PhD-FSTM-2022-058
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 06/05/2022 in Esch-Sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN BIOLOGIE

by

Romain TCHING CHI YEN
Born on 02 March 1993 in Papeete (France)

SYSTEMS METHODS FOR ANALYSIS OF HETEROGENEOUS GLIOBLASTOMA DATASETS TOWARDS ELUCIDATION OF INTER-TUMOURAL RESISTANCE PATHWAYS AND NEW THERAPEUTIC TARGETS

Dissertation defence committee

Dr. Reinhard SCHNEIDER, dissertation supervisor
Professor, Université du Luxembourg

Dr. Emma SCHYMANSKI, Chair
Associate Professor, Université du Luxembourg

Dr. Enrico GLAAB, Vice Chairman
Assistant Professor, Université du Luxembourg

Dr. Simone NICLOU
Professor, Luxembourg Institute of Health


ITTM
Information Technology
for Translational Medicine


Luxembourg Centre
for Systems Biomedicine

Affidavit

I hereby confirm that my thesis entitled ***Systems methods for analysis of heterogeneous Glioblastoma datasets towards elucidation of inter-tumoural resistance pathways and new therapeutic targets*** is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Esch-Sur-Alzette, _____

TCHING CHI YEN Romain

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 766069. I am grateful for the opportunity I was provided with, and for the experience I earned from it.

I would like to thank my colleagues at ITTM and LCSB, who provided scientific and technical assistance throughout my PhD. In particular, Kerstin Neininger, Nils Christian and Serge Eifes who have been especially present and helpful on all fronts. In addition, I'm grateful for the help Marek Ostaszewski provided for the Glioblastoma Disease Map, and for Roland Krause's assistance in guiding the writing process of the present manuscript.

I'm also addressing special thanks to my fellow GLIOTRAIN ESRs and their PIs, for working with me on this journey. The work presented here would not have been the same without the contribution of Gonca Dilcan, who performed the sequencing of the GLIOTRAIN samples at VIB and worked hard with me to extensively characterize and document the resulting data, and without Ioannis Ntafoulis and Martine Lamfers who provided the data and pharmacological knowledge necessary for our conjoint work in the EMC collaboration.

I am grateful as well to members of my CET committee Enrico Glaab and Emma Schymanski for the feedback and follow-up they provided over the four years of my studies.

Finally, this project was only possible thanks to the guidance provided by my co-supervisors Reinhard Schneider and Andreas Kremer, to whom I'm indebted for offering me this PhD position.

Dedications

A mes amis proches, Etienne, Clément, Maxence, Laurence, Juliette, Alice et Jason, qui m'ont soutenu et encouragé tout au long de mes travaux.

A ma sœur et mon frère, Tiphaine et Evan, qui m'ont permis de me changer les idées alors que j'en avais bien besoin.

A Morgane, qui m'a fait garder le sourire jusqu'au bout.

A ma mère Katia, qui a cru en moi, m'a relevé à chaque mauvaise passe et ne m'a pas laissé abandonner.

A mon père Bernard, qui nous a quitté trop tôt, et à la fierté qu'il aurait de voir ces mots.

Table of Contents

List of illustrations, figures, tables

Figures

Tables

List of abbreviations

AKT	**AKT serine/threonine kinase**
AUC	Area under the Curve
ACSN	Atlas of Cancer Signaling Network
C_max	maximum nontoxic dose for a drug
CDM	Common Data Model
CNS	Central Nervous System
CNV	Copy Number Variation
CRF	Case Report Form
CTI	Cancer Trials Ireland
DASL	cDNA-mediated Annealing, Selection, extension, and Ligation
DEA	Differential Expression Analysis
DMSO	Dimethyl sulfoxide
DNA	Desoxyribo-Nucleic Acid
EMC	Erasmus Medical Center
EMT	Epithelial-to-Mesenchymal Transition
ESR	Early Stage Researcher
ETL	Extract, Transform, Load
FAIR	Findability, Accessibility, Interoperability, Reusability
FDA	Food and Drugs Administration
FDR	False Discovery Rate
FF	Fresh Frozen
GDPR	General Data Protection Regulation
GLIOTRAIN	Exploiting GLIOblastoma intractability to address European research TRAINing needs in translational brain tumour research, cancer systems medicine and integrative multi-omics
IC50	half maximal Inhibitory Concentration, concentration of the drug at which 50% of the cell population has died following drug exposure
ICM	Institut du Cerveau et de la Moelle épinière
IDH	Isocitrate Dehydrogenase
IPA	Ingenuity Pathway Analysis
IT	Intermediate-Term [survivors]
ITN	Innovative Training Network
ITTM	Information Technologies for Translational Medicine
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	Knockout
LASSO	Least Absolute Shrinkage and Selection Operator
LCSB BioCore	Luxembourg Centre for Systems Biomedicine - Bioinformatics Core
LIH	Luxembourg Institute of Health
LT	Long-Term [survivors]
MGMT	O-6-Methylguanine-DNA methyltransferase
MINERVA	Molecular Interaction NEtwoRks VisuAlization
miRNA	micro-Ribo-Nucleic Acid
mRNA	messenger Ribo-Nucleic Acid
MSCA	Marie Sklodowska-Curie Action
NGS	Next Generation Sequencing
OMOP	Observational Medical Outcomes Partnership

OHDSI	Observational Health Data Sciences and Informatics
OS	Overall Survival
PCA	Principal Component Analysis
PFS	Progression Free Survival
PhD	Philosophical Doctorate
PI3K	Phosphatidylinositol-4,5-biphosphate 3-kinase
PMID	PubMed ID
RB	Retinoblastoma
RCSI	Royal College of Surgeons in Ireland
RPPA	Reverse Phase Protein Array
RTK	Receptor Tyrosine Kinase
SBGN	Systems Biology Graphical Notation
SBML	Systems Biology Markup Language
SIGNOR	SIGnaling Network Open Resource
SNP	Single Nucleotide Polymorphism
SOC	Standard Of Care
SOP	Standard Operations Procedure
ST	Short-Term [survivors]
TCGA	The Cancer Genome Atlas
TMZ	Temozolomide
TP53	Tumor Protein p53
WGCNA	Weighted Gene Co-expression Network Analysis
WGS	Whole Genome Sequencing
WHO	World Health Organization

1 Summary

Glioblastoma is the most common and lethal type of brain cancer. Despite an aggressive treatment of maximal resection followed by radiotherapy and concomitant chemotherapy, the median survival time of patients after diagnosis is of 14.6 months, and there have not been any major advances in treatment to improve prognostics or the patients' quality of life since 2005. As a consequence, there is an urgent need to improve knowledge about this cancer and understand the mechanisms involved in its resistance to treatment. To that end, in this thesis I set out to uncover the molecular mechanisms underlying IDH-wildtype Glioblastoma resistance to treatment, through a multi-omics integrative analysis making use of both quantitative data and established knowledge about the disease.

First, a major part of these studies involved the manual screening and curation of the literature to identify the core driver alterations of functional pathways in IDH-wildtype Glioblastoma, focusing on the RTK/PI3K/AKT signaling cascade, the RB pathway and the TP53 pathway. The results of that investigation were compiled into a Glioblastoma Disease Map, a visual and interactive network representation of these pathways in the context of Glioblastoma made publicly available on the MINERVA platform. Furthermore, this work led to the definition of novel modelling standards for genetic alterations in the Disease Map framework, which may be further developed and used by the community.

This Glioblastoma Disease Map network was analyzed alongside whole-transcriptome and whole-genome sequencing data to investigate resistance mechanisms of Glioblastoma. From that analysis, interconnection patterns between the three pathways in the disease maps could be highlighted and discussed, and the emergence of cell motility as a critical part of resistance mechanisms was proposed. In addition, a collaboration with a laboratory from the Erasmus Medical Center in Rotterdam led to multiple analyses of transcriptomics towards the identification of biomarkers predictive of drug response, and the publication of an article as a co-author.

Finally, over the PhD multiple database systems and frameworks such as the tranSMART data warehouse or the OMOP Common Data Model were encountered and used to develop and implement data management processes in line with FAIR data principles, data privacy, and downstream data analysis considerations. The results and experience acquired thanks to that work provided valuable insights for the implementation of quantitative analyses, but also more broadly for the proper conduct research.

2 Introduction

2.1 Glioblastoma

Glioblastoma is the most common and lethal type of brain cancer. Over the period of 2012-2016 in the United States[1], Glioblastoma represented 48.3% of all malignant brain and other Central Nervous System (CNS) tumors, corresponding to an incidence of 3.42 per 100,000 population. The median survival rate of untreated patients is barely 3 months[2], while treatment allows to increase it to 14.6 months[3]. Still, only 6.8% of patients reach the 5-years mark[1], making the disease extremely fast and brutal.

Glioblastoma, classified as a World Health Organization (WHO) Grade IV tumour[4,5], is mainly diagnosed in older people with a median age of diagnosis of 65[1], and affect males slightly more as they represent an average of 57.8% of new yearly cases[1].

Molecular profiling of Glioblastoma tumours have greatly impacted knowledge about the disease, to the extent that the WHO recommends including both histological and molecular considerations for Glioblastoma diagnosis[4,5].

An important classification first reported by Kleihues and Ohgaki (1999)[6] is the distinction between Primary and Secondary glioblastomas. Note that in the latest classification of the World Health Organization from 2016[4] and 2021[5], the terminology of Primary and Secondary Glioblastoma has been abandoned, the terms of IDH-wildtype and IDH-mutant being preferred, respectively, as Isocitrate Dehydrogenase (IDH) mutation commonly allows discrimination between the two types. However, since these labels were used in the context of my studies in the data that I worked with, in the present document the terminology will be kept when it was explicitly used in data and documents that were provided for the PhD work. IDH-wildtype Glioblastomas are tumours emerging *de novo* in the brain. They represent the large majority of Glioblastoma cases[7], occur in patients on average 64 years old, and is the more aggressive type of Glioblastoma. These tumours typically present overexpression and mutations of the *EGFR* gene, deletion of the *CDKN2A* locus and amplification of the *MDM2* gene. IDH-mutant Glioblastomas are tumours that evolved from lower-grade gliomas into more aggressive lesions, although they still present a better prognostic than IDH-wildtype Glioblastomas, and affect a younger part of the population as well since the patients diagnosed with it are 45 years old on average. IDH-mutant Glioblastomas are characterized by mutations of the *TP53, IDH1* and *IDH2* genes.

Besides, research into molecular pathways and characteristics of Glioblastoma has been productive in the description of genes and pathways relevant to the development and survival

of Glioblastoma tumours[8–13]. In particular for IDH-wildtype Glioblastomas, these findings point to the Receptor Tyrosine Kinase (RTK)/PI3K/AKT, Retinoblastoma (RB) and TP53 pathways as key drivers of the disease.

The first step for the Standard Of Care (SOC) for Glioblastoma consists in surgical resection of the tumour at diagnosis. Resection of the tumour presents a huge challenge since brain surgery is a delicate operation which typically does not allow for full removal of the highly diffuse tumour tissue[14,15]. As a result, tumour cells remain in the patient's brain even after surgery and invariably lead to recurrence of the cancer, *i.e.* emergence of a new tumour, which explains the short median survival rate of Glioblastoma patients. The surgery is then followed by treatment of radiotherapy and concomitant Temozolomide (TMZ) chemotherapy[16], called the Stupp regimen, which delays recurrence of the tumor but does not prevent it. TMZ is an alkylating agent capable of methylating DNA leading to cell death[17]. In the context of TMZ therapy, the methylation of the MGMT (O-6-Methylguanine-DNA methyltransferase) promoter constitute an important biomarker for positive prognostic of the patient. Indeed, MGMT is a methyl-transferase able to revert the effects of TMZ on DNA[18,19], thus mitigating the efficacy of the therapy.

However, while there are extensive efforts to develop new treatments[15,20–22] in particular targeting specific molecular mechanisms mentioned above, since establishment of concomitant TMZ chemotherapy into the standard of care in 2005, there has been no major improvements in Glioblastoma treatments and patients' survival. As a consequence, there is a significant need to improve our understanding of Glioblastoma mechanisms and identify potential therapeutic targets.

