This process delivered the called selected gene-models. The size of the sets for Clusters 1 and 2 was over 200, while for the rest of the clusters less than models 70 were selected.

The analysis of the selected gene-models showed that their intersection between clusters was small. This small set of shared gene-models was studied in deep. The significant subpopulations for the models of each cluster were compared and a subpopulation grouping was found. Clusters 1 and 6 presented an independent profile, while the rest of the clusters had the same subpopulation for the shared models.

The features of the selected gene-models were also analyzed for each cluster. This and other results are described below for each cluster.

CLUSTER 1 This group of patients showed an independent profile in most of the results, from their clustering to the selection of their gene-models and shared samples mostly with the Classical and Proneural subtypes. Mutation of several genes were found in this clusters, for most of them no link to GBM was found. A large number of selected gene-models were obtained. The features analyzed for this cluster showed an interesting presence of ubiquitin-related genes and NRF1, along to GBM related genes.

CLUSTER 2 This cluster presented some unique features, such as the mutation of the genes p53 and IDH1, which link this cluster to the Proneural subtype and to secondary GBM. Over 200 gene-models were selected for this cluster. NRF1 and EGFR have a important presence in the features of this cluster.

CLUSTER 3 A significant amount of Mesenchymal samples are allocated in this cluster. Similar to Cluster 1, several genes were found significantly mutated in this cluster, but no link to GBM was found. It presents a small amount of selected models with no high frequency features and that presented a profile similar to the whole-set profile.

CLUSTER 4 Several relevant genes were found mutated in this cluster, but not statistically significant. This is the only only group with a significant feature of the methylation layer, which is related to another ubiquitin-related modifier.

CLUSTER 5 Cluster with a hypermutated profile, in addition several genes not related to GBM were found with significant amount of mutations. This cluster appears as a unique and novel group. The distribution of the source of its covariates was unique. Only the expression of the gene APP had a significance appearance in this cluster.

CLUSTER 6 Few mutated genes, but no significant ones were discovered in this cluster. This cluster has the smallest set of selected gene-models, but several features were found with a high appearance ratio, mostly novel genes such as APP, ELAVL1, NRF1 and UBC.

In many cases, the expression and CNV of the genes EGFR and NRF1, along the expression of the UBC gene had significant appearances in many clusters. This frequent presence was studied and it was discovered that the features were located in different subpopulation for most of the gene-models and clusters, and thus the features were not actually share between clusters.

Finally, the use of an gene-set enrichment approach proved to be not useful with our settings. Due to the large number of significant enriched sets obtained no comprehensive profiling was done over the clusters using this methodology.

Table 19: Main characteristics of the discovered clusters. NS refers to no significant value found.

	Clusters					
Characteristic	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Patients' age	Young	Young	NS	Old	Old	Old
Gender	NS	NS	NS	NS	NS	NS
Non de novo presence	High	Middle	High	Low	Low	Low
Karnofsky score	High	High	NS	Low	Low	Low
Neoadjuvant treatment presence	High	Low	High	Low	Low	Low
Survival rate	High	High	Low	Low	Low	Low
Verhaak subtypes	Classical/ Proneural	Possibly Proneural	Mesenchymal	NS	NS	NS
Significant mutations	CHEK2, HIF1A, IRS4, ZNF429	IDH1, TP53	ANK2, ASPM, CHEK2, SLC25A13		C15orf2, FBLN2, LAMA1, LRP2, TP53	

Part III

DISCUSSION AND CONCLUSION

6

DISCUSSION

In this work a novel methodology for the clustering of patients based on a mixture of linear models was presented. In addition, the resulting clusters were used to define subgroups of GBM patients. Each group was analyzed based on clinical and biological data to profile them and to discover putative markers.

In order to accomplish such objectives, several complementary methods were implemented and different analysis techniques were used to interpret the results. This complex route resulted in a plethora of independent results that must be put into order to understand and discuss their significance.

DATA OBTAINMENT AND PREPROCESSING

The first stage of this work was performed in a straightforward manner for most of the datasets, where the selection and preparation of them was done to maximize the number of samples and biological layers used in the model learning. This process presented the first critical decision in this work: the exclusion of the somatic mutations as input data for the model.

Exclusion of somatic mutation data in the models

SNP data has been an hallmark of molecular disease analysis and due to this its inclusion in the study was desired. Unfortunately, this was not possible in the proposed linear model. Even when mutation data can be coded as categorical variables and integrated to our model using a dummy coding, the number of elements (possible mutated bases) to consider and the additional variables added by the coding make the models much more complex and difficult to train. This characteristic led the addition of the genetic variation data only after the model fitting.

After this decision, the preprocessing of the sets was done for the gene expression and CNV data without major issues. The microRNA dataset had to be preprocessed from the raw data, which allowed us to eliminate any doubts about the quality of the data. Finally, the CpG sites methylation data had the particularity of having two sources.

Methylation data and dual platform source

Due to the dual source of the CpG sites methylation data, the mixture model was modified adding a dummy variable for the analysis of the effect of the platforms. No significant variation was found between samples from different sources and the evidence indicates that the platforms could be used together without major issues.

With all the datasets already preprocessed its integration and use in the model was next, but it presented an obstacle: the different scales of the variables.

Covariates scaling

An initial run of the model fitting showed that variables with larger variances were over-selected by the penalized model, because of this the layers were scaled using the min-max scaling. This process worked and no layer was over-selected as shown in the analysis of the post-penalization design matrix. This procedure permits the integration of different types of biological data but the effect of this process over the fitting could not be measured. A possible negative effect of this procedure is the scaling of sets with outliers, where most of the points will be scaled down and the only significant values are those belonging to the outliers. This effect was not tested but based on the results obtained no small constant set of features with non-zero coefficients (the outliers) was found in the trained gene-models. Instead a diverse set of covariates was selected with non-zero coefficients.

TRAINING OF MIXTURE OF REGRESSION MODELS

Over 10 000 models were trained and for most of them ($> 7\,000$) only one subpopulation was obtained, which is in concordance with our expectation over the number of significant relationships for the analysis of complex diseases, see MOTIVATION. With the use of the biological prior knowledge and *lasso* penalization, the final number of covariates for all of them was less than 200. This result showed the importance of the double penalization scheme, where thanks to the biological prior knowledge it was possible to obtained models with a number of covariates small enough to be trained in a reasonable amount of time, and the *lasso* penalization allowed to reduce the number of variables even more by selecting the most significant ones.

Correlated variables and lasso

An additional characteristic of the *lasso* penalization is the elimination of highly correlated features. This penalization method will select ran-

domly one of the correlated variables and penalize the rest of them. This effect helps to deal with the correlation found in biological systems and is considered to be one of the main reasons for the heavy penalization of the methylation layer due to their bimodal profiles.

This random selection of correlated features, even though useful, is not the best approach for this issue. One proposed solution is to perform a selection prior to the fitting of the models. But this approach does not solve the issue of how to deal with covariates that are correlated but do not necessarily belong to the same system, such as CpG sites. Another approach could be the use of independent penalization terms for each layer, which would increase the fitting difficulty and would not solve the previously mentioned issue. Because of these problems our approach is considered an initial functional solution that must be studied further.

The time used for the training of this large amount of models was always a major consideration that was partially solved by the implementation of the algorithm in `C++`.

Algorithm speed and issue with number of features

The new implementation of the MFLRMP algorithm showed a gain of speed ranging from 10 times faster for smaller sets to 2 times faster for more complex data. The complexity of the data is related to the number of subpopulation trained, the number of samples and covariates. The latter was found to be the most significant parameter in the speed gain and for large models (over 500 covariates) the gain would be insignificant. This effect shows once again the major role that the penalization plays in the fitting of the models, where it has, a direct relationship with the cost of running the algorithm.

It has been considered that the loss in speed of the new implementation could be due to the movement of data between R and `C++` caused by the use of a wrapping R function over the `C++` code. This and other considerations are the object of future work of this project.

After the fitting of the gene-models the hyperparameters (λ and $\mathbf{K}$) selected for them were analyzed. It was found that a main relationship between the number of subpopulations and the strength of the penalization term existed. For models with a larger number of subpopulations, a stronger penalization was applied. This correlation was understood as a selective definition of the subpopulations: for models with one or few subpopulations, a strong penalization did not have a large effect (hence is not chosen), while for models with multiple subpopulations, the penalization helps to define each subpopulation. Additionally, an interesting issue with the analysis of the hyperparameters was found.