In this context, the GLIOTRAIN ("Exploiting **GLIO**blastoma intractability to address European research **TRAIN**ing needs in translational brain tumour research, cancer systems medicine and integrative multi-omics") project was initiated. It is a Horizon 2020, Marie Sklodowska-Curie Action (MSCA) Innovative Training Network (ITN) involving 8 beneficiaries and 12 partner organizations with the specific goal of training 15 Early-Stage Researchers (ESRs) on research on IDH-wildtype Glioblastoma towards the development of new treatments and the elucidation of Glioblastoma resistance mechanisms. Furthermore, although each individual ESR has their own PhD project, they all fit within the overarching objectives of GLIOTRAIN, and were designed to be interconnected. In particular, one of the major milestones for the GLIOTRAIN project was the generation of Next Generation Sequencing (NGS) data from Glioblastoma samples coming from several beneficiaries in the project. In addition to the clinical characterization of the patients from which those samples were extracted, sequencing should

lead to the production of transcriptomics RNA-Seq data, proteomics RPPA (Reverse Phase Protein Array) data, low-coverage whole-genome sequencing (WGS) data and methylation data. The whole pipeline, from selection of the samples to the creation of the final GLIOTRAIN database, involved several ESRs. This data was collected for analysis in several PhD projects, including mine.

Within GLIOTRAIN, my PhD project aimed at defining and performing an integrative analysis combining both the knowledge already available about the disease and the data generated by the project in order to improve understanding of the molecular resistance mechanisms developed by the tumour. My studies led me to focus on three main axes of work: the compilation of literature knowledge through the construction of a Glioblastoma Disease Map, the concurrent analysis of multiple sources of data, and implementation of good data curation and data management practices.

2.2 Disease Map

In the era of high-throughput technologies producing large amount of data and leading to an ever-increasing accumulation of biological knowledge, representation of that knowledge as networks of biological entities have gained attention[23,24] as a solution to compile, interconnect, visualize and analyze information at different levels. Indeed, biological network can represent interactions at the level of organs, tissues, cells, proteins, DNA, or even between amino acids. Moreover, biological networks can be inferred or enriched based on quantitative data[25,26], or analyzed as a standalone resource[27,28]. Even more interesting, computational methods exist to account for the biological knowledge in the form of networks as prior knowledge in the integrative analysis of multiple omics source of data[29–33], resulting in more context-relevant findings.

Importantly, such networks can be published and shared for the scientific community to use in research. Notable resources for biological network at the molecular level include the STRING[34,35] and KEGG[36,37] (Kyoto Encyclopedia of Genes and Genomes) databases, the SIGnaling Network Open Resource (SIGNOR) 2.0[38] or the Atlas of Cancer Signaling Network (ACSN)[39]. These networks were created through the extensive curation of a high number of publications, resulting in the systematic compilation of knowledge following clearly defined formats and standards. As a result, anyone who uses one of these networks would need to understand and align on the corresponding framework.

One such framework is the Disease Maps project[40,41]. Disease Maps are a type of integrated and highly-curated molecular network representation of signaling and metabolic pathways for specific diseases, relying on standard formats such as the Systems Biology Graphical Notation[42] (SBGN) which defines guidelines for graphical representation of biological entities and their interactions, and the Systems Biology Markup Language[43] (SBML) which is a verbose syntax for representation of biological networks, which can be used by softwares to read or export network definitions. Construction of Disease Maps is a community-driven effort, requiring both extensive screening of the literature and input from experts on the disease. Moreover, the usage of clearly defined standards and support from an active community allows for the development of powerful network exploration and analysis tools[44–50].

On the downside, the Disease Maps framework is not well equipped to extensively represent mutations. Indeed, literature about genetic alterations in Disease Maps did not bear productive results, and consultation with experts on the framework confirmed that Disease Maps projects usually focus on protein-level interactions and mechanisms. Alterations at the genetic level are considered as implicit since their effect are typically echoed as mutated proteins or up- or downregulated expression profile. However, that approach not only makes mutated genes implicit, but also does not support standalone representation of transcriptional rates modifications, which can then only be visualized by projecting quantitative data onto the network. In the context of Glioblastoma and cancer in general, where genetic alterations such as point mutations or copy number variation are key drivers of the disease, this absence of consensus on genetic alterations representation is a drawback.

Despite this challenge, the Disease Maps standard and associated tools still appeared as the most appropriate and accessible framework to undertake compilation of molecular interactions driving Glioblastoma development and survival, in order to visualize, explore, characterize, and integrate literature knowledge about Glioblastoma into a data analysis for the PhD project.

2.3 Data management methods and systems

With the rise of NGS data where large volumes of data are generated from many different sources and formats, data curation and management are essential to proper conduct of research. Indeed, the number of tools and methods, including statistical tests, machine learning algorithms, *etc.* available to analyze data is also in constant growth, and any given analysis often requires specific formatting, normalization, or pre-processing of the initial data before it can be used as an input. As a result, it is essential to apply proper transformations to the data and keep track of it, while also being aware of the impact such transformations can have on the data and analysis results.

In addition, proper data management should follow Findability, Accessibility, Interoperability, Reusability (FAIR) data principles[51], which provide a useful framework to ensure data quality and characterization, subsequently allowing for more reliable and robust results. Indeed, following these principles ensures that any results from an analysis can be reproduced and validated by reproducing the analysis with the same data, and that this data may be further exploited for new investigations. With a well-characterized database, research can be conducted more transparently and be easier to validate and advanced further. Simultaneously, to preserve patients' right to data privacy, data management processes are also required to include considerations of anonymization of data and secure data storing and sharing practices.

As part of the training both under the European ITN and under the ITTM (Information Technologies for Translational Medicine) company, during the PhD three different data management systems were encountered and used: the clinical trials databases from Cancer Trials Ireland, the tranSMART data warehouse for hosting the GLIOTRAIN data, and the rising *Observational Medical Outcomes Partnership* Common Data Model (OMOP CDM) used in data harmonization projects. By learning about these systems and how to handle and use them, valuable insights about the importance of data management in the conduct of research were gained.

2.3.1 Cancer Trials Ireland

A requirement of the GLIOTRAIN grant was for all ESRs to spend time away from their host institution to experience different working environments in an organization partner of the project. This was an opportunity which was used to discover and learn more about the very specific processes of clinical trials. As a consequence, a secondment was organized to Cancer Trials Ireland (CTI).

CTI is a non-profit company based in Dublin, Ireland, which supports and coordinates cancer-related clinical trials in Ireland and across Europe, and was a partner organization of GLIOTRAIN. The objectives of the secondment were to learn about how data was collected, stored and analyzed in the context of clinical trials, to use that knowledge in the processing of data from the GLIOTRAIN project. During that time, in addition to the general processes of conducting trials, specific training on data management methods as performed in CTI was received and applied hands-on to help directly on two CTI projects in particular: a study about identifying glioma biomarkers, and a trial to compare breast cancer treatments.

2.3.1.1 Glioma Biomarker study

This project was an observational study in which specific biomarkers data was collected from patients, in order to compare it with data from healthy controls and try to determine whether the level of any of these biomarkers could suggest onset of glioma in a non-invasive way and be included in routine checks.

2.3.1.2 Breast Cancer Trial

The breast cancer trial was a two-arms comparison of breast cancer treatment regimen, meant for statistical analysis. In this case, the CTI database for that trial was already relatively complete and well maintained, and the study was at the stage of early analysis through the characterization of the data.

2.3.2 TranSMART

The tranSMART system[52,53] can be considered as a data warehouse, which includes both a database to store and access data, as well as a graphical interface on top of it that allows for exploration and statistical analysis of the data from the database. It is a patient-centric system, meaning that every datapoint needs to be associated to a single patient characterized in the database.

TranSMART is designed for storing either “low-dimensional” data, *i.e.* single variables linked directly to a given patient in the database such as the clinical information of each patient, or “high-dimensional” data[54] which typically correspond to the different types of -omics data[55] that are supported and require a mapping between the variables and the samples in the dataset, and between the samples of the dataset and the patients in the database.

Beyond the database aspect of tranSMART, the system also provides a graphical interface platform to visualize, explore and analyze the data stored. For the sake of this interface, organization of the data in the database is required. Indeed, although tranSMART databases

have a fixed structure with schemas and tables pre-defined, when data is inserted in the database it has to be clearly annotated so that different datasets of the same data type can easily be differentiated, visualized and analyzed in the interface. Included in these required annotations is the position in a tree-like structure of the data, called “tranSMART data tree”, used for navigation and selection of specific data in the graphical interface.

As part of the GLIOTRAIN project, ITTM was tasked with the hosting of the data generated in the project for the purpose of analysis in multiple ESR PhD projects. The data was to be sequenced from samples taken from the biobank of four of the GLIOTRAIN beneficiaries: Royal College of Surgeons in Ireland (RCSI), Luxembourg institute of Health (LIH), Erasmus Medical Center, Netherlands (EMC) and Institut du Cerveau et de la Moelle epiniere, France (ICM). However, following definition of the materials required to contribute samples, the LIH was not able to provide paraffin sections for their samples which was a consortium requirement, and consequently they were not able to contribute samples. Fortunately, ICM and EMC were able to compensate for the number of samples that LIH was initially supposed to contribute. The resulting data to be made available on this GLIOTRAIN database included clinical information about the patients from which the samples were extracted, low-coverage WGS data, RNA-Seq data, Methylation data and Reverse Phase Protein Array (RPPA) data, thus providing data at the genomic, transcriptomic and proteomic level to the consortium. However, the RPPA dataset was eventually removed from the list and is absent from the final database.

2.3.3 The OMOP CDM

The *Observational Medical Outcomes Partnership* Common Data Model[56,57] (OMOP CDM) is a data model increasingly used to harmonize medical, clinical, and healthcare registries and databases[58–61] to align them with the same structures and standard, thus making them comparable and includable in the same analyses despite being from different sources. The OMOP CDM is maintained, developed and promoted by the Observational Health Data Sciences and Informatics (OHDSI) community[62,63], which provides resources, training, tools and methods to use the OMOP CDM and analyze harmonized data. In particular, the OHDSI community is a strong proponent of FAIR data principles, since it facilitates communication and meeting of registry representants who may be interested in conducting research conjointly on their data, and even develops tools to perform “network studies”[64–67], *i.e.* data analyses which can be performed in a federated, non-centralized way, where data of individual registries is analyzed locally instead of being shared, and only aggregated results are returned for interpretation, thus ensuring data security and anonymity of patients.

The OMOP CDM is characterized by both the structure of its database and the terminology it used for the data in it.

The CDM structure[68] has been designed to be able to store most medical and healthcare-related data, including diagnoses, treatments, lab measurements, but also billing or healthcare provider data. In addition, in order to limit loss of information in the process of converting data from the source database to the OMOP format, the OMOP tables typically include fields where the values as they appear in the source data can be inserted.

Furthermore, the OMOP CDM uses its own terminology, called the OMOP "Standardized Vocabularies"[69], to represent data. This terminology is constituted by numerical IDs called "concept IDs" representing the many biomedical concepts which can be encountered in converted databases. This ontology was constituted by assigning unique numerical IDs (*i.e.* the concept IDs) to all the codes in widely used standard ontologies such as ICD10[70], RxNorm[71], *etc.*

As a consequence, in order for a source registry to be included in network studies, it first needs to be mapped to the OMOP CDM both in structure and in terminology, a process called the Extract, Transform, Load (ETL) process. The OHDSI community has developed tools such as the Rabbit-in-a-Hat[72] and Usagi[73] softwares to perform mapping from the source structure and terminology to the OMOP structure and Standardized Vocabularies, respectively, as well as extensive documentation and guidelines to define these mappings[74,75].

Looking into the OMOP CDM was instigated for two reasons. First, it was briefly considered as an alternative to store the clinical data part of the GLIOTRAIN data. However, since that clinical data was relatively simple (only one timepoint) and could easily be included in the tranSMART database, it made little sense to separate the data into two different systems with different structures and environments.

Secondly, with the increasing number of projects that used the OMOP CDM or wanted a database to be converted to the CDM, ITTM as a company also took the opportunity to get closer to the OHDSI community and take part to OMOP-related projects pertaining to converting databases to the OMOP CDM. As such, these projects represented a great opportunity to get hands-on experience on Data Management and its impact on how data is shared and analyzed by multiple parties.

Through that involvement and experience, an understanding of key aspects of the process of mapping data to the OMOP CDM was acquired, documented, and used to lead efforts to

develop a methodology and tools to be used within ITTM to improve quality and speed of the team on mapping projects.

2.3.3.1 OMOP Processes Documentation

Through the work achieved on data management for the CTI, GLIOTRAIN tranSMART and ITTM OMOP database, the fact that standardized, reproducible, and well-documented processes were essential to ensure quality and reliability of the data was made clear. As a consequence, a major contribution to the efforts of streamlining OMOP mapping projects work achieved the identification of key parts of the process, bottlenecks, time- and resources-consuming tasks, *etc.* and the development of a methodology to standardize the OMOP mapping pipeline, mitigate bottlenecks and issue-triggering events, and document and track progress to make onboarding of new staff on a project easier. These suggested methods and processes were written as Standard Operating Procedures (SOPs) for the team to agree on a common work pipeline.

2.3.3.2 ETL Software

One important bottleneck in OMOP mapping projects is the implementation of an ETL script, *i.e.* the code that executes the mapping defined from the source data to the OMOP CDM. Since every source database is different, in theory each ETL needs to be tailored specifically to that source's data. This is a very stringent issue, since it means that

- new code needs to be implemented for each project,
- the person coding the ETL needs to know in detail the source data and mappings to the OMOP CDM defined, which results in either
    - the same person must both define the mappings and code the ETL, which is an issue since having the skills to do both is not trivial and requires extensive training
    - or there needs to be extensive knowledge transfer between the mapper and the ETL implementer, which is a very inefficient and time-consuming solution.

To overcome this issue, the ITTM team has implemented a Python program which takes a matrix containing mapping definitions as an input and automatically apply them across the source data. This provided a solution where both mapper and ETL implementer relied on the agreed structure of the matrix for the definition of the mappings, allowing for the handover from mapping to ETL execution to require minimal knowledge transfer.

However, that initial software still required significant adjustments based on the source data structure, and was not adapted to the Rabbit-in-a-Hat/Usagi tools pushed by the OHDSI community and used as part of the new ITTM SOPs for OMOP mapping project. As a result, within this thesis the program was developed further to make it compatible with the newly adopted methodology and tools.

2.3.3.3 Machine-readable Syntax

In particular, an important part of the work on making the ETL software more modular and independent from the source database structure of any given project was in the adaptation of a formal structure to define mappings, which was carried out through the matrix in the first version of the program. This structure had to have the same role, namely provide a standardized framework for defining the mappings that the mapper would apply, writing the mappings in a format usable as input by the ETL software.

The solution that was adopted was the creation of a programing-like syntax, which should be simple enough that mappers could write it even without extensive programming knowledge, while also strictly structured so that it may be read and interpreted by the ETL program, hence significantly reducing the work needed from a developer to execute the defined mappings.

2.4 Data Analysis

2.4.1 Identification of predictive biomarkers of drug response

Due to delays in the production of the GLIOTRAIN data following the European GDPR implementation, a collaboration was started with another GLIOTRAIN ESR from the EMC in Rotterdam, in the Netherlands. Through that collaboration, computational statistical analyses were performed to identify transcriptomics biomarkers predictive of drugs responses, and these analyses brought in the dimension of drugs response which was absent from the main GLIOTRAIN data. In addition, through this collaboration I obtained data to analyze and produce results to present should the delays in the production of the main GLIOTRAIN data extend for too long.

2.4.1.1 Drugs Repurposing Project

The collaboration initially focused on a study towards the repurposing towards Glioblastoma of drugs used for other diseases. The principle of the drug repurposing study was to expose 45 Glioblastoma (35 Primary and 10 Recurrent) cell cultures to several concentrations of a given drug, quantify the number of cells that survived at each concentration, compute the IC50 (concentration of the drug at which 50% of the cell population has died from drug exposure), and use these IC50 values to determine whether the cell cultures responded to the drug or not. This process was performed for 109 drugs that are not currently used for Glioblastoma treatment, with the goal of identifying which, if any, of these drugs had the potential to be used against Glioblastoma. Furthermore, analysis of the gene expression profile of the cell cultures was planned, with the hope of identifying predictive biomarkers for drug response, *i.e.* genes for which the expression profile in the tumour would suggest a drug would be effective for treating the patient.

2.4.1.2 Berkeley LASSO Analysis review

Initially, support on the analytical side of the Drugs Repurposing Project was provided by a partner of the EMC team at the University of California Berkeley, where a Least Absolute Shrinkage and Selection Operator (LASSO) analysis[76], a linear regression approach that allows for feature selection, was performed on the data . Since the exact methodology and parameters used in there had not been disclosed, it was decided that the results of that analysis would be compared to the results produced during the collaboration, to potentially validate the approach. To do so, as well as determine whether the Berkeley methodology could be reproduced from the limited knowledge available, the results of the Berkeley analysis were reviewed, and reproduction was attempted at two levels:

- using the Berkeley LASSO-selected genes to try and identify the same genes of interest
- compare the functional pathways corresponding to the Berkeley LASSO-selected genes to the pathways identified in the Drug Repurposing Project analysis

2.4.1.3 Validation of Glioblastoma cell culture models

Later during the PhD project, the collaboration with EMC was extended to another project, in order to produce analyses and figures to use in an article that had been in preparation for a long time, where the team argues for their models of Glioblastoma cell culture and shows that these cultures are representative of the parental tumor they are derived from, in particular in response to TMZ exposure which is the current standard of care drug for Glioblastoma.

For this, they mainly wanted to show correlation between the cell cultures' response to TMZ and response from the patients to treatment, represented by their Overall Survival (OS) and Progression Free Survival (PFS).

Furthermore, same as for the drug repurposing project, a secondary objective for this study was to see if predictive biomarkers to TMZ response with new updated drug response data could be identified.

2.4.2 GLIOTRAIN Data Analysis

Finally, an integrative analysis using the GLIOTRAIN and EMC data as well as the Glioblastoma Disease Map was planned to investigate resistance mechanisms of Glioblastoma. That analysis was set to take place in two stages.

First, analysis of each resource individually would allow to detect information about resistance mechanisms characteristic to each data type. To do so, Differential Expression Analyses (DEA) of each dataset was performed to determine significant differences in the molecular profiles of Short-Term (ST) survivors (overall survival < 12 months) and Long-Term (LT) survivors (overall survival > 36 months) of Glioblastoma. In addition, a network topology analysis should be performed on the Glioblastoma Disease Map in order to identify potential key components of the network that may be worth targeting for treatment of Glioblastoma.

Secondly, an analysis relying on network topology to orient data analysis should be performed to allow for the emergence of synergistic patterns between the datasets undetected by individual analysis.

However, due to time constraints only the DEA part of the analysis and a qualitative rather than computational analysis of the Glioblastoma Disease Maps network were effectively performed.

3 Scope and Aims of the Thesis

Through this thesis I am presenting the work and research I conducted during my PhD studies, towards elucidation of resistance mechanisms of Glioblastoma via an integrative multi-omics data analysis. This is done along three axes of research and reflection: the creation of a Glioblastoma Disease Map; the analysis of multiple Glioblastoma-related datasets of different omics type; and the documentation and automatization of data management-related processes and methods.

The first axis, about the Disease Map, describes how literature about the molecular mechanisms involved in Glioblastoma was screened to identify relevant pathways and interactions that should be included in a Glioblastoma-specific disease map. In this part of the thesis are included the methodology to select and extract molecular interactions, the choices on how to represent it within the Disease Map standards, the resulting network and insights that could be derived from the disease map on its own. It also presents efforts to extend the Disease Map framework to establish yet poorly covered considerations about representation of genetic alterations models.

Secondly, in the reflection about data management, concepts of extensive transparency, documentation and communication in the handling of data at the level of data management and how they may impact downstream analyses are explored via implementation of data management methods under three different systems environments: clinical trials databases, the tranSMART data warehouse and the OMOP CDM. It also includes both the steps taken for the management of the GLIOTRAIN data and in the improvement of efficiency of the Data Management processes.

Finally, the Data analysis part of the manuscript presents the processing, normalization and other transformations applied to the GLIOTRAIN and EMC data before analysis, details the different analyses methods considered and applied for the analysis for my PhD project as well as for the collaboration with EMC, and discusses the results these analyses produced in the perspective of uncovering Glioblastoma resistance mechanisms.

Figure 1 illustrates the different parts of this project and how they come together towards the objective of the thesis.

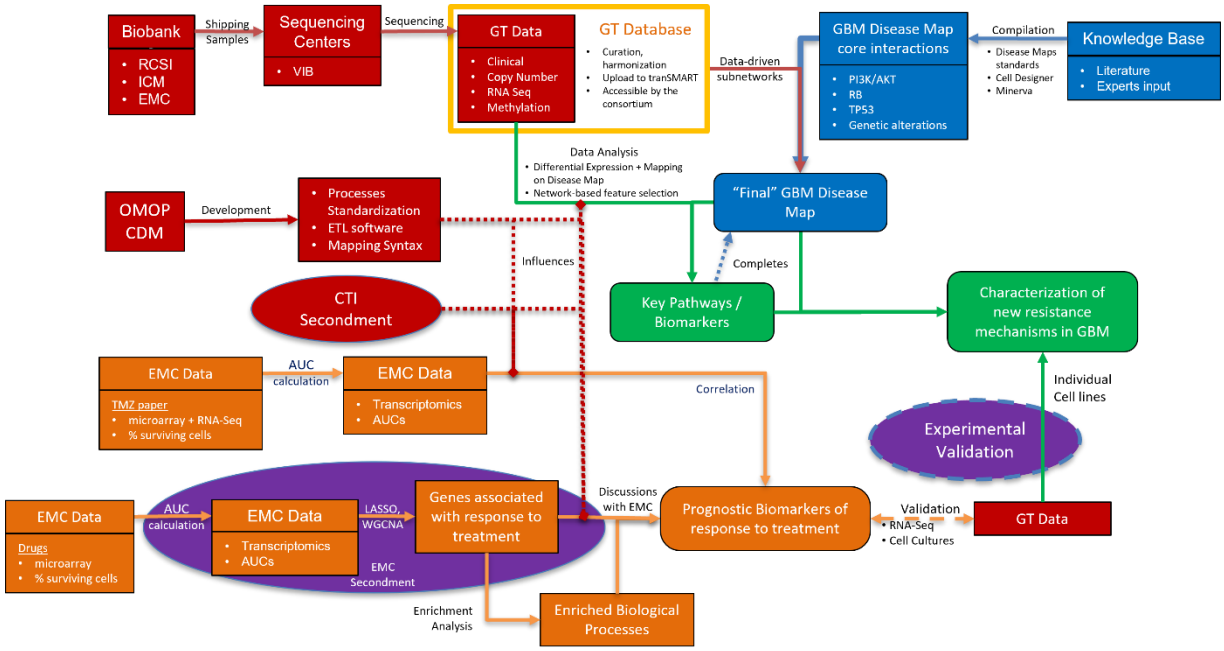


Figure 1: Overview of the components of the PhD. *Red: Data Management. Blue: Glioblastoma Disease Map. Orange: Identification of predictive biomarkers of drug response Analysis. Green: Integrative analysis. Oval shapes: Hosted away abroad*

4 Materials and Methods

All data manipulations and analyses were performed using the R language[77] in the RStudio software[78], unless specified otherwise.

Implementation of the described functions and analyses has been made publicly available on a GitHub repository at "https://github.com/RomainTching/phd" (this does not include the software described in the section 4.2.3.2 ETL software, which is ITTM Intellectual Property and cannot be published in this context). The corresponding files and functions will be referenced where relevant in the 4 Materials and Methods section.

4.1 Glioblastoma Disease Map

4.1.1 Literature screening

Articles used as a basis to build the Glioblastoma Disease Map were found through the PubMed website[79], searching for keywords relevant to the topic of interest, such as "Glioblastoma pathways", "mTorC1 targets" of "TP53 in Glioblastoma" for instance. For each search, up to three pages of the results were screened to select and read articles that appeared to be most relevant based on both title and abstract of the article.

To determine a starting point for building the Glioblastoma Disease Map, broad review papers about the molecular profile of Glioblastoma[8–13] were read. From there, the Glioblastoma Disease Map was started focusing on the three pathways recurrently identified in the articles as key drivers of Glioblastoma: RTK/PI3K/AKT pathway, the RB pathway and the TP53 pathway.