Difficulties in the analysis and comparison of the hyperparameters

The distribution of the hyperparameters found for each gene-model was compared and studied. As discussed in the result section, the focus in some models with a particular large number of covariates made the analysis of the distribution of the hyperparameters erratic. These models, although significant, represented a small portion of the total models and do not represent the general distribution of them. Making them a focus of the analysis without considering the vast majority of models was the source of the erratic analysis.

This effect must be taken into consideration for future analyses, in particular, when a small set of models or variables are focused on and the vast majority of the elements are not taken into account.

CLUSTERING OF PATIENTS AND PROFILING

After the fitting of the gene-models, the co-occurrence probabilities were defined for each pair of samples in order to study if they belong to the same subpopulation for most of the models. This scheme was used to group the patients into 6 clusters. A problem presented by this approach in gene-models with more than 2 subpopulations is discussed below.

Co-occurrence in multiple models

The definition of the co-occurrence probability is given by a quadratic expression as shown in Section 3.5.3. Given this profile the values are heavily skewed towards 0,5. This effect is stronger for models with more than 2 subpopulations, where the values are closer to 0,3. Even with its difficulties this value is considered the right approach from a technical point of view to calculate the co-occurrence of samples in the gene-models, because it uses the probabilistic nature of the resulting gene-models. Possible modifications to solve this issue are a major objective in the future work of this project, but due to the small number of gene-models with large amount of subpopulations this phenomena did not have a widespread effect in the analysis.

During the clustering process, a particular group of samples appear well-defined for the different setting of the clustering algorithm. This group hindered the obtainment of the other clusters and was denominated Cluster 1. Due to this effect, the criteria for the number of clusters selected was based on the visual analysis of the dendrogram and not in the silhouette values.

Multiple linear model evidence

The main hypothesis of this work is that the different groups of patients with GBM present particular inter-molecular relationships. Which means that if a linear model is built with a gene expression as output variable then the covariates and coefficients of that model can differ between patients.

This main hypothesis was probed by the obtainment of several clusters, especially one well-defined compact cluster (Cluster 1) and a group of closely related clusters. The clinical and model analysis of these clusters showed a small intersection between them and allowed us to characterized them.

Clinical analyses of the cluster

The characterization of the clusters of patients was done using the clinical data from the TCGA repository. One problem faced was the incapacity to access all the clinical data due to restriction of the clinical information of the patients. This can be seen in the difficulty to annotate the samples as secondary GBM, which was not entirely possible. The clinical analyses showed that the clusters could be grouped into two groups. The first one comprised by Clusters 1, 2 and 3 had younger patients with better prognosis, while the other clusters were comprised by older patients with more negative prognosis. This gave us a first approach the general organization and profiling of the clusters.

A second profiling approach was the comparison of the clusters defined in this work with those defined in Verhaak et al.

Cluster and subtypes comparison

This comparison was not completely successful due to two issues. The first one was the small number of samples present in both studies. Most of the samples used in Verhaak et al. could not be used in this work because they lack information in some layers. Considering this problem the use of samples with full information (paired data) is crucial for the training of the models and cannot be changed. The second issue was the allocation of the shared samples. In this case, no subtype was completely allocated in a cluster and, in fact, most of the subtypes were assigned to multiple clusters. Because of these characteristic, a partial comparison was done, where clusters 1 and 3 were identify as possible analogs to the Classical and Mesenchymal subtypes respectively and Cluster 2 to the Proneural subtype in a less stringent manner. The rest of the clusters presented almost no shared samples and are considered novel groups.

The last profiling performed to the samples was the somatic mutations analysis, where the SNP data not used previously as input data for the models was utilized to study the mutation profiles of the clusters.

Somatic mutations profiles

This analysis was the first proof of concept of the importance of certain genes (EGFR, PTEN and TP53) in the profiling of GBM. Mutations for these genes appeared in several clusters, but only the gene TP53 had a statistical significant presence in some clusters (2 and 5). In a similar fashion, gene IDH1, which is key in the study of secondary GBM, appeared only in the Cluster 2 in a significant amount, reinforcing its profile similar to the Proneural subtype.

Several genes showed a significant mutation profile in a particular cluster, but for the most of them no connection to GBM was possible based on the literature. Besides announcing their potential role as biomarkers for the different clusters, the implication of the presence of these genes is something that could not be completely studied in this work. Further analysis of their status in additional samples is needed.

GENE-MODELS AND FEATURES ANALYSIS FOR EACH CLUSTER

With the profiling of the patients for each clusters done, the analysis was put on the gene-models that define each cluster and their features. The first approach for this was the selection of the significant gene-models for each cluster and their comparison.

Gene-models selection and analysis

The selection of gene-models for each cluster gave rise to a set of unique models with very small intersection between clusters. Clusters 1 and 2 obtained a large number of selected gene-models, over 200, while for the rest of the clusters this amount was under 100. This result came as a surprise, considering that Cluster 1 is the smallest cluster, but the one with the highest co-occurrence between samples. The difference in the selected models size can be explained due to the cohesion of the members of the cluster. Clusters, such as the first one, appear with high co-occurrence in several models, while a group of samples with low cohesion would appear in few gene-models with high co-occurrence.

In addition, the analysis of the selected gene-models was done between clusters, which resulted in a very small number of models shared by 4 or more clusters. The analysis of the subpopulation of these shared models showed that the samples from the Clusters 2, 3, 4 and 5 were allocated in the same subpopulations for these models,

while Clusters 1 and 6 presented unique profiles. This was the first result were Clusters 1 and 2 were allocated in different groups and Cluster 6 presented an unique profile.

A short gene set enrichment analysis of the gene-models selected for each cluster was performed. This study gave no interesting results and for the clusters with the smaller sets of selected models almost no significant pathway (adjust p-value $\leqslant 0{,}05$) were retrieved.

Features aggregation for each cluster

Having performed the analysis of the gene-models the attention was then given to the features of these models, where the aggregation of them for each cluster let us discover relevant covariates for several models in a cluster. During the analysis, the number of appearances of each feature was normalized based on the amount of gene sets selected for that cluster. The normalization was crucial to discriminate covariates in clusters with a large number of selected models.

Most of the features retrieved with the highest appearance between clusters were directly related to GBM, such as EGFR, PIK3R1, TP53. This not only confirmed their importance in these tumors but also proved that their signatures can define essential gene-models. While CNV was not relevant in most results presented previously, features belonging to this cluster appeared highly significant in this analysis.

Besides the genes and proteins families involved in GBM, a set of covariates that appeared repeatedly are NRF1 and the ubiquitin-related genes.

The gene NRF1 appears as a significant feature in several clusters, its gene expression and CNV signatures were crucial in the definition of the significant gene-models. No linked has been made between this gene and GBM before. The appearances of ubiquitin-related genes in several clusters and from 3 different genes show that this family of genes have a crucial role in the definition of the gene-models and clusters. No relationship between ubiquitin functions and GBM was found in literature, which opens a door to a new focus in the treatment of GBM.

The last analysis applied was the gene set enrichment analysis of the features, which presented important issues.

Gene set enrichment application and issues

Gene sets enrichment analysis over different sets of results was used several times in this work. A major complication was found in the application of it to the models' features due to the large number of genes to study (from 200 to over 1 000). This large amount of genes led to a large number of significant enriched pathways. Making any distinc-

tion between the pathways was not possible and the study in general delivered no important results.

The main issue in the gene set enrichment analysis was the incapability of selecting or weighting the features. In other studies, a statistic or p-value is used along a set of genes to filter or enrich them. This problem is discussed below in the relation of how to analyze the features and their coefficients.

Coefficients significance

One main issue was the difficulty to study the coefficients of the covariates. This issues was partially solved by analyzing the appearances of covariates over the selected gene-models of a cluster, but there was no direct way to discriminate between the different features with non-zero coefficients. One possible, but erroneous approach, would be to discriminate them based on their absolute value, but this method would not take into consideration the statistical significance of the coefficient.

A better approach would be to calculate a statistical for each coefficient. In general, is possible to calculate a t-statistic for regression models where the null hypothesis of the coefficient being 0, is tested. This approach is not yet possible in our case, just a few years ago a method was proposed to perform such analysis in simple linear models with *lasso* penalization (Lockhart et al., 2014). This methodology contemplates an iteration of the penalization over individual features, calculating on each step the importance of that feature. A method such as this would not only be very costly in our framework but also it would have to consider the use of the EM-algorithm in our method. This problem is significant, but its resolution escapes the scope of this work.

microRNA layer presence issue

For the previous analyses, the microRNA layer was not used. This decision was based on the difficulties to study the covariates belonging to this layer due to their high appearances in most of the gene-models. This issues arises from the small amount of elements belonging to this layer and the large number of interactions between them and the gene layer. Due to these characteristics, the small set of microRNAs were present as covariates for most of the gene-models, making them indistinguishable based on their appearances.

To solve this issue, additional experimentally validated interactions information are necessarily. With these additional datasets it would be possible to have a larger number of microRNAs included in the design matrices and only true positive interactions.

7

CONCLUSION

This work focused in the non-sequential integrative analysis of GBM samples and the discovery of novel subgroups of patients with unique profiles and mechanics. This project not only showed the potential of the mixture of linear models approach for the integration of heterogeneous data for the study of complex diseases, but was also able to discover 6 different subgroups, half of them resembling previously defined subtypes and the other half as novel clusters.

Key genomic aberrations and genes differential expression in the study of GBM were found as critical markers in the clusters profiling. In addition, ubiquitin-related genes appeared along these previously known markers and have opened a door to a new set of research targets.

RESEARCH OBJECTIVES AND METHODOLOGY

The main objective for this project was the definition and implementation of an integrative framework of heterogeneous data from GBM patients and the detection of subgroups of patients in it. This objective was accompanied with the use of the subgroups to analyze the molecular and clinical signatures of the patients to discover possible novel research targets.

The main objective presented itself as a complex problem which had to be solved by a pipeline of different methodologies. These methods range from the obtainment and preprocessing of the datasets to be integrated (gene and microRNA expression, CNV and CpG sites methylation), to the definition, performance optimization and execution of the mixture of regression model with heterogeneous data, and finally the clustering of samples based on the resulting mixture models.

The second objective, the analysis of the resulting clusters, was performed to discover clinical and molecular enrichments in the subgroups. Then, the most significant gene-models for each cluster were selected and their structure studied and compare between subgroups.

The enrichment analysis of the clusters allowed us to discover particular profiles and the putative research target for each cluster.

MAIN RESULTS AND IMPLICATIONS

Through the integration of the multi-'omic' datasets, over 10 000 gene-models were trained and thanks to the double penalization scheme,

all these models had less than 200 covariates. From them, over 2 000 presented subpopulations and were used in the clustering of the GBM samples.

This first set of results was studied in order to observe the behavior of the MFLRMP algorithm and whether the resulting gene-models followed the expected configuration. In line with our hypothesis only over a fifth of the gene-models presented subpopulations, and for models with more subpopulations, the penalization strength was stronger and thus smaller and better defined subpopulations were obtained.

The clustering of the samples using the gene-models allowed us to discover 6 distinct clusters. Using this grouping of the samples, the characterization of the groups was done using the clinical information and the somatic mutations profiles of the samples. This profiling of the groups allowed us to give an initial characterization to them, finding in addition supergroups of clusters that share significant characteristics. The comparison to previously studies was difficult due to the small number of samples shared between them.

By the additional analysis of the significant models for each cluster, it was possible to study the gene-models and the features that define these groups. A significant role of GBM related aberrations, such as CNV and differential expression of EGFR and TP53, was discovered. In addition, it was found a strong involvement of elements not previously related to GBMs such as NRF1 and ubiquitin-related genes.

FUTURE WORK

Some of the several methodologies comprising the pipeline developed in this work are the direct focus of future development in order to achieve an analysis of higher quality. These methodologies are the re-definition of the co-occurrence values for a better clustering of the samples and the inclusion of a significance test for the gene-models' coefficients. These improvements are not easily achieved and can be considered projects on their own, in particular the latter one.

Additionally, several smaller elements can be optimized for an improvement in the execution of this pipeline, such as the revision of the C++ code and the inclusion of the microRNA layer in the feature analysis, which should be possible with the appearance of larger validated interaction databases.

Finally, it would be in our interest to apply this framework in other complex diseases for their analysis. Thanks to the TCGA repository is possible to obtain the datasets necessarily for other types of cancer, making them the most obvious targets of new studies.

This work has developed a new analytic framework to study complex diseases. This novel framework is capable of solving three main issues: integration of heterogeneous data, patients' subgroups discovery and translation of these results into research target and patients profiling.

The used methodology allows to integrate the different layers of information from the beginning. This is a crucial difference with the most common sequential methodologies to analyze heterogeneous data. Furthermore, our framework allows us to group the patients using these multi-'omic' gene-models and not to compare or integrate the clustering results of each individual layer of information. And finally, the gene-models that define the clusters are comprised not only by expected genetic aberrations, but new elements also arise as defining elements in the gene-models and proposed research targets.

Part IV

SUMMARY

SUMMARY

Glioblastoma multiforme (GMB) is an extremely aggressive and invasive brain cancer with a median survival of less than one year. In addition, due to its anaplastic nature the histological classification of this cancer is not simple. These characteristics make this disease an interesting and important target for new methodologies of analysis and classification. In recent years, molecular information has been used to segregate and analyze GBM patients, but generally this methodology utilizes single-'omic' data to perform the classification or multi-'omic' data in a sequential manner. In this project, a novel approach for the classification and analysis of patients with GBM is presented. The main objective of this work is to find clusters of patients with distinctive profiles using multi-'omic' data with a real integrative methodology.

During the last years, the TCGA consortium has made publicly available thousands of multi-'omic' samples for multiple cancer types. Thanks to this, it was possible to obtain numerous GBM samples (> 300) with data for gene and microRNA expression, CpG sites methylation and copy-number variation (CNV). To achieve our objective, a mixture of linear models were built for each gene using its expression as output and a mixture of multi-'omic' data as covariates. Each model was coupled with a lasso penalization scheme, and thanks to the mixture nature of the model, it was possible to fit multiple submodels to discover different linear relationships in the same model. This complex but interpretable method was used to train over 10 000 models. For ~2 400 cases, two or more submodels were obtained.

Using the models and their submodels, 6 different clusters of patients were discovered. The clusters were profiled based on clinical information and gene mutations. Through this analysis, a clear separation between the younger patients and with higher survival rate (Clusters 1, 2 and 3) and those from older patients and lower survival rate (Clusters 4, 5 and 6) was found. Mutations in the gene IDH1 were found almost exclusively in Cluster 2, additionally, Cluster 5 presented a hypermutated profile. Finally, several genes not previously related to GBM showed a significant presence in the clusters, such as C15orf2 and CHEK2.

The most significant models for each clusters were studied, with a special focus on their covariants. It was discovered that the number of shared significant models were very small and that the well known GBM related genes appeared as significant covariates for plenty of models, such as EGFR1 and TP53. Along with them, ubiquitin-related

genes (UBC and UBD) and NRF1, which have not been linked to GBM previously, had a very significant role.

This work showed the potential of using a mixture of linear models to integrate multi-'omic' data and to group patients in order to profile them and find novel markers. The resulting clusters showed unique profiles and their significant models and covariates were comprised by well known GBM related genes and novel markers, which present the possibility for new approaches to study and attack this disease. The next step of the project is to improve several elements of the methodology to achieve a more detail analysis of the models and covariates, in particular taking into account the regression coefficients of the submodels.

ZUSAMMENFASSUNG

Glioblastoma multiforme (GMB) ist eine extrem aggressive und invasive Form von Hirntumor mit einer medianen Überlebenszeit von unter einem Jahr. Des weiteren ist, aufgrund seiner anaplastischen Natur, eine histologische Klassifikation nicht einfach. Aufgrund dieser Charakteristika ist GMB ein interessantes und wichtiges Ziel für neue Methoden der Analyse und Klassifizierung. In jüngster Zeit wurden molekulare Informationen zur Segregation und Analyse von GMB Patienten verwendet, allerdings verwendet diese Methode meist nur einen Datentyp zur Klasifizierung, oder multiple-“omics“ Daten in einer sequenziellen Weise. In dieser Arbeit wird ein neuer Ansatz zur Klassifizierung und Analyse von Patienten mit GMB vorgestellt. Hauptziel ist die Identifikation von Patienten-Clustern mit charakteristischen Profilen, unter Verwendung multipler-“omic“ Daten mit einer echten integrativen Methode.