For each of these pathways, review articles on the pathway were sought out to model it in the Disease Map, starting with the proteins that are known to be mutated or have altered expression profiles in Glioblastoma. Moreover, the references of each review were also screened to identify other papers, both original research and review, that would be relevant to read afterwards.

Therefore, the pathways were first modelled based on reviews that described the interactions cascade on a broader level, generally in the way they normally unfold in a healthy context rather than in Glioblastoma. Glioblastoma-specific alterations were then integrated afterwards, when they concerned mutated proteins, since alterations in transcription rate is not directly representable at the protein-protein interactions level.

Furthermore, because the process of screening literature and building a disease map is very time-consuming, the decision was made to mainly focus on review articles to build the core of the Disease Map, and only look into original research papers afterwards if time allowed it, since progress on other aspects of the PhD would have been greatly impeded otherwise.

The RTK/PI3K/AKT pathway was started with how the *PI3K* and *AKT* proteins are activated[12,13,80,81], before looking into downstream effectors of the signaling cascade. In particular, pathways that may lead to typical cancer-altered functions[82,83] such as apoptosis, cell proliferation or angiogenesis for instance, were sought out and focused on. This led to the investigations on the mTORC1 pathway[84,85] which influences cell growth, the *FOXO* transcription factors family[86–88] which regulate transcription of many targets, some of which are involved in regulation of apoptosis and of the cell cycle, as well as papers describing interactions with effectors of the RTK/PI3K/AKT cascade that bridge to the TP53[89] and RB[90] pathways.
Later, the RAS/RAF/MEK/ERK signaling cascade[91,92] was identified to be of great interest and integrated in the RTK/PI3K/AKT submap since it is also initiated by RTK activation, and may act as an alternative path to activate *PI3K* as well.

For the RB pathway, focus was given to reviews[93–95] that described both regulation of RB and its functions to inhibit *E2F* transcription factors family.

Finally, while regulation mechanisms of *TP53*[89,96–98] are relatively consistent and well characterized, screening literature for the *TP53* transcriptional targets was tricky since it has many known targets but studies that would investigate Glioblastoma-specific relevant targets could not be found. As a result, transcription by *TP53* was assumed to be equally likely for all of its targets, and they were identified and inserted in the map based on the review from Bieging *et al.* (2014)[99]. Indeed, this review lists many of the *TP53* targets and categorizes them into the cellular function they are involved with, which allowed to produce a basis for the TP53 pathway with clear separation of the different downstream functions it regulates, so that the submap may easily be adjusted by adding or removing *TP53* transcriptional targets should a Glioblastoma-specific list of relevant targets be identified.

Furthermore, during research and reading it became clear that in addition to these three pathways, explicit representation of genetic alterations in Glioblastoma was also needed. Thus, genetic alterations were also integrated, based mainly on findings from the Glioblastoma-specific TCGA (The Cancer Genome Atlas) datasets[8,9], but also on other articles[10,12,13,100,101].

4.1.2 Genetic Alterations Models definition

In order to represent Glioblastoma-specific genetic alterations, new standardized models in line with the broader Disease Map standards needed to be developed. Since a common consensus and good practices on how to represent these genetic alterations could not be found, models that would fit the needs of this project, but also general enough that it could be recognized and considered for integration by the Disease Map community, had to be defined.

Thus, to define this model, the Systems Biology Graphical Notation (SBGN) standard was followed closely and representation of genetic alterations were defined within that framework. Three issues were solved in the definition of that model: the identification of genetic alteration types relevant to Glioblastoma, the representation of biological entities involved, and how to convey the genetic alterations as interactions between entities.

Genetic alterations relevant to Glioblastoma were determined to include chromosomal or gene amplifications or deletions, point mutations and patterns of mutually exclusive or co-occurring mutations. Point mutations can have a wide variety of effects with different amplitudes, however given the qualitative nature of Disease Map diagrams, modelling the relative strength of the effect of mutations is not possible, and representable effects themselves can be classified into four categories based on outcome of the mutation: increased transcription rate, increase protein efficiency, decreased transcription rate and decreased protein efficiency.

Biological entities involved in these alterations are either chromosomal regions, genes, mRNAs for transcription rate aberrations, or small-scale mutation (SNPs, small insertions or deletions) sites.

Chromosomal regions in copy number variations were represented as hypothetical complexes, whereas they should be containers when modelling other mutations involving genes contained in those regions.

Mutation sites are defined as a single modification site on the gene, which serves to represent all mutations of this gene that result in the same outcome. If other mutations of the gene lead to a different outcome (for instance, decreased transcription rate and decrease of protein efficiency), they are represented by a distinct modification site. Presence of any of the mutation linked to the modification site is represented by a ‘*’ symbol. This is a slight divergence from the standard, since this ‘*’ symbol in SBGN is supposed to represent that the modification site presents any of the alterations: ‘phosphorylation’, ‘methylation’ or ‘acetylation’. But since there is no dedicated ‘mutation’ annotation, this was chosen as placeholder. Furthermore, the list of

mutations corresponding to that outcome may be added as notes to the mutated gene, to retain that information in the map.

With these definitions, determining a representation for mutation-modelling interaction becomes simpler. A small-scale mutation can be the transition from wild-type to mutated alleles. Chromosomal aberrations are represented by a transition from the normal locus as a hypothetical complex to one which is in an “Amplified” or “Knockout” (KO) state. In addition, to convey that these are not natural transitions a phenotype node annotated with the corresponding genetic aberration, *i.e.* “Chromosome Duplication”, “Loss of heterozygosity”, “Gene Duplication”, “Gain of Function Mutation” and “Loss of Function Mutation” can be used as a catalyzer of the transition. As for mutually exclusive or co-occurring mutations, they can be represented as respectively negative or positive influences between the corresponding genes or loci.
As for the qualitative effects of the genetic alterations, an increase or decrease in the transcription rate of the gene has been defined as respectively a positive or negative influence of the mutated gene on its mRNA; increased or decreased protein efficiency or activity can be represented either by a direct positive or negative influence from the mutated protein on its wild-type counterpart in the pathway if an exact characterization of the mutation is not available, and if it is then the corresponding interactions must be represented.

Following definition of these different models for genetic alterations representation, they were presented, discussed and revised by active members of the Disease Map community from the LCSB BioCore.

4.1.3 Methodology of building the Glioblastoma Disease Map

In order to build the molecular interaction networks corresponding to each of these pathways, work was conducted in parallel on different files, or “submaps”, one for each pathway.

The submaps and final Disease Map were built using the CellDesigner software[102,103] version 4.4.2, and the final Glioblastoma Disease Map was then uploaded to the MINERVA platform[44].

To transcribe literature knowledge into a given submap, three different documents were involved to compile information about molecular interactions:

1. A “pre-curation” file, listing all statements encountered in articles read that may potentially be of use to describe and justify molecular interactions. This file is a spreadsheet containing two tabs. The first tab contains two columns: the statements of interest, and the publication that the review is using as a source for that statement. The

second tab contains, for each article read, a mapping between the citation system it uses (*e.g.* numbers: [1, 2, …]; author names and date: Aldape *et al.*, 2015…) and the PMID of the cited publication.

2. A "curation" file, which only includes the statements from the pre-curation file that describe interactions that are effectively modelled in the submap. This file is a table based on a template provided by the LCSB Biocore, and contains several columns:
    - "Evidence text" that contains the statement that describes a given interaction
    - "Title" of the publication the evidence text is extracted from
    - "Authors" of the publication the evidence text is extracted from
    - "Journal name / Publication date" of the publication the evidence text is extracted from
    - "PMID" the PubMed ID of the publication the evidence text is extracted from
    - "Cited resource" the PubMed ID(s) of the publications referenced as the source associated with the evidence text, if relevant
    - "Disease context" to indicate under which setting the described interaction takes place. As have been mentioned before, this could be in Glioblastoma, cancer in general, healthy tissue functions, but also a specific part of the pathway if it's the focus of the publication, such as regulation of a specific protein or influence of a protein on a downstream cellular function.
    - "ReactionID" that contains the ID of the reaction in the CellDesigner submap
    - "ID changes" that was used if an interaction was deleted or broken down into several, so that the original ReactionID is still recorded.
3. the CellDesigner file which contains the submap network for the pathway

The typical pipeline used to add new interactions to a given submap was to:

1. Read several publications concerning a specific topic. For each publication I would:
    a. read the publication a first time
    b. add the publication title and PMID as a new row in the precuration file, and highlight it to identify it as a title row
    c. read the publication a second time, and highlight all statements in it that may be relevant to describe molecular interactions; these statements should be relatively short, usually a summarization of findings which would otherwise take several sentences to characterize in detail
    d. add all these statements as new rows in the precuration file

    e. go through the references list of the publication to find potentially interesting papers that be relevant to read later, and add these references to a “Read later” folder in a bibliography references manager
2. Once the first step is completed, review the new statements in the precuration file across the publications added, and identify interactions to integrate to the submap. Preferably, for a given interaction there should be statements from at least two publications, which would strengthen confidence in their reliability.
3. For a given molecular interaction to add to the submap:
    a. the interaction was represented on the CellDesigner submap
    b. the PMID of the publication(s) the associated statements came from were added to the submap as a “isDescribedBy” relation of the “PubMed” data type
    c. the associated statements were added to the curation file, which was completed with all relevant information about the publications the statements came from
    d. the rows for associated statements in the precuration file were colored in blue to indicate the corresponding interaction had been modelled. If a statement described more than one interaction, the row was colored in green and the text for only the part about this specific interaction was colored in blue
    e. redundant statements from a same publication which were not used as reference statements in the curation file were greyed out to also indicate the corresponding interaction was already modelled, but not using these statements as reference description.

This process ensured that all potentially relevant information was identified from the publications read, stored in a centralized place to avoid the need to go back multiple times to the full text of a publication, and tracked to identify which pieces of that information had been used and which had yet to be represented.

The way molecular interactions were modelled depended on how detailed their descriptions was in the literature and on my interpretation of these descriptions. Generally speaking:

a. (de)phosphorylations and other types of post-translational modifications were typically represented by a transition of the (de)phosphorylated protein from one state to the other, catalyzed by the appropriate enzyme when relevant
b. however, if the description of the interaction clearly stated that the enzyme binds to the protein, or that they form a complex, or other similar formulations that explicitly mentions binding as a step of the process, the interaction was represented in two steps, with the

formation of the complex first, then its disassembly with the protein released in its (de)phosphorylated state

c. since (de)phosphorylation or other modifications may result in (in)activation of a protein, these cases were represented accordingly since the modification itself does not provide a clear indication that it corresponds to an (in)active state of the protein
d. broad cellular functions such as DNA damage response, apoptosis, cell growth, *etc.* were represented with the phenotype type of node in the network
e. there were occurrences where the description of a transition was unclear or where the process was suspected to be more complex and involving some additional steps, such as how DNA damage leads to activation of certain proteins for example, but for which a better characterization could not be found directly in the articles and would require finding articles specific to it. Since that interaction still needed to be represented with the information at hand, at least for the time being, the activating component of the transition was represented as “triggering” the transition as a way to indicate its role was more ambiguous than a clear catalyzer
f. when a protein was described as being sequestrated, or inhibited following binding to another, it was represented as a free, active protein forming a complex with the other protein, leading to its inactive state in the complex
g. otherwise, inhibition was represented as an action of the inhibiting agent on the interaction it inhibits, or as a negative influence on the protein inhibited if the specific interaction(s) inhibited were not identified
h. transcription factor-mediated transcription was modelled as a positive (or negative) influence of the transcription factor on its target’s mRNA
i. in cases where members of a protein family, or targets of a transcription factor, were involved in similar interactions with the same nodes, they were grouped together under a hypothetical complex, so that interactions involving them were represented by interactions only with the hypothetical complex, rather than separately with each of the components it contains. However, if one of the components of these hypothetical complexes was involved in other interactions only specific to that component, these interactions had to be represented with another instance of that component outside of the hypothetical complex, reformulated to "in accordance with the SBGN standard"

Once the submaps could be considered as finalized, *i.e.* the most important interactions for each submap were represented and investigating their downstream effectors further would take more time than available for the Disease Map, they were combined into one. This was done by

copy-pasting the networks from each submap into a new CellDesigner file. Then, the interactions that were found in more than one submap were identified and merged. Finally, the layout of the whole network was re-defined to make it more readable and easier to navigate.

4.2 Data management methods and systems

4.2.1 Cancer Trials Ireland

4.2.1.1 Glioma Biomarkers Study

For the Glioma Biomarkers Study, which was an observational trial seeking to identify blood biomarkers of glioma, I was tasked with inputting the trial data into a database. The data for each participant of this study was collected on individual Case Report Forms (CRFs) that had been previously scanned to be available numerically, and needed to be compiled and inputted in the CTI database for that study.

Each scanned CRF was screened to input the values and information it contained into the corresponding field of the CTI database. Once all the data from that form had been transcribed, the file of the CRF was placed into a different subfolder.

Issues in the data, such as unreadable data due to low quality of the scan, incorrect data type, out-of-range values, *etc.,* were referenced in a document dedicated to these issues, recording issues encountered as well as the corresponding CRF file, patient ID, and field. This file was regularly submitted to CTI supervisors, who then transferred the information to the staff members in contact with the hospitals participating to the study, so that they may request an update or clarification for that data.

4.2.1.2 Breast Cancer Trial

For the Breast Cancer Trial which was a two-arms, treatments comparison study, a fully populated database was available containing data about demographics (number of participants in different age ranges, geographical location…), biomarkers (*ERBB2, HER*, …), adverse events (frequency, severity, outcome of the different know adverse events…), response to treatment (quality of life, overall survival, progression free survival…) *etc.* collected for the trial.

The first stage for this work consisted mainly in getting familiar with the database and tools used for the exploration and analysis of the data. Several meetings with a supervisor as well as the database maintainer took place to discuss and ensure proper understanding of the database structure, its contents and how to query it. Extensive discussions also helped to grasp exactly what information was needed in the tables and reports output, how it was supposed to be formatted, and what information was expected to be conveyed through them. Meanwhile, SAS software[104] syntax was learned through online tutorials and looking at the CTI code for a few other similar trials.

Following the preparation phase, implementation of scripts to produce summary statistics tables and reports about the database contents for comparison between the two arms of the study began. This process started with the simpler tables which only required counting a few variables instead of intricate formatting of the output, to get comfortable with programming with SAS software before looking into more complex tables requiring to calculate outputs from several variables, compute summary statistics or output nested multi-level tables for instance. In doing so the example of existing projects was followed and one script per table or report was written.

Following observations of the repetitive usage of a few code snippets for multiple tables and reports, these snippets were turned into macros, the SAS software equivalent to programming functions, which were compiled into a separate dedicated script.
Among the more notable macros defined there were:

- one for each recurrent summary statistic (average, variance, …)
- one that would take data and a keyword for table structure as input to produce the table in the selected format
- one that output, for a given table, the information about missing data (patient, field…) that was encountered during the computation of that table. That output may then be used to request update or completion of the data to the source hospital

Scripts were then further compiled to gather the code for all tables and reports of a same category (such as demographics, biomarkers, or adverse events) into one script per category. These scripts ended up larger than any individual script for one table or report, but centralized the code into a more organized and easy to navigate structure than a multitude of small script files would have been.

4.2.2 TranSMART and the GLIOTRAIN Data

The data quality control and curation steps performed towards upload of data to the GLIOTRAIN tranSMART database were conducted in close collaboration with the GLIOTRAIN ESR who oversaw the sequencing of the GLIOTRAIN biobank samples, at VIB, to clarify issues and ambiguities that were encountered in the provided datasets. The code used to investigate the data and transform it to a tranSMART-compatible format can be found on the GitHub repository in the *GLIOTRAIN_Data_Analysis/GTdata_Curation.R* script.

4.2.2.1 Provided data

To ensure a relative homogeneity of the clinical profile of the patients they came from, the samples contributed by ICM, EMC and RCSI were selected based on the same inclusion criteria:

- a Primary Glioblastoma diagnosis, with wildtype Isocitrate Dehydrogenase (IDH) genes status
- patients were less than 70 years old at resection
- a Karnofsky performance status[105] (measuring ability to perform ordinary tasks) of over 70, which indicates a patient can care for themselves but is unable to carry on normal activity or work.
- patients were under only first line medical Stupp regimen: radiotherapy at 60Gy + Temodal[106] (Temozolomide-containing drug)
- frozen tissue from the tumours was available

In addition to the tumour tissue, EMC also provided 26 cell cultures for sequencing. These cell cultures were derived from 26 of the 56 tumour samples they provided, thus offering the possibility to compare expression profiles of tumours and derived cell cultures, but also to potentially use their cell cultures to validate findings from analysis of the tumour samples. ***Table* 1** summarizes the number of samples from each organization that were sent for each type of sequencing.

Table 1: Origin of samples from the GLIOTRAIN database. *ST / IT / LT: Short-Term / Intermediate-Term / Long-Term [survivors]*

Number of Samples	Sample type	Originating Institute	ST / IT / LT
15	Glioblastoma tissue (extracted nucleic acids)	RCSI (Brain Tumour Biobank, Beaumont Hospital, Dublin, Ireland)	- / 15 / -
57	Glioblastoma tissue (extracted nucleic acids)	ICM ("Onconeurotek" Pitié-Salpêtrière Hospital, Paris, France)	5 / 29 / 23
56	Glioblastoma tissue (FF)	EMC (Neuro-oncology Biobank, Erasmus Medical Centre, Rotterdam, The Netherlands)	9 / 37 / 10
26	Glioblastoma cell culture pellets	EMC (Neuro-oncology Biobank, Erasmus Medical Centre, Rotterdam, The Netherlands)	7 / 16 / 3
Total: 154			

***Table* 2** lists the collected clinical information associated to the GLIOTRAIN samples.

Table 2: Collected clinical data associated with samples in the GLIOTRAIN database. *Table extracted and adapted from the GLIOTRAIN project Data Management Plan.*

Data	Values
Patient ID	GLIOTRAIN-specific pseudonymized ID
Diagnosis Glioblastoma	Yes [Inclusion Criterion], No
Sex	Male, Female
Age at diagnosis	<70 [Inclusion Criterion]
Surgical Procedure	Biopsy, Partial Resection, Complete Resection
Location of the tumour (lobe)	Butterfly, Frontal, Parietal, Temporal, Occipital, Subcortical, and combinations of these
Location of the tumour (side)	Left, Right, Corpus Callosum
IDH Status	Wildtype [Inclusion Criterion]
MGMT promoter methylation status	Methylated, Unmethylated
Karnofsky performance status	>70 [Inclusion Criterion]
First line medical Stupp regimen (60Gy+Temodal)	Yes [Inclusion Criterion]
First tumor progression	Number of months since diagnosis
Second line medical treatment	Description of second line treatment
Second tumor progression	Number of months since diagnosis
Third line medical treatment	Description of third line treatment
Third tumor progression	Number of months since diagnosis
Fourth line medical treatment	Description of fourth line treatment
Fourth tumor progression	Number of months since diagnosis
Fifth line medical treatment	Description of fifth line treatment
Fifth tumor progression	Number of months since diagnosis
Sixth line medical treatment	Description of sixth line treatment
Death	Yes, No
FFPE available	Yes, No
Frozen tissue available	Yes [Inclusion Criterion]

The selected samples were then shipped to VIB (Leuven, Belgium), where they underwent sequencing by another GLIOTRAIN PhD student to produce RNA-Seq, WGS and Methylation data, following the pipelines described below.

The **RNA-Seq data** for the GLIOTRAIN database was produced using the following pipeline:

- Illumina HiSeq 4000, single end
- Removing duplicates with Clumpify version 37.28
- Removing adapters with Fastx clipper version 0.0.13
- Quality check with FastQC version 0.11.4
- Mapping the reads with STAR version 2.6

- Manipulating alignments (sorting, indexing) with Samtools
- Gene count with HTSeq version 0.10.0

In addition, seven of the samples were re-sequenced. Five of them because they had almost no reads in the output, and a standard output after re-sequencing lead to the conclusion that these five samples were not properly loaded during the initial sequencing run. For the other two samples, they had low read counts both before and after re-sequencing, which lead to the conclusion that the issue was likely due to poor quality of the library for these samples.

As a result of this process the dataset that was provided for processing and upload to the database contained the RNA-Seq read counts for 58,278 transcripts, associated to ENSEMBL IDs, in 154 samples.

Whole Genome Sequencing (WGS) data was produced for 151 samples from the GLIOTRAIN biobank. This was low-coverage sequencing data, therefore appropriate to derive Copy Number Variations, but not SNPs and variants information. Sequencing was performed through the following pipeline:

- Shot-gun whole genome libraries were prepared with KAPA library prep kit
- Illumina HiSeq 4000, 0.1X coverage
- Mapping the reads with BWA-mem version 0.7.12 to human reference genome hg19
- Picard to remove PCR duplicates v1.43
- Indexing and manipulating the reads with Samtools 0.1.18
- CNAs to identified by binning reads in 50kb windows with QDNASeq (R package)
- ASCAT algorithm v2.0.7 is used to segment the raw data

The resulting dataset contained the *log_R* value for a given chromosomal fragment or contig, which corresponds to the log2 transformation of the estimated copy number of that fragment. Each row in the dataset concerned one such fragment, indicating the chromosome and sample to which it belongs, the start and end positions of the fragment on the chromosome, and the *log_R* value. **Table 3** and **Figure 2Error! Reference source not found.** illustrate the format of that dataset, which contained data on a total of 11,829 individual chromosomal fragments.

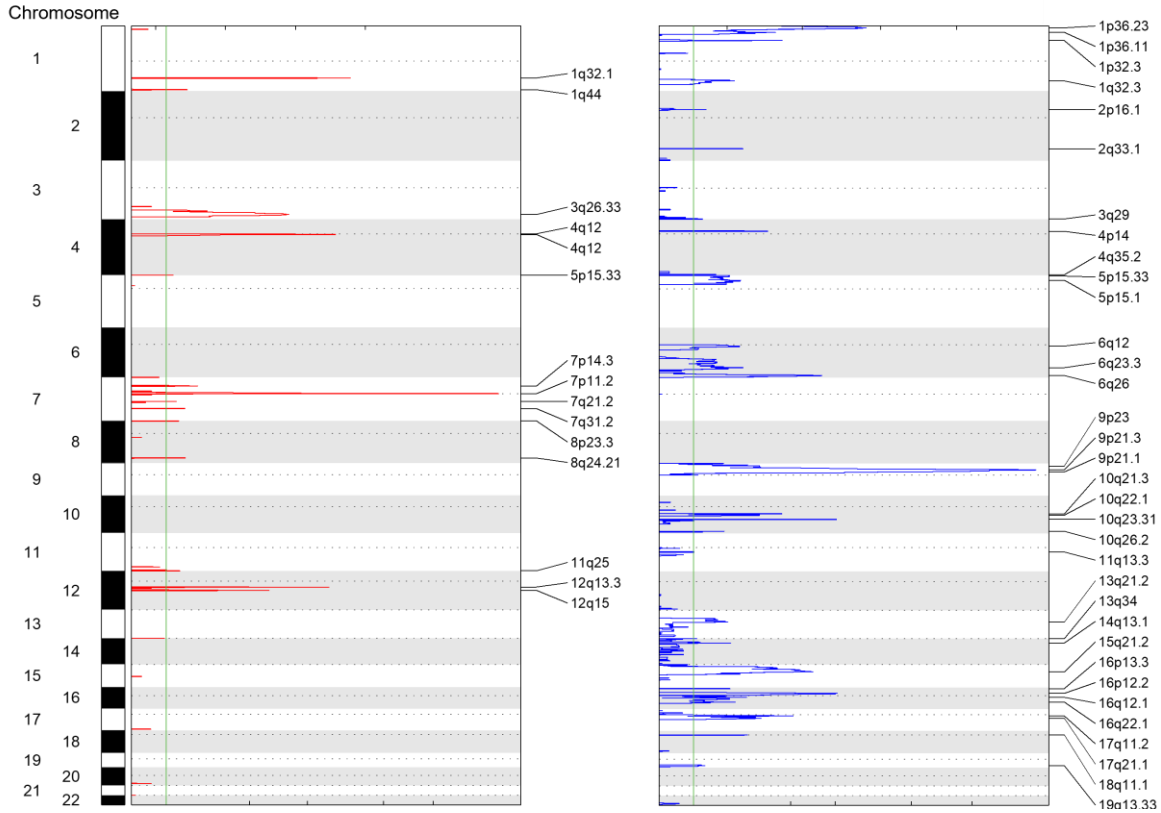
Sample	Chromosome	Start	End	*Log_R*
Sample 1	i	Position 1	Position 3	0.3208
Sample 1	i	Position 3	Position 5	0.0171
Sample 2	i	Position 1	Position 2	0.1625
Sample 2	i	Position 3	Position 4	1.2542
Sample 2	i	Position4	Position5	0.0593
Sample 3	i	Position 1	Position 5	0.0811

Table 3: Example of the format of the processed seg file data for an imaginary chromosome i. *Start and End give the pair-base positions of the limits of the considered fragment on the chromosome. Log_R is the measured value for that fragment*

Figure 2: Representation of the fragments present in *Error! Reference source not found.*, aligned per sample. Each horizontal black line represents a single continuous fragment of the chromosome 'i' for the corresponding sample. Dashed vertical green lines represent the start or end position of at least one of the

In addition, **focal events** were computed from the WGS data using the GISTIC2.0 software[107] at VIB, and were also shared as potentially interesting to the consortium. **Figure 3** represents the focal events identified in this dataset. It contained, for each focal event identified, the cytoband it corresponds to, its start and end positions, and for each sample whether the copy number of the region is strongly amplified, amplified, unchanged, or reduced, with values [2, 1, 0, -1] respectively.