In den letzten Jahren wurden vom TCGA Konsortium tausende multi-“omic“ Proben verschiedener Krebstypen öffentlich zur Verfügung gestellt. Dank diesem war es möglich zahlreiche (>300) GMB Proben mit Daten für: Gen- und microRNA Expression, CpG-Dinukleotid Methylierung, und copy-number variation (CNV). Um unsere Zielsetzung zu erreichen wurde eine Mischung von linearen Modellen für jedes Gen erzeugt, die Genexpression als Ausgabe und eine Mischung von multi-“omics“ Daten als Kovariaten verwendend. Jedes Modell wurde mit einem lasso penalization Schema gekoppelt, und Dank der gemischten Natur des Modells war es möglich multiple Submodelle zu fitten, um verschiedene lineare Beziehungen im selben Modell zu entdecken. Diese komplexe aber interpretierbare Methode wurde verwendet um über 10 000 Modele zu trainieren. Wobei für ~2 400 Fälle zwei oder mehr Submodelle erhalten wurden.

Die Modelle und ihre Submodelle verwendend, wurden 6 verschiedene Patienten-Cluster entdeckt. Diese Cluster wurden anhand klinischer Informationen und Genmutationen profiliert. Dabei zeigte sich eine klare Trennung zwischen jüngeren Patienten mit höherer Überlebensrate (Cluster 1, 2 und 3) und älteren Patienten mit niedrigerer Überlebensrate (Cluster 4, 5 und 6). Mutationen im IDH1 Gen wurden fast ausschließlich in Cluster 2 gefunden und Cluster 5 präsentierte ein hypermutiertes Profil. Zusätzlich zeigte sich eine signifikante Präsenz von bisher nicht mit GMB in Verbindung gebrachten Genen (wie C15orf2 und CHEK2) in den Clustern.

Das signifikantesten Modelle jedes Clusters wurde studiert, wobei ein spezieller Fokus auf ihre Kovarianten gelegt wurde. Es wurde ent-

deckt, dass die Anzahl geteilter signifikanter Modelle sehr klein war und das die bekannten mit GMB zusammenhängenden Gene (wie EGFR1 und TP53) in vielen Modellen als Covariablen auftauchen. Zusammen mit diesen spielten Ubiquitin-verwandte Gene (UBC und UBD) sowie NERF1, welche bisher nicht mit GMB in Zusammenhang gebracht wurden, eine sehr signifikante Rolle.

Diese Arbeit zeigt das Potential einer Mischung linearer Modelle um multi-“omics“ Daten zu integrieren sowie Patienten zu gruppieren um sie zu profilieren und neue Marker zu finden. Die resultierenden Cluster zeigten einzigartige Profile und ihre signifikanten Modelle bestanden aus bekannten mit GMB zusammenhängenden (assoziierten) Genen und neuen Markern, welche die Möglichkeit für neue Ansätze zum Studium und Bekämpfung dieser Krankheit eröffnen. Der nächste Schritt dieses Projektes ist es mehrere Elemente der Methoden zu verbessern um eine detailliertere Analyse der Modelle zu ermöglichen, im Speziellen unter Berücksichtigung der Regressionskoeffizienten der Kovariaten.

Part V

APPENDIX

A

DATA

A.1 SHARED PATIENTS SAMPLES BETWEEN DATASET

Patients' id	Patients' id	Patients' id
TCGA-02-0001-01	TCGA-12-0826-01	TCGA-19-2631-01
TCGA-02-0003-01	TCGA-12-0827-01	TCGA-19-5947-01
TCGA-02-0007-01	TCGA-12-0828-01	TCGA-19-5950-01
TCGA-02-0009-01	TCGA-12-0829-01	TCGA-19-5952-01
TCGA-02-0010-01	TCGA-12-1088-01	TCGA-19-5954-01
TCGA-02-0011-01	TCGA-12-1089-01	TCGA-19-5955-01
TCGA-02-0014-01	TCGA-12-1090-01	TCGA-19-5956-01
TCGA-02-0021-01	TCGA-12-1091-01	TCGA-19-5958-01
TCGA-02-0024-01	TCGA-12-1092-01	TCGA-19-5959-01
TCGA-02-0027-01	TCGA-12-1093-01	TCGA-19-5960-01
TCGA-02-0028-01	TCGA-12-1094-01	TCGA-26-1438-01
TCGA-02-0033-01	TCGA-12-1095-01	TCGA-26-1440-01
TCGA-02-0034-01	TCGA-12-1096-01	TCGA-26-1442-01
TCGA-02-0038-01	TCGA-12-1097-01	TCGA-26-1443-01
TCGA-02-0043-01	TCGA-12-1098-01	TCGA-26-1799-01
TCGA-02-0047-01	TCGA-12-1099-01	TCGA-26-5133-01
TCGA-02-0052-01	TCGA-12-1598-01	TCGA-26-5134-01
TCGA-02-0054-01	TCGA-12-1599-01	TCGA-26-5135-01
TCGA-02-0057-01	TCGA-12-1600-01	TCGA-26-5136-01
TCGA-02-0058-01	TCGA-12-1602-01	TCGA-26-5139-01
TCGA-02-0060-01	TCGA-12-3644-01	TCGA-27-1830-01
TCGA-02-0064-01	TCGA-12-3646-01	TCGA-27-1831-01
TCGA-02-0069-01	TCGA-12-3648-01	TCGA-27-1832-01
TCGA-02-0071-01	TCGA-12-3649-01	TCGA-27-1833-01
TCGA-02-0074-01	TCGA-12-3650-01	TCGA-27-1834-01
TCGA-02-0075-01	TCGA-12-3651-01	TCGA-27-1835-01
TCGA-02-0080-01	TCGA-12-3652-01	TCGA-27-1836-01
TCGA-02-0083-01	TCGA-12-3653-01	TCGA-27-1837-01

Continued on next page

Patients' id	Patients' id	Patients' id
TCGA-02-0085-01	TCGA-12-5295-01	TCGA-27-1838-01
TCGA-02-0086-01	TCGA-12-5301-01	TCGA-27-2518-01
TCGA-02-0089-01	TCGA-14-0736-01	TCGA-27-2521-01
TCGA-02-0099-01	TCGA-14-0781-01	TCGA-27-2523-01
TCGA-02-0102-01	TCGA-14-0783-01	TCGA-27-2524-01
TCGA-02-0107-01	TCGA-14-0786-01	TCGA-27-2526-01
TCGA-02-0113-01	TCGA-14-0787-01	TCGA-27-2527-01
TCGA-02-0114-01	TCGA-14-0789-01	TCGA-27-2528-01
TCGA-02-0115-01	TCGA-14-0790-01	TCGA-28-1746-01
TCGA-02-0116-01	TCGA-14-0812-01	TCGA-28-1747-01
TCGA-02-2466-01	TCGA-14-0813-01	TCGA-28-1749-01
TCGA-02-2470-01	TCGA-14-0817-01	TCGA-28-1750-01
TCGA-02-2483-01	TCGA-14-0865-01	TCGA-28-1751-01
TCGA-02-2485-01	TCGA-14-0866-01	TCGA-28-1752-01
TCGA-02-2486-01	TCGA-14-0867-01	TCGA-28-1753-01
TCGA-06-0122-01	TCGA-14-0871-01	TCGA-28-1755-01
TCGA-06-0124-01	TCGA-14-1034-01	TCGA-28-1756-01
TCGA-06-0125-01	TCGA-14-1037-01	TCGA-28-1757-01
TCGA-06-0126-01	TCGA-14-1396-01	TCGA-28-2502-01
TCGA-06-0128-01	TCGA-14-1401-01	TCGA-28-2506-01
TCGA-06-0129-01	TCGA-14-1402-01	TCGA-28-2509-01
TCGA-06-0130-01	TCGA-14-1451-01	TCGA-28-2510-01
TCGA-06-0133-01	TCGA-14-1452-01	TCGA-28-2513-01
TCGA-06-0137-01	TCGA-14-1453-01	TCGA-28-2514-01
TCGA-06-0139-01	TCGA-14-1454-01	TCGA-28-5204-01
TCGA-06-0140-01	TCGA-14-1455-01	TCGA-28-5207-01
TCGA-06-0141-01	TCGA-14-1456-01	TCGA-28-5208-01
TCGA-06-0142-01	TCGA-14-1458-01	TCGA-28-5209-01
TCGA-06-0143-01	TCGA-14-1459-01	TCGA-28-5213-01
TCGA-06-0145-01	TCGA-14-1794-01	TCGA-28-5214-01
TCGA-06-0147-01	TCGA-14-1795-01	TCGA-28-5215-01
TCGA-06-0155-01	TCGA-14-1821-01	TCGA-28-5216-01
TCGA-06-0169-01	TCGA-14-1823-01	TCGA-28-5218-01
TCGA-06-0650-01	TCGA-14-1825-01	TCGA-28-5219-01
TCGA-06-0875-01	TCGA-14-1827-01	TCGA-28-5220-01