Figure 3: Amplified (red) and deleted (blue) cytobands of identified focal events, as detected by the GISTIC software. *Y axis corresponds to chromosomal position of the focal event, X axis represents its amplitude. Credit for analysis and figures to Gonca Dilcan (GLIOTRAIN ESR at VIB, Belgium)*



Methylation data was generated only for 15 samples, due to higher cost and limited tissue available for RNA-Seq, WGS and Methylation sequencing. The samples were selected based on two criteria:

- There needed to be an even distribution of short-term (OS < 9), intermediate-term (9 < OS < 36) and long-term (OS > 36) survivors-derived samples in the dataset, so five of each.
- There was enough of the sample left after both WGS and RNA-Seq analyses were performed.

Note that as a result of this selection, the distribution of methylation profiles is likely not representative of what can be observed in a random selection of samples. The selection of the samples was performed by three GLIOTRAIN ESRs together at VIB.

Contrary to RNA-Seq and WGS, the Methylation data was sequenced by a member of the VIB team other than the GLIOTRAIN PhD student, who was not able to provide a detailed description of the sequencing pipeline, besides that it also relied on an Illumina HiSeq4000 sequencer.

The dataset resulting from Methylation sequencing contained the beta (β) values for individual CpG sites, that is to say the estimation of methylation level of the sites for the patient, between 0 for unmethylated and 1 for methylated. Furthermore, data was not available for all site in all 15 samples, resulting in a large yet extremely sparse matrix. As a consequence, to limit the size of the data to be loaded on the database while avoiding removing potentially relevant information, the dataset was filtered out to keep only CpG sites for which data was not missing in at least 11 (75%) of the samples. The final matrix contained data for 250,959 CpG sites.

Following sequencing, the data was sent to ITTM for processing and upload to a tranSMART database for sharing and access across the consortium. The sequencing datasets all used sample IDs internal to VIB rather than the GLIOTRAIN IDs.

4.2.2.2 Data Privacy and Security

In order to respect data privacy and security, several measures were taken and aligned across the whole consortium.

First, pseudonymization considerations were included during the process of determining the clinical information that would be collected and shared with the consortium about each sample. As a result, an ID system internal to GLIOTRAIN and based on the example of the The Cancer Genome Atlas (TCGA) barcodes[108] was defined. This labelling system contained information to identify each sample based on the institute that provided it, a sample number, the tissue origin of the sample (tumour or cell culture), the type of analyte contained in the sample (DNA, RNA or Protein extracts, or whole tissue), and the sequencing that was performed to generate the data (WGS, RNA-Seq or Methylation). These barcodes were agreed on by the consortium and used directly to label samples at the collection phase, so that the mapping between GLIOTRAIN IDs and the source institute IDs would be known only to the source institute.
Furthermore, any type of information that could easily allow to identify the patient also had to be excluded or modified to remove that possibility, such as recording the age at diagnosis of the patient rather than the exact date of diagnosis.

Secondly, secure channels to access and share data were needed. For that, several solutions had to be implemented:

- a private ownCloud file sharing instance was set up and accessible via ITTM-provided credentials, the password of which had to be changed upon first connection. The ownCloud was used to share and work on documents within the consortium. It was also

used to transfer small data files between users, with the agreement that the data had to be immediately collected and removed from ownCloud folders by the people involved

- for larger omics files that had to be sent from VIB to ITTM for curation and upload to the tranSMART database, a SSH connection was provided by VIB for the transfer
- the tranSMART database itself was hosted on secure servers accessible only with credentials, provided by ITTM upon request and approval by the GLIOTRAIN directing body, and included a temporary password that was required to be changed at first login.

The usage of these channels, as well as the fact that other ways to share sensitive information such as emails were inappropriate as potentially easy to breach, was extensively and repeatedly explained to the consortium members through mails, presentations, and in the internal GLIOTRAIN Data Management Plan.

These considerations and subsequent implemented solutions ensured that the GLIOTRAIN data was handled in a secure way, respecting both data privacy of the patients and GDPR regulations.

4.2.2.3 TranSMART Data Tree

The organization of the data in the graphical interface of tranSMART took into consideration that the data should be organized logically and intuitively, by grouping together data that would likely be needed at the same time. As such, the data tree nodes devolved from broad categories such as “Clinical Data” or “Biomarker Data” into more and more specific labels like “First Line of Treatment” or “RNA-Seq Data”.

While they were initially sequenced together, the choice was made to separate the cell cultures omics datasets (RNA-Seq, WGS and focal events) from the parental tumour data in the final database, since cell cultures were an addition to the main body of patient data, and may present a different profile than a normal tumour, even if slightly. As such, consortium member may want to exclude cultures data from their analyses, which would have been difficult to do if they were part of the same dataset in tranSMART, while including them if they are loaded as a different dataset was much simpler. In addition, discrepancies between the molecular profiles of a few of the cell cultures and their parental tumour counterparts (see subsection 5.2.2.1 Provided Data Characterization) comforted this decision.

4.2.2.4 Characterization of the Sequencing Data

In order to map ENSEMBL IDs of transcripts from **RNA-Seq** sequencing to the corresponding gene names, two issues were solved through inquiry to VIB:

- the version of the ENSEMBL database used in their sequencing pipeline to annotate transcripts turned out to be the release 89 (April 2017) of the ENSEMBL database
- a "*_PAR_Y*" suffix present for 45 transcripts after an ENSEMBL ID which also existed in the data without the suffix and in the ENSEMBL database, was identified as an annotation to differentiate transcripts derived from the Y chromosome allele of a gene from the transcripts coming from the gene on the X chromosome, which were the ones without the suffix

The initial WGS dataset provided by VIB was a matrix where each row contained the log2-transformed copy number estimation of a given chromosomal fragment from a sample. It contained data for 151 samples out of the 154 total included in the GLIOTRAIN biobank, the three missing being because they didn't have enough biological material for both RNA-Seq and WGS sequencing.

First the coverage of the data was investigated. For that purpose, a graphical view of the data for each chromosome was produced, by lining up horizontally all the fragments of a given sample for that chromosome, and presenting this representation of the chromosome for all samples on top of each other, much like what is represented in the example of **Figure 2**. In addition, each fragment was colored based on its associated copy number value.

Furthermore, a comparison of the copy number variation profiles between the cell cultures samples and the parental tumour samples they were derived from was performed at VIB.

For the focal events dataset, it was noted that the GISTIC analysis to identify focal events did not cover X/Y chromosomes.

Next, the RNA-Seq and WGS data was analyzed to evaluate its quality and detect potential biases and artifacts from source institute, sequencing batch, or source tissue type. This was achieved by performing a Principal Component Analysis (PCA) on the data, using the *ade4*[109] R package, version 1.7-16.

The PCA was performed on the RNA-Seq data including all samples, where any 0 value was incremented to 1 and all values were then log-transformed.

To analyze the WGS data, it was transformed to a more adapted format. For each chromosome, the limits of all fragments from all samples for that chromosome were considered to be shared across all samples, and subsequently the regions between two consecutive limit positions were defined as new fragments. Thus, many fragments from the initial data were broken down into several smaller ones sharing the same value as the original. The columns in this new dataset, hereafter called "WGS revised dataset", were then named as a combination

of the chromosome, start and end positions of the fragments they correspond to. Using the example defined by **Figure 2** and **Table 3**, **Table 4** shows what the transformed data looks like.

Sample	chri_Position1_ Position2	chri_Position2_ Position3	chri_Position3_ Position4	chri_Position4_ Position5
Sample 1	0.3208	0.3208	0.0171	0.0171
Sample 2	0.1625	NA	1.2542	0.0593
Sample 3	0.0811	0.0811	0.0811	0.0811

Table 4: Format of the WGS revised dataset**, based on the example data from **Table 3

4.2.2.5 ETL Processing of the Data

In order to load the GLIOTRAIN data to the tranSMART database, a pipeline using software internal to ITTM as well as data loading tools[110] recognized by the tranSMART foundation was used. This pipeline required several files as input.

For the clinical and low-dimensional data:

- A file containing the tranSMART data tree structure, the name of each node, as well as the path to the file containing the corresponding clinical and low-dimensional data
- One file or more containing the clinical and low-dimensional data to be loaded to the database

For each omic dataset loaded to the database:

- The data itself, formatted so that samples are columns and features (genes, chromosomal regions, transcripts, *etc.*) as rows, and the symbol '.' (a dot) as missing values.
- A file mapping the samples from the dataset to the corresponding patient in the clinical data.
- A file containing the list of all features present in the dataset and the gene they correspond to. But this association to a gene is not a requirement and the corresponding column can be left empty, when the features are large chromosomal regions for instance.

Before anything, the VIB sample IDs were replaced in the data by the GLIOTRAIN IDs.

Then, the files to load the clinical data for the patients, as well as the sequencing batches information for the RNA-Seq and WGS datasets as low-dimensional data were prepared.

A mapping file indicating the gene name corresponding to each ENSEMBL ID was produced.

The WGS revised dataset format was found to be the best solution as a tranSMART-compatible format for that data.

In the Methylation dataset the missing values “NA” were replaced with ‘.’, for tranSMART to recognize them as such.

Furthermore, as previously mentioned in the final database the parental tumour and cell lines samples should be available in distinct datasets. As a consequence, after transformation to fit the tranSMART requirements, the RNA-Seq dataset, WGS revised dataset, and focal events dataset were divided accordingly.

Once the files associated to each of these six datasets and to the methylation data, as well as the database structure, clinical data and batch information files were completed, the ETL program was executed and the data loaded to a tranSMART v16.3 database.

Following uploading, the tranSMART user interface was used to test and explore the data and make sure everything was in order, accessible and worked as expected, including download of the data. Only then, finally, emails were sent to the consortium members to notify them of the availability of the database, and provide them with credentials to access it.
The database was accessed and used by several members of the consortium over the course of the project to explore and download data.

4.2.2.6 Documentation

All of the findings produced through the characterization of the GLIOTRAIN data, as well as transformation performed to reach the format of the data as available in the final GLIOTRAIN database were communicated repeatedly to the consortium, through emails and presentations.

Furthermore, the writing of an internal “data booklet” was started to compile all that information and the contents of the database.

4.2.3 The OMOP CDM

4.2.3.1 Writing SOPs

The definition of standard processes for mapping projects to migrate a database to an OMOP-compatible format required in-depth evaluations of the steps needed in such projects, how to perform them in a standardized, reproducible and efficient way, and how to provide a clear and thorough documentation to describe them.

First, the work needed in a mapping project was broken down into broad processes. These processes were defined partly from experience of mapping projects, and by reflecting on a way to logically divide the work into coherent and non-overlapping categories. Once these categories were roughly characterized, their precise definition was established by identifying milestones and major objectives, defined as documents or work required to be completed before another process may be started.

Following identification of broad processes, the tasks involved in the completion of each process were listed. In this endeavor, solutions to optimize completion of work were sought out, especially for bottlenecks and particularly time- and resource-consuming tasks. For instance, templates were created for reports, project documentation and code, wherever possible.