Continued on next page

Patients' id	Patients' id	Patients' id
TCGA-06-0876-01	TCGA-14-1829-01	TCGA-28-6450-01
TCGA-06-0877-01	TCGA-14-2554-01	TCGA-32-1970-01
TCGA-06-0878-01	TCGA-14-2555-01	TCGA-32-1973-01
TCGA-06-0879-01	TCGA-14-3477-01	TCGA-32-1976-01
TCGA-06-0881-01	TCGA-14-4157-01	TCGA-32-1977-01
TCGA-06-0882-01	TCGA-15-1444-01	TCGA-32-1978-01
TCGA-06-0939-01	TCGA-15-1446-01	TCGA-32-1979-01
TCGA-06-1084-01	TCGA-15-1447-01	TCGA-32-1980-01
TCGA-06-1086-01	TCGA-15-1449-01	TCGA-32-1982-01
TCGA-06-1087-01	TCGA-16-0846-01	TCGA-32-1986-01
TCGA-06-1800-01	TCGA-16-0848-01	TCGA-32-1987-01
TCGA-06-1801-01	TCGA-16-0849-01	TCGA-32-1991-01
TCGA-06-1802-01	TCGA-16-0850-01	TCGA-32-2491-01
TCGA-06-1804-01	TCGA-16-0861-01	TCGA-32-2494-01
TCGA-06-1805-01	TCGA-16-1045-01	TCGA-32-2495-01
TCGA-06-2557-01	TCGA-16-1047-01	TCGA-32-2498-01
TCGA-06-2558-01	TCGA-16-1048-01	TCGA-32-2615-01
TCGA-06-2559-01	TCGA-16-1055-01	TCGA-32-2616-01
TCGA-06-2561-01	TCGA-16-1056-01	TCGA-32-2632-01
TCGA-06-2562-01	TCGA-16-1060-01	TCGA-32-2634-01
TCGA-06-2563-01	TCGA-16-1062-01	TCGA-32-2638-01
TCGA-06-2564-01	TCGA-16-1063-01	TCGA-32-4208-01
TCGA-06-2565-01	TCGA-16-1460-01	TCGA-32-4211-01
TCGA-06-2566-01	TCGA-19-0955-01	TCGA-32-4213-01
TCGA-06-2567-01	TCGA-19-0957-01	TCGA-32-4719-01
TCGA-06-2569-01	TCGA-19-0960-01	TCGA-32-5222-01
TCGA-06-2570-01	TCGA-19-0962-01	TCGA-41-2571-01
TCGA-06-5408-01	TCGA-19-0963-01	TCGA-41-2572-01
TCGA-06-5410-01	TCGA-19-0964-01	TCGA-41-2573-01
TCGA-06-5411-01	TCGA-19-1385-01	TCGA-41-3392-01
TCGA-06-5412-01	TCGA-19-1386-01	TCGA-41-3393-01
TCGA-06-5413-01	TCGA-19-1387-01	TCGA-41-3915-01
TCGA-06-5414-01	TCGA-19-1388-01	TCGA-41-5651-01
TCGA-06-5415-01	TCGA-19-1389-01	TCGA-76-4925-01
TCGA-06-5416-01	TCGA-19-1390-01	TCGA-76-4926-01

Continued on next page

Patients' id	Patients' id	Patients' id
TCGA-06-5418-01	TCGA-19-1392-01	TCGA-76-4931-01
TCGA-06-5856-01	TCGA-19-1786-01	TCGA-76-4934-01
TCGA-06-5858-01	TCGA-19-1787-01	TCGA-76-4935-01
TCGA-06-5859-01	TCGA-19-1789-01	TCGA-76-6191-01
TCGA-06-6389-01	TCGA-19-1791-01	TCGA-76-6192-01
TCGA-06-6391-01	TCGA-19-2620-01	TCGA-76-6193-01
TCGA-12-0670-01	TCGA-19-2623-01	TCGA-76-6282-01
TCGA-12-0820-01	TCGA-19-2624-01	TCGA-76-6285-01
TCGA-12-0821-01	TCGA-19-2625-01	TCGA-81-5910-01
TCGA-12-0822-01	TCGA-19-2629-01	TCGA-87-5896-01

Table 20: Set of shared patients between datasets.

A.2 FEATURES WITH THE HIGHEST APPEARANCES RATIOS ON EACH CLUSTER

Cluster	Layer	Name	Appearances	Appearances Ratio
1	Gene Expr.	UBC	15	6.20
1	CNV	EGFR	11	4.55
1	CNV	NRF1	10	4.13
1	Gene Expr.	APP	9	3.72
1	Gene Expr.	EGFR	8	3.31
1	Gene Expr.	NRF1	7	2.89
1	Gene Expr.	PIK3R1	7	2.89
1	Gene Expr.	SFN	6	2.48
1	Gene Expr.	UBD	6	2.48
1	Gene Expr.	FYN	5	2.07
1	Gene Expr.	TP53	5	2.07
2	CNV	NRF1	9	3.37
2	Gene Expr.	ELAVL1	9	3.37
2	Gene Expr.	UBC	7	2.62
2	CNV	EGFR	6	2.25
2	Gene Expr.	APP	6	2.25
2	Gene Expr.	EGFR	6	2.25
4	CNV	EGFR	3	4.41
4	Gene Expr.	UBC	3	4.41
4	Methyation	cg21053323	3	4.41
5	Gene Expr.	APP	3	4.29
6	Gene Expr.	ELAVL1	5	12.20
6	Gene Expr.	UBC	4	9.76
6	Gene Expr.	NRF1	3	7.32

Table 21: Features with a significant number of appearances for each cluster considering the selected gene-models.

B

IMAGES

B.1 ANALYSIS OF THE SILHOUETTE VALUES OF THE CO-OCCURRENCE MATRIX

Histograms of the silhouette values for the clustering of the co-occurrence matrix for different number of clusters.

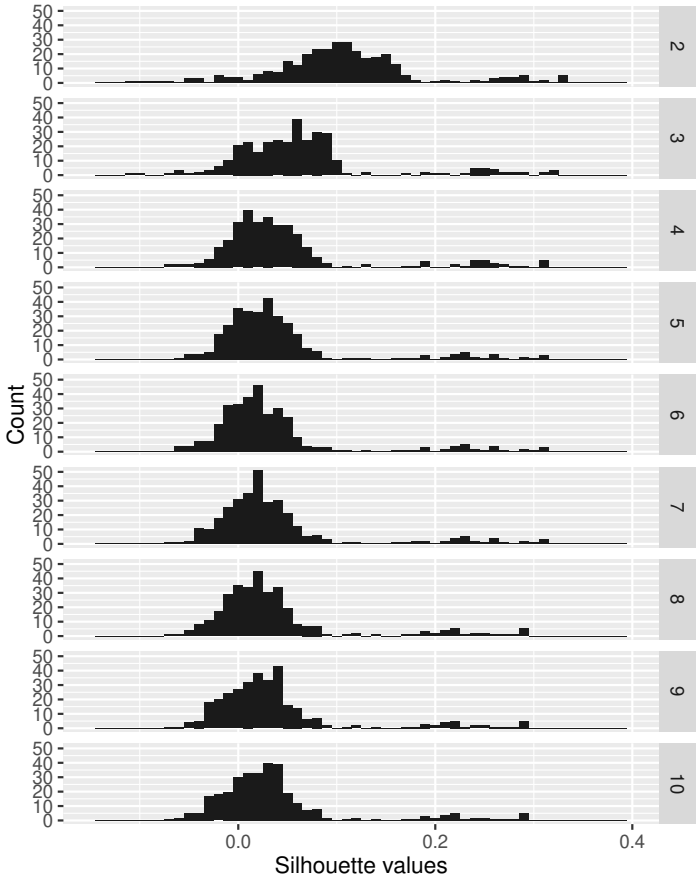
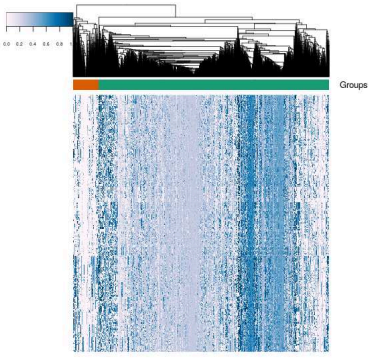
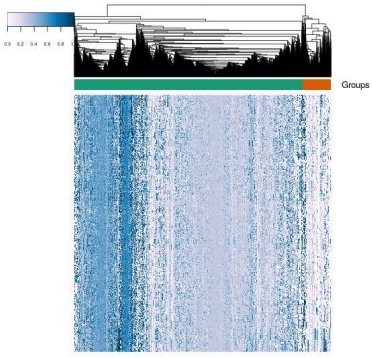


Figure 29: Histogram of the silhouette values of the co-occurrence matrix for different number of clusters.

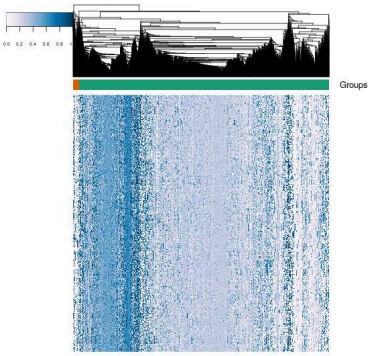
B.2 HEATMAPS OF CO-OCCURRENCE DISTANCES FOR THE GENE-MODELS GROUPING FOR THE PATIENTS' CLUSTERS



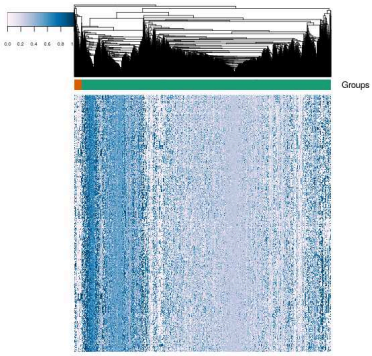
(a) Heatmap cluster 1



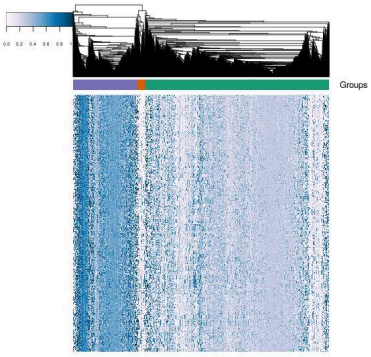
(b) Heatmap cluster 2



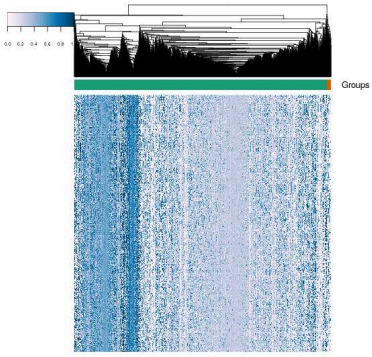
(c) Heatmap cluster 3



(d) Heatmap cluster 4



(e) Heatmap cluster 5



(f) Heatmap cluster 6

Figure 30: Heatmaps and dendrograms of the co-occurrence values to group gene-models for each cluster, the clustering of gene-models was performed with $K = 2$ for the clusters $1, 2, 4$ and 5 and $K = 3$ for the rest.

B.3 VENN DIAGRAM OF THE SELECTED GENE-MODELS FOR THE PATIENTS CLUSTERS

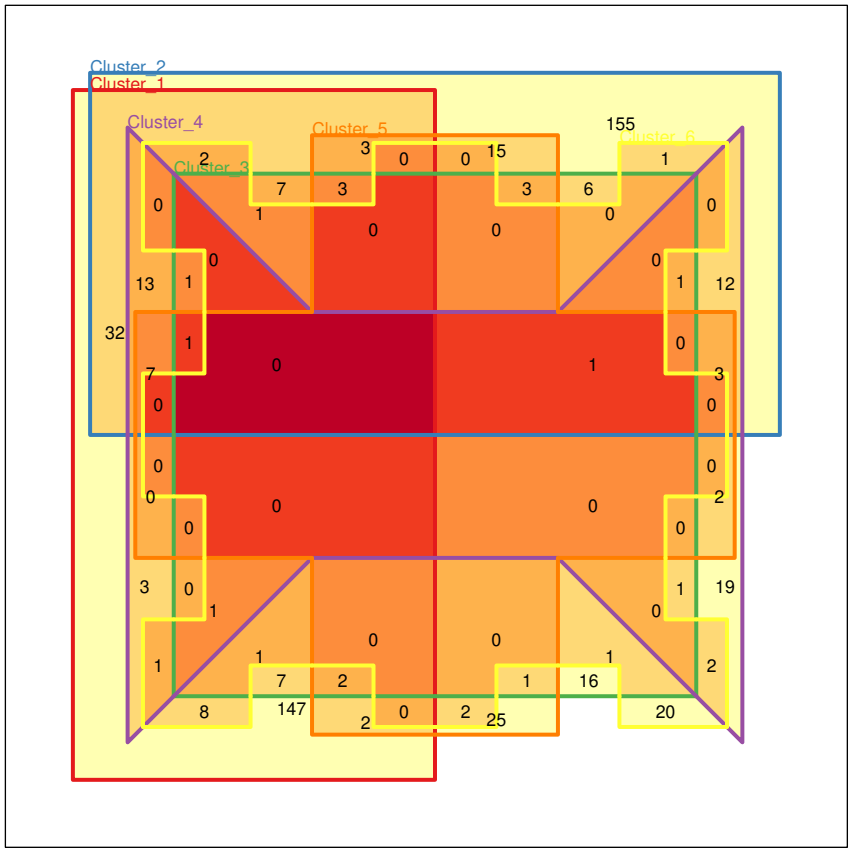


Figure 31: Venn diagram of the 6 clusters over their selected gene-models.

BIBLIOGRAPHY

Abatangelo L, Maglietta R, Distaso A, D'Addabbo A, Creanza T, Mukherjee S, and Ancona N. 2009. Comparative study of gene set enrichment methods. *BMC Bioinformatics* 10.1:275. DOI: 10.1186/1471-2105-10-275.

Agnihotri S, Burrell KE, Wolf A, Jalali S, Hawkins C, Rutka JT, and Zadeh G. 2013. Glioblastoma, a Brief Review of History, Molecular Genetics, Animal Models and Novel Therapeutic Strategies. *Arch. Immunol. Ther. Exp.* 61.1:25–41. DOI: 10.1007/s00005-012-0203-0.

Berg RA van den, Hoefsloot HC, Westerhuis JA, Smilde AK, and Werf MJ van der. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 7.1:142. DOI: 10.1186/1471-2164-7-142.

Bishop CM. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 0387310738.

Bolstad BM. 2015. *preprocessCore: A collection of pre-processing functions*. R package version 1.28.0.

Borgognone MG, Bussi J, and Hough G. 2001. Principal component analysis in sensory analysis: covariance or correlation matrix? *Food Quality and Preference* 12.5-7:323–326. DOI: 10.1016/s0950-3293(01)00017-9.