Finally, this exercise was completed by writing extensive documentation of these processes as internal SOPs, describing in detail each task, and providing guidelines on how to complete it. For this, a uniform and coherent terminology also had to be defined and used throughout the SOPs.
Once written each SOP was submitted to the ITTM team involved in OMOP mapping projects for discussion and validation, and was adjusted based on feedback until a consensus was reached.

4.2.3.2 ETL software

The first version of the ITTM ETL program was investigated, reorganized and further developed.

A first step of this work consisted in the review of the code to understand its structure, how it operates, and identify and document missing features as well as potential bottlenecks of execution which would be resource consuming.

Secondly, the code was refactored to increase modularity: recurrent code snippets were turned into functions called where relevant, large functions were broken down into smaller functions calls, and code that was scattered across the program but related to the same feature was extracted to dedicated modules and classes of the program.

Finally, additional features for the software were developed to align it with the newly standardized OMOP mapping project pipeline documented in the SOPs. Among these new features is the intake and interpretation of machine-readable mapping syntax described below.

Each new code refactoring or development was tested to ensure proper implementation and tracked using the ITTM internal Gitlab versioning platform.

4.2.3.3 Creating the machine-readable syntax

An important bottleneck in completing an OMOP mapping project lies in the programmatical implementation of defined mappings to execute them. To mitigate the impact of this issue a machine-readable syntax for mapping was designed. The objective in the creation of this syntax was to define a way to describe mappings that would be structured enough that it may be read and interpreted by a program, while also simple and intuitive so that people unfamiliar with programming may use it without extensive training.

To design the syntax, typical mapping definitions were reviewed to identify recurrent patterns and structures that required to be formalized in a machine-readable way. Simultaneously, in order to keep it intuitive to non-programmers, formalization of the syntax was limited as much as possible to include only the structures indispensable to handle mapping definitions. Furthermore, to limit the number of instructions necessary to define any given mapping, operations that were frequently used in mappings were identified and associated to explicit keywords, which would then be used in the interpretation of the syntax to refer to these operations.

However, it was also recognized that some of the more complex mapping operations would not be covered by the syntax confined to the strict minimum of formalization. Nevertheless, the gain from keeping it that way was judged optimal. Indeed, regardless of mapping definition with the syntax, a programmer would still need to be involved for customization of the ETL launcher and its execution. The machine-readable syntax would allow mappers to define most mappings themselves, and the more complex operations would need to be entrusted to the programmer for implementation. As a consequence, this would limit the additional workload of the programmer to only a few operations which can be extensively described by the mapper, instead of having to read through and implement every mapping definition individually.

The resulting syntax, *i.e.* the set of guidelines to define mappings in a structured yet intuitive fashion, was submitted to the ITTM team involved in OMOP mapping projects, and particularly people without programming skills, to improve and adjust it.

Once validated, the code to read and interpret it was implemented and integrated to the ETL software.

Finally, the syntax and how to use it in mapping projects was extensively described in documentation integrated to the ITTM OMOP SOPs.

4.3 Data Analysis

4.3.1 Enrichment Analysis Methods

In order to characterize the results from the different analyses performed during the PhD studies and to understand their biological relevance in the context of Glioblastoma, it was necessary to determine whether the genes identified as a result of a given analysis presented synergies in terms of the cellular processes and functional pathways to which they belonged. For this purpose, mainly two methodologies have been used.

4.3.1.1 IPA functional pathways investigation

The QIAGEN Ingenuity Pathway Analysis[111] (IPA, QIAGEN Inc., https://digitalinsights.qiagen.com/IPA) software version 01-16 was used in the earlier analyses of the Identification of predictive biomarkers of drug response work, to investigate the genes found to be associated with a given drug by generating and exploring functional pathways networks. Using it was encouraged by the EMC team since it was also the tool used for the Berkeley LASSO analysis to identify genes of interest out of the LASSO results, and as such would be part of the reproducibility efforts.

As such, the functional pathways overly represented among these genes were identified using the following steps:

- a new empty "pathway" was created, where the genes under investigation as well as the drug to which genes were associated and its known direct target(s) were added
- the investigated genes were placed together and away from the drug and its targets
- the *Connect* function was used to determine if there was any known direct connection(s) between any of these molecules
- the *Grow* and *Path Explorer* functions, adding 50 molecules at a time, were used up to two times to expand the set of nodes that may link them with each other
- the *Canonical Pathways* function overlay was used to determine if any biological function was strongly represented in the resulting sets of genes
- the network was manually modified to present a more structured and readable layout, with the investigated genes on one side, the drug and its target on the other, and the nodes that were added as a result of network expansion functions in-between.

However, that approach was manual, very time consuming, there was no way to control or define if and why a gene should be added over another when expanding the network, and the molecular interactions and associated canonical pathways are proprietary and manually

curated, which may lead to an incomplete or biased network during the process. For instance, canonical pathways appeared biased as mostly cancer-related pathways would show up, even with only a handful of genes in the network while non-cancer pathways were scarcely represented.

4.3.1.2 Gene Ontology enrichment function

As a result, a computational alternative was implemented. An R function was coded to:

- take a list of genes of interest as input
- perform an enrichment test against Gene Ontology Biological Processes terms thanks to the topGO[112] R package, version 2.42.0

output the top 20 pathways found enriched in the genes of interest along with the p-value of the enrichment test. Note that these were not adjusted for multiple testing and used mainly for ranking purposes. Adjustment was not attempted as it would have required a complex approach due to the multi-step nature of the enrichment tests for limited additional insights, as described in the R package documentation. This function named *runTogGO()* can be found on the GitHub repository in both *Identification_of_predictive_biomarkers_of_drug_response/ EMCAnalysis_IoannisDrugs_runningAnalyses.R* and *GLIOTRAIN_Data_Analysis/ GTAnalysis_DEA.R* scripts.

4.3.2 Identification of predictive biomarkers of drug response

4.3.2.1 Materials

4.3.2.1.1 Drugs Repurposing Project

For the Drugs Repurposing Project, the available data was:

- The whole-transcriptome data from a DASL (cDNA-mediated Annealing, Selection, extension, and Ligation) assay, collected from first resection Glioblastoma tumors, *i.e.* before any drug therapy against the cancer. It came in the form of four files, one per sequencing batch, and each of them presenting expression data of 29,377 probesets for a varying number of samples, with a total of 88 samples across the four batches. In addition, this data was not collected specifically for this project: it was from a previous, completely independent study[113] which used it for a different purpose. It is worth noting that since that study took place several years before, we had no control or way to verify the methods and quality of the process of generating that data.
- The clinical data about the patients for which DASL data is available, about their tumors and the cell cultures derived from it. That data included:

    - Overall Survival of the patients in months after surgery.
    - Pathological diagnosis of the tumor, of which there were 35 Primary Glioblastoma, 10 Recurrent Glioblastoma, 12 oligodendrogliomas, 9 astrocytomas, and the rest were other types of brain malignancies.
    - The WHO grade, of which there were 64 grade IV, 19 grade III, and the rest was lower or unspecified.
    - The MGMT promoter methylation status of the cell cultures.
- The IC50 values for all 109 drugs screened on the 45 Glioblastoma cell cultures. These values were extrapolated from the percentage of cell population survival at different concentrations of the drugs for all 45 cultures. This dataset, hereafter called "initial IC50s dataset" contained a mix of quantitative values (*e.g.* 9.63, 0.0081, 175.10), categorical values (*e.g.* "< 0.0256", "> 160") and missing values marked with an 'X'. As such, there were clear inconsistencies within the dataset, where some IC50s were marked with a categorical estimation (">160") right next to numerical values beyond that same category (*e.g.* 175.10). This was the first drug response dataset provided.
- The "original IC50s dataset", from which the initial IC50s dataset was derived by replacing extreme values such as 10^{144} or 10^{-18} by the aforementioned qualitative values of ">160" and "<0.0008", respectively. Such extreme values were the result of extrapolation of IC50 even for drugs for which the tested range of concentrations did not contain the IC50 and was thus inappropriate. This dataset was provided following request for the fully numerical dataset instead of the mix of numerical and categorical values.
- The raw data for cell cultures survival after exposure to different concentrations of each screened drug, provided along the original IC50s dataset. This data contained, for each cell culture, spectrophotometry measurements of solutions of cell cultures following exposure to a given concentration of a drug. For each cell culture and each drug, measurements were collected after exposure to six different concentrations of the drug, twice. Similarly, measurements were also taken after exposure to six different concentrations of DMSO, which was the solvent for cell cultures, to use as control. Note that the concentrations the cell cultures were exposed to were not the same for all drugs, since drugs may be more or less potent. ***Table* 5** summarizes the six concentrations tested for each drug, but it should be mentioned that in a few isolated cases, there were cell cultures which were tested with a different set of concentrations than the other cell cultures for a given drug. However, these isolated cases will not be described in detail

here, since the difference in concentrations tested was limited on the log scale where the data was used for analyses, and that difference was handled in the downstream analysis so that the cell cultures may still be compared to each other.

Table 5: Concentrations used to test survival of most cell cultures against each drug.

Concentrations (µM)	Drugs tested with these concentrations
160, 16, 1.6, 0.16, 0.016, 0.0016	Allopurinol, Altretamine, Aminolevulinic acid hydrochloride, Anastrozole, Azacitidine, Bendamustine hydrochloride, Busulfan, Capecitabine, Carmustine, Celecoxib, Chlorambucil, Cisplatin, Cyclophosphamide, Dabrafenib mesylate, Dacarbazine, Decitabine, Dexrazoxane, DMSO, Enzalutamide, Estramustine phosphate sodium, Exemestane, Floxuridine, Fludarabine phosphate, Fluorouracil, Fulvestrant, Hydroxyurea, Ifosfamide, Lenalidomide, Letrozole, Lomustine, Mechlorethamine hydrochloride, Megestrol acetate, Mercaptopurine, Methotrexate, Methoxsalen, Mitotane, Nelarabine, Pazopanib hydrochloride, Pemetrexed, Pentostatin, Pipobroman, Pomalidomide, Procarbazine hydrochloride, Streptozocin, Sunitinib, Temozolomide, Thalidomide, Thioguanine, Thiotepa, Tretinoin, Uracil mustard, Vismodegib
80, 8, 0.8, 0.08, 0.008, 0.0008	Afatinib, Amiodarone hydrochloride, Arsenic trioxide, Axitinib, Bleomycin sulfate, Bosutinib, Cabozantinib, Carboplatin, Cladribine, Clofarabine, Crizotinib, Cytarabine hydrochloride, Dasatinib, Daunorubicin hydrochloride, Doxorubicin hydrochloride, Epirubicin hydrochloride, Erlotinib hydrochloride, Etoposide, Everolimus, Gefitinib, Gemcitabine hydrochloride, Idarubicin hydrochloride, Imatinib, Irinotecan hydrochloride, Lapatinib, Melphalan hydrochloride, Mitomycin, Mitoxantrone, Nilotinib, Oxaliplatin, Plicamycin, Ponatinib, Pralatrexate, Raloxifene, Regorafenib, Sirolimus, Sorafenib, Tamoxifen citrate, Temsirolimus, Teniposide, Topotecan hydrochloride, Trametinib, Valrubicin, Vandetanib, Vemurafenib, Vorinostat
8, 0.8, 0.08, 0.008, 0.0008, 0.00008	Bortezomib, Cabazitaxel, Carfilzomib, Dactinomycin, Docetaxel, Ixabepilone, Omacetaxine mepesuccinate, Paclitaxel, Romidepsin, Vinblastine sulfate, Vincristine sulfate, Vinorelbine tartrate

4.3.2.1.2 Berkeley LASSO Analysis

In addition to the aforementioned datasets, the analysis performed at Berkeley University yielded:

- The Berkeley LASSO analysis results, that is to say the genes for which the expression profile appeared to be significantly connected to a drug response. Such genes were identified for 36 drugs out of the 109 screened, the other components did not produce significant results. These results can be found in **Table 21** from annex 8.1 Berkeley LASSO Analysis Results.
- From these results, 10 drugs out of the initial 109 were shortlisted at Berkeley University for further investigation, using the IPA software to assess functional relevance of the

associated genes. **Table 6** presents the subset of the 10 shortlisted drugs from **Table 21**, and highlights genes that were found to be of particular interest following the IPA investigation.

- Information about the IPA-identified genes of interest regarding their full name, protein class, and whether they are known to be prognostic biomarkers in any cancer type.