Brennan C, Verhaak R, McKenna A, Campos B, Noushmehr H, Salama S, Zheng S, Chakravarty D, Sanborn JZ, Berman S, Beroukhim R, Bernard B, Wu CJ, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla S, Ciriello G, Yung W, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner D, Van Meir E, Prados M, Sloan A, Black K, Eschbacher J, Finocchiaro G, Friedman W, Andrews D, Guha A, Iacocca M, O'Neill B, Foltz G, Myers J, Weisenberger D, Penny R, Kucherlapati R, Perou C, Hayes DN, Gibbs R, Marra M, Mills G, Lander E, Spellman P, Wilson R, Sander C, Weinstein J, Meyerson M, Gabriel S, Laird P, Haussler D, Getz G, and Chin L. 2013. The Somatic Genomic Landscape of Glioblastoma. *Cell* 155.2:462–477. DOI: 10.1016/j.cell.2013.09.034.

Bühlmann P, Rütimann P, Geer S van de, and Zhang CH. 2013. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference* 143.11:1835–1858. DOI: 10.1016/j.jspi.2013.05.019.

Campos-Valenzuela JA. 2015. *Fast FMRLasso Repository*. URL: https://github.com/tugash/fastfmrlasso (visited on 08/01/2016).

Carlson M. 2016. *org.Hs.eg.db: Genome wide annotation for Human*. R package version 3.3.0.

Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, and Tyers M. 2015. The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 43.Database issue:D470–D478. DOI: 10.1093/nar/gku1204.

Cho DY, Kim YA, and Przytycka TM. 2012. Network biology approach to complex diseases. *PLoS Comput Biol* 8.12:e1002820. DOI: 10.1371/journal.pcbi.1002820.

Chuang HY, Lee E, Liu YT, Lee D, and Ideker T. 2007. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3.1. DOI: 10.1038/msb4100180.

Cloughesy TF, Cavenee WK, and Mischel PS. 2014. Glioblastoma: From Molecular Pathology to Targeted Treatment. *Annu. Rev. Pathol. Mech. Dis.* 9.1:1–25. DOI: 10.1146/annurev-pathol-011110-130324.

Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, and D'Eustachio P. 2014. The Reactome pathway knowledgebase. *Nucleic Acids Res* 42.Database issue:D472–D477. DOI: 10.1093/nar/gkt1102.

Eddelbuettel D. 2013. *Seamless R and C++ integration with Rcpp*. Springer.

Eddelbuettel D and François R. 2011. Rcpp : Seamless R and C++ Integration. *Journal of Statistical Software* 40.8:1–18. DOI: 10.18637/jss.v040.i08.

Eddelbuettel D and Sanderson C. 2014. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis* 71:1054–1063. DOI: 10.1016/j.csda.2013.02.005.

Eisenreich S, Abou-El-Ardat K, Szafranski K, Campos Valenzuela JA, Rump A, Nigro JM, Bjerkvig R, Gerlach EM, Hackmann K, Schröck E, Krex D, Kaderali L, Schackert G, Platzer M, and Klink B. 2013. Novel CIC point mutations and an exon-spanning, homozygous deletion identified in oligodendroglial tumors by a comprehensive genomic approach including transcriptome sequencing. *PLoS One* 8.9:e76623. DOI: 10.1371/journal.pone.0076623.

Faraway JJ. 2014. *Linear models with R*. CRC Press.

Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, Teague JW, Stratton MR, McDermott U, and Campbell PJ. 2015. COSMIC: exploring the world's knowledge of somatic muta-

tions in human cancer. *Nucleic Acids Res* 43.D1:D805–D811. DOI: 10.1093/nar/gku1075.

Goodenberger ML and Jenkins RB. 2012. Genetics of adult glioma. *Cancer Genetics* 205.12:613–621. DOI: 10.1016/j.cancergen.2012.10.009.

Greenblum SI, Efroni S, Schaefer CF, and Buetow KH. 2011. The PathOlogist: an automated tool for pathway-centric analysis. *BMC Bioinformatics* 12.1:133. DOI: 10.1186/1471-2105-12-133.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, deFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, and Stratton MR. 2007. Patterns of somatic mutation in human cancer genomes. *Nature* 446.7132:153–158. DOI: 10.1038/nature05610.

Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, Calderaro J, Bioulac-Sage P, Letexier M, Degos F, Clément B, Balabaud C, Chevet E, Laurent A, Couchy G, Letouzé E, Calvo F, and Zucman-Rossi J. 2012. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nature Genetics* 44.6:694–698. DOI: 10.1038/ng.2256.

Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, Chu CF, Huang HY, Lin CM, Ho SY, Jian TY, Lin FM, Chang TH, Weng SL, Liao KW, Liao IE, Liu CC, and Huang HD. 2014. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* 42.Database issue:D78–D85. DOI: 10.1093/nar/gkt1266.

Huang N, Shah PK, and Li C. 2012. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Briefings in Bioinformatics* 13.3:305–316. DOI: 10.1093/bib/bbr056.

Illumina Inc. 2014. *Infinium HumanMethylation450K BeadChip Support - Downloads*. URL: http://support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit/downloads.html (visited on 08/01/2016).

Kamburov A, Stelzl U, Lehrach H, and Herwig R. 2012. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* 41.D1:D793–D800. DOI: 10.1093/nar/gks1055.

Khalili A and Chen J. 2007. Variable Selection in Finite Mixture of Regression Models. *Journal of the American Statistical Association* 102.479:1025–1038. DOI: 10.1198/016214507000000590.

Kleihues P and Ohgaki H. 1999. Primary and secondary glioblastomas: From concept to clinical diagnosis. *Neuro-Oncology* 1.1:44–51. DOI: 10.1215/15228517-1-1-44.

Klink B, Miletic H, Stieber D, Huszthy PC, Campos Valenzuela JA, Valenzuela JAC, Balss J, Wang J, Schubert M, Sakariassen PØ, Sundstrøm T, Torsvik A, Aarhus M, Mahesparan R, von Deimling A, Kaderali L, Niclou SP, Schröck E, Bjerkvig R, and Nigro JM. 2013. A novel, diffusely infiltrative xenograft model of human anaplastic oligodendroglioma with mutations in FUBP1, CIC, and IDH1. *PLoS One* 8.3:e59773. DOI: 10.1371/journal.pone.0059773.

Kotsiantis S, Kanellopoulos D, and Pintelas P. 2006. Data preprocessing for supervised learning. *International Journal of Computer Science* 1.2:111–117.

Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, and Børresen-Dale AL. 2014. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer* 14.5:299–313. DOI: 10.1038/nrc3721.

Lockhart R, Taylor J, Tibshirani RJ, and Tibshirani R. 2014. A significance test for the lasso. *Ann. Statist.* 42.2:413–468. DOI: 10.1214/13-aos1175.

Lopez-Romero P. 2016. *AgiMicroRna: Processing and Differential Expression Analysis of Agilent microRNA chips*. R package version 2.16.0.

Louhimo R, Lepikhova T, Monni O, and Hautaniemi S. 2012. Comparative analysis of algorithms for integration of copy number and expression data. *Nature Methods* 9.4:351–355. DOI: 10.1038/nmeth.1893.

Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, Scheithauer BW, and Kleihues P. 2007. The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathologica* 114.2:97–109. DOI: 10.1007/s00401-007-0278-6.

Low J, Blosser W, Dowless M, Ricci-Vitiani L, Pallini R, Maria R de, and Stancato L. 2012. Knockdown of Ubiquitin Ligases in Glioblastoma Cancer Stem Cells Leads to Cell Death and Differentiation. *Journal of Biomolecular Screening* 17.2:152–162. DOI: 10.1177/1087057111422565.

Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, D'Eustachio P, and Stein L. 2012. Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers (Basel)* 4.4:1180–1211. DOI: 10.3390/cancers4041180.

Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloski CE, Sulman EP, Bhat KP, Verhaak RG, Hoadley KA, Hayes DN, Perou CM, Schmidt HK, Ding L, Wil-

son RK, Berg DVD, Shen H, Bengtsson H, Neuvial P, Cope LM, Buckley J, Herman JG, Baylin SB, Laird PW, and Aldape K. 2010. Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer Cell* 17.5:510–522. DOI: 10.1016/j.ccr.2010.03.017.

Ohgaki H. 2005. Genetic pathways to glioblastomas. *Neuropathology* 25.1:1–7. DOI: 10.1111/j.1440-1789.2004.00600.x.

Ohgaki H and Kleihues P. 2007. Genetic Pathways to Primary and Secondary Glioblastoma. *The American Journal of Pathology* 170.5:1445–1453. DOI: 10.2353/ajpath.2007.070011.

– 2013. The Definition of Primary and Secondary Glioblastoma. *Clinical Cancer Research* 19.4:764–772. DOI: 10.1158/1078-0432.CCR-12-3002.

Ostrom QT, Bauchet L, Davis FG, Deltour I, Fisher JL, Langer CE, Pekmezci M, Schwartzbaum JA, Turner MC, Walsh KM, Wrensch MR, and Barnholtz-Sloan JS. 2014. The epidemiology of glioma in adults: a "state of the science" review. *Neuro-Oncology* 16.7:896–913. DOI: 10.1093/neuonc/nou087.

Ostrom QT, Gittleman H, Fulop J, Liu M, Blanda R, Kromer C, Wolinsky Y, Kruchko C, and Barnholtz-Sloan JS. 2015. CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2008-2012. *Neuro-Oncology* 17.suppl 4:iv1–iv62. DOI: 10.1093/neuonc/nov189.

Peng J, Zhu J, Bergamaschi A, Han W, Noh DY, Pollack JR, and Wang P. 2010. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* 4.1:53–77. DOI: 10.1214/09-aoas271.

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordóñez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, Cox AJ, Edkins S, Kokko-Gonzales PI, Gormley NA, Grocock RJ, Haudenschild CD, Hims MM, James T, Jia M, Kingsbury Z, Leroy C, Marshall J, Menzies A, Mudie LJ, Ning Z, Royce T, Schulz-Trieglaff OB, Spiridou A, Stebbings LA, Szajkowski L, Teague J, Williamson D, Chin L, Ross MT, Campbell PJ, Bentley DR, Futreal PA, and Stratton MR. 2010. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463.7278:191–196. DOI: 10.1038/nature08658.

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, and Eisenberg D. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 32.Database issue:D449–D451. DOI: 10.1093/nar/gkh086.

Savage RS, Ghahramani Z, Griffin JE, Kirk P, and Wild DL. 2013. Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data. *arXiv preprint arXiv:1304.3577*.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, and Buetow KH. 2009. PID: the Pathway Interaction Database. *Nucleic Acids Res* 37.Database issue:D674–D679. DOI: 10.1093/nar/gkn653.

Scherer H. 1940. A critical review: the pathology of cerebral gliomas. *Journal of Neurology, Neurosurgery & Psychiatry* 3.2:147–177. DOI: 10.1136/jnnp.3.2.147.

Setty M, Helmy K, Khan AA, Silber J, Arvey A, Neezen F, Agius P, Huse JT, Holland EC, and Leslie CS. 2012. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. *Mol Syst Biol* 8.1. DOI: 10.1038/msb.2012.37.

Shen L and Tan EC. 2005. Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data. *IEEE/ACM Trans. Comput. Biol. and Bioinf.* 2.2:166–175. DOI: 10.1109/tcbb.2005.22.

Siegel RL, Miller KD, and Jemal A. 2015. Cancer statistics, 2015. *CA: A Cancer Journal for Clinicians* 65.1:5–29. DOI: 10.3322/caac.21254.

Städler N, Bühlmann P, and Van De Geer S. 2010. ℓ1-penalization for mixture regression models. *TEST* 19.2:209–256. DOI: 10.1007/s11749-010-0197-z.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, and Mesirov JP. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102.43:15545–15550. DOI: 10.1073/pnas.0506580102.

Sun Z, Asmann YW, Kalari KR, Bot B, Eckel-Passow JE, Baker TR, Carr JM, Khrebtukova I, Luo S, Zhang L, Schroth GP, Perez EA, and Thompson EA. 2011. Integrated Analysis of Gene Expression, CpG Island Methylation, and Gene Copy Number in Breast Cancer Cells by Deep Sequencing. *PLoS ONE* 6.2. Ed. by Z Zhao:e17490. DOI: 10.1371/journal.pone.0017490.

Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim Js, Kim CJ, Kusanovic JP, and Romero R. 2009. A novel signaling pathway impact analysis. *Bioinformatics* 25.1:75–82. DOI: 10.1093/bioinformatics/btn577.

The Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455.7216:1061–1068. DOI: 10.1038/nature07385.

– 2014. *The Cancer Genome Atlas*. URL: http://cancergenome.nih.gov/ (visited on 08/01/2016).

The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmule-

vich I, Sander C, and Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 45.10:1113–1120. DOI: 10.1038/ng.2764.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*:267–288.

Van Meir EG, Hadjipanayis CG, Norden AD, Shu HK, Wen PY, and Olson JJ. 2010. Exciting New Advances in Neuro-Oncology: The Avenue to a Cure for Malignant Glioma. *CA: A Cancer Journal for Clinicians* 60.3:166–193. DOI: 10.3322/caac.20069.

Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, and Hayes DN. 2010. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17.1:98–110. DOI: 10.1016/j.ccr.2009.12.020.

Vlachostergios PJ, Voutsadakis IA, and Papandreou CN. 2012. The ubiquitin-proteasome system in glioma cell cycle control. *Cell Div* 7.1:18. DOI: 10.1186/1747-1028-7-18.

Wong N and Wang X. 2015. miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res* 43.Database issue:D146–D152. DOI: 10.1093/nar/gku1104.

Xue H, Xian B, Dong D, Xia K, Zhu S, Zhang Z, Hou L, Zhang Q, Zhang Y, and Han JDJ. 2007. A modular network model of aging. *Mol Syst Biol* 3.1:147. DOI: 10.1038/msb4100189.

Zhang J. 2016. *CNTools: Convert segment data into a region by sample matrix to allow for other high level computational analyses.* R package version 1.28.0.

Zhang X, Zhang W, Cao WD, Cheng G, and Zhang YQ. 2012. Glioblastoma multiforme: Molecular characterization and current treatment strategy (Review). *Experimental and Therapeutic Medicine* 3.1:9–14. DOI: 10.3892/etm.2011.367.

Zhu J and Hastie T. 2004. Classification of gene microarrays by penalized logistic regression. *Biostatistics* 5.3:427–443. DOI: 10.1093/biostatistics/kxg046.

ANLAGE 1

Technische Universität Dresden
Medizinische Fakultät Carl Gustav Carus
Promotionsordnung vom 24. Juli 2011

Erklärungen zur Eröffnung des Promotionsverfahrens

1 Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

2 Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts habe ich Unterstützungsleistungen von folgenden Personen erhalten:

- Prof. Dr. rer. nat. Lars Kaderali
- Dr. Diana Clausznitzer
- Allen Lister
- Stefan Neumann

3 Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

4 Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

5 Die Inhalte dieser Dissertation wurden in folgender Form veröffentlicht:

- -

6 Ich bestätige, dass es keine zurückliegenden erfolglosen Promotionsverfahren gab.

7 Ich bestätige, dass ich die Promotionsordnung der Medizinischen Fakultät der Technischen Universität Dresden anerkenne.

8 Ich habe die Zitierrichtlinien für Dissertationen an der Medizinischen Fakultät der Technischen Universität Dresden zur Kenntnis genommen und befolgt.

Dresden, 28. August 2018

Unterschrift des Doktoranden

ANLAGE 2

Hiermit bestätige ich die Einhaltung der folgenden aktuellen gesetzlichen Vorgaben im Rahmen meiner Dissertation (Nicht angekreuzte Punkte sind für meine Dissertation nicht relevant.)

☐ das zustimmende Votum der Ethikkommission bei Klinischen Studien, epidemiologischen Untersuchungen mit Personenbezug oder Sachverhalten, die das Medizinproduktegesetz betreffen Aktenzeichen der zuständigen Ethikkommission.

☐ die Einhaltung der Bestimmungen des Tierschutzgesetzes Aktenzeichen der Genehmigungsbehörde zum Vorhaben/zur Mitwirkung.

☐ die Einhaltung des Gentechnikgesetzes.

☒ die Einhaltung von Datenschutzbestimmungen der Medizinischen Fakultät und des Universitätsklinikums Carl Gustav Carus.

Dresden, 28. August 2018

Unterschrift des Doktoranden