Table 6: Drugs and associated LASSO-selected genes shortlisted for IPA investigation at Berkeley University. *Bold, underlined gene names are the genes of interest identified in IPA.*

Drug	Mechanism of Action	Targets	Gene Names
Vemurafenib	B-RAF inhibitor	MAP4K5, SRMS, BRAF, ARAF, RAF1, TNK2, FGR	**CTSG**, **DSP.2**, HBD, SLC22A2.2, SLC6A20, TNNT2
Tretinoin	Retinol analogue	RARA, RARB, RARG, RXRA, RXRB, RXRG	**DAO**, DNAI2, **KIF19**, **PTPN3**, **RNF7.1**, SEMA3E, SLC39A12, **TRPM3.1**
Dexrazoxane	Topoisomerase II inhibitor	TOP2A, TOP2B	**ASB12**, CC2D1B.1, **ICAM5**, LPHN1, MRM1, OLA1.1, RTBDN.2, SEMA6D.4, SPPL2B
Cytarabine hydrochloride	Antimetabolite	POLA1, POLB, POLD1	**PTPN20B**
Mitotane			**CALCA.2**
Bortezomib	Proteasome inhibitor	26-Proteasome, PSMB1, PSMB2, PSMB5, PSMD1, PSMD2	**CTAG2**
Imatinib	Bcr-Abl	KIT, PDGFRA, PDGFRB, CSF1R, DDR2, DDR1, ABL1, RET	MIRLET7D, OR2L13, PRPF40B.1, Sep.04, SLC8A3.2, **SLITRK1.1**, **XRN2**
Hydroxyurea		WNT3, RRM1, RRM2, RRM2B, Ribonucleotide reductase	**ATG12**, LOC286238, **WNT3**, ZPBP2.1
Sorafenib	Multi-TK inhibitor		**PDE1A**, SLC01A2.2
Pazopanib hydrochloride	multi-targeted kinase inhibitor	FLT4, KIT, PDGFRA, PDGFRB, FGFR1, FGFR2, CSF1R, FGFR3, FLT1, KDR, RET, LCK, ITK	**ALOX12**, C15orf27, CASQ2, CBLN4, CPLX3, CT45A5, CTXN3.1, FAM81A, FAM9A, GABRB2, **XRN2**, ZNF676, GPR128, LCN6, **MAPK1**, MYH11, NECAB1, OPRK1, **OPRM1.5**, PLN.1, TMEM144, **TPSD1**, TROVE2.3, UCN3

Validation of Glioblastoma cell culture models 51 cell cultures were used in this study, including the GLIOTRAIN cell lines except for the questionable ones, as well 13 cell cultures present in the Drugs Repurposing Project. The datasets used in this study and corresponding to these cell lines include:

- New datapoints of cell cultures survival rates after exposure to TMZ, screened over the concentrations [6.25, 12.25, 25, 50, 100, 200, 400] µM, since it came out during the drugs repurposing project that the initial concentrations tested were too low to capture response to TMZ properly.
- The DASL data and RNA-Seq data subsets for the cell lines that had this transcriptomic data available

- The OS and PFS data for the patients from which the cell cultures were derived.
- The MGMT promoter methylation status of both the parental tumours and of the cell cultures.

4.3.2.2 Data Preparation

Before performing any analysis with the provided data, it had to be processed into a format fit for analysis. The corresponding code can be found on the GitHub repository in the *Identification_of_predictive_biomarkers_of_drug_response/ EMCAnalysis_IoannisDrugs_prepocessingAndExtraction.R* script.

4.3.2.2.1 DASL microarray

The first steps of transformation and normalization of the DASL data were taken following directions from the Berkeley University team that had previously performed a LASSO analysis with the data. Using the Lumi[114],[115],[116] R package version 2.42.0, the four raw intensity files were imported and that data was converted into expression values, using the *lumiExpresso* method applying quantile normalization of the data at the same time. The four datasets were then joined and corrected for batch effect using the *ComBat*[117] function from the sva[118] R package version 3.38.0. Next, probesets for which 75% or more of the samples presented a detection p-value above a threshold of 0.05 were removed. Finally, the probesets IDs were replaced with the gene name they corresponded to. Since there were genes that were represented by several probesets, the gene name was followed by a number starting at 1 and increasing with every new probeset referring to the same gene. For instance, if three probesets corresponded to the *TP53* gene, they would be named *TP53.1*, *TP53.2* and *TP53.3*. This processed dataset, with all 88 samples and all probesets except for those of poor data quality, will hereafter be referred to as the "DASL normalized dataset".
The probesets in this dataset were then filtered for the needs of the various analyses performed during the PhD project, both for Identification of predictive biomarkers of drug response analyses and the GLIOTRAIN data analyses.

For the first of these filters, in cases where multiple probesets corresponded to the same gene, only the one with the highest variance was kept and the others were removed. Subsequently, the ".n" suffix to the gene name was dropped since there was then only one probeset per gene. After that, the mean and variances of all remaining probesets were computed, and the probesets that were in the lower quartile of both distributions were removed, in order to limit noise in the dataset. The list of probesets, and by extension genes, that resulted from this filtering process will hereafter be called the "unbiased genes set".

Another filter applied to the DASL normalized dataset was defined with the intent to investigate the data specifically under the oncogenic angle, the idea being to only keep probesets corresponding to genes known to be associated with cancer. For this purpose, a list of such genes was established by looking into arrays specialized for cancer:

- the nanoString nCounter Pan-Cancer Pathways Panel[119] and the nanoString nCounter Pan-Cancer Progression Panel[120] which both list about 770 genes involved in pathways related to angiogenesis, PI3K, EMT...
- the lists from Illumina TruSight Oncology 500[121] and Illumina Ampliseq Cancer Panel[122] of 523 and 409 genes, respectively
- the lists from Qiagen Cancer PathwayFinder RT^2 Profiler PCR Array[123] and Comprehensive Cancer GeneRead DNAseq Gene Panels[124] of 84 and 124 genes, respectively
- the list from Agilent ClearSeq Cancer Research Panels[125] of 151 genes

Then, going back to the DASL normalized dataset, probesets associated with genes that were present in any of these lists were kept and the others were removed from the dataset.
The probesets from the DASL normalized dataset corresponding to any of the genes from this list were kept, and the other probesets were removed. Since this resulted in a much smaller number of probesets than the unbiased genes set, probesets were not further filtered out. This list of probesets will be referred to as the “cancer genes set” from now on.

The last filter was defined and used specifically in the validation of Glioblastoma cell culture models analysis, and included only protein-coding genes. In this case, annotations from the ENSEMBL database were collected to identify protein-coding genes. The probesets which did not refer to any of these genes were filtered out from the DASL normalized dataset. Among the remaining probesets, the data from probesets which referred to a same gene was averaged per sample. Note that this filter was also applied to the GLIOTRAIN RNA-Seq data used for the cell culture models validation study. These genes will be referred to along with the corresponding dataset as the “protein-coding genes set”.

4.3.2.2.2 Drug response data

The initial IC50s dataset provided was not appropriate to use in quantitative analyses. Discussions about this issue led to the original IC50s dataset and spectrophotometry measurements dataset to be sought out and found. Instead of categorical values, boundaries were defined and applied on the entire original IC50s dataset to avoid inconsistencies in the data and mitigate the skewing introduced by these extreme values in subsequent analyses:

- since the smallest dose that was tested in the drug exposure experiments was 0.00008uM, the lower bound was set to that value divided by 3, and any IC50 lower than that was replaced by this lower bound value;
- for very high values, the C_max (maximum nontoxic dose for a drug in humans) value for Hydroxyurea which was the highest (795 microMolar) out of all the screened drugs was multiplied by 3 to define the higher bound for the dataset. Any IC50 higher than that was set to this value.

Although these are also arbitrary values, at least they provided a more consistent approach to the problem and would allow to run quantitative analyses, although the impact of this shortcoming should also be considered in any result making use of this dataset, which will be called the "bounded IC50s dataset".

Later an approach taken by Tiriac *et al.* (2018)[126], inspired the EMC team to suggest using the spectrophotometry measurements to calculate the Area Under the Curve (AUC) of cell survival rates at different concentrations of a drug, and use that AUC as drug response variable instead of the IC50. Considering that the IC50s values were already approximated and rather unreliable, the suggestion was adopted, and a method in R was implemented to calculate the AUC given the datapoints.

The expected input of this code is a dataframe containing the viability rate of the cells at the different tested dosages of a given drug, where the column names should be the concentration value (just the number, without unit), and the row names should be the cell culture denomination. An argument of the function allows to indicate whether the data contains duplicates for cell cultures, in which case the duplicates should be identified with the suffix "*_dupN"* (*_dup1, _dup2, _dup3*…).

The first step in the function is to apply a curve-fitting algorithm to fit a specific curve model to the data using the *drc*[127] R package version 3.0-1. For each cell culture, three types of models were tested for fitting to the data: a log-logistic model, linear model, and exponential decay model. These three potential models were chosen based on observations of plotted cells survival data on the log scale, where shapes that could match these models could be seen. **Figure 4** provides an example of such plots, where the log-logistic shape is visible in the data from Thioguanine exposure experiments, and curves from the Gemcitabine hydrochloride data evolve in a way closer to either exponential decay or linear decrease.

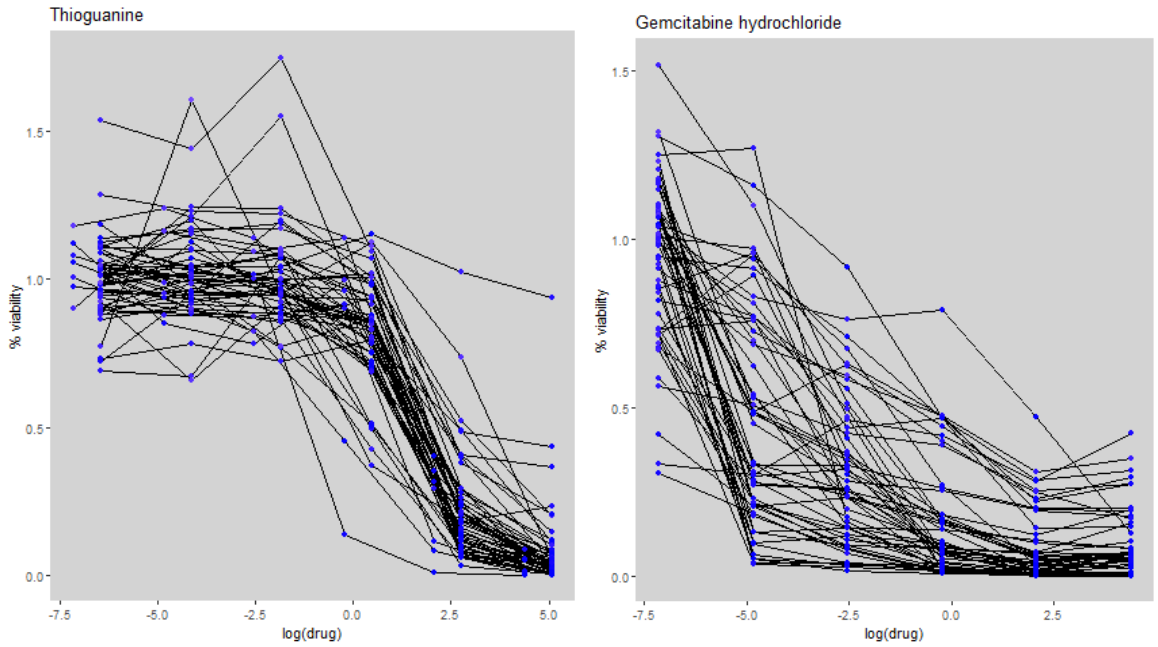


Figure 4: Example of plots of cell survival rates (Y axis) at different concentrations (X axis) of Thioguanine (left) and Gemcitabine hydrochloride (right). *Blue dots correspond to the percentage of cell population survival at the given drug concentration, averaged across all replicates for a given cell culture. Black lines represent evolution of the cell culture response at different concentrations by linking the blue dots of a given cell culture.*

The model with the lowest residual standard error out of the three is selected for that cell culture. If the exponential model is selected but the estimated value of the model is above 3 at the lowest concentration, meaning that using this model the cell population would supposedly be more than three times larger than when not exposed to the drug which is an unreasonable assumption, then the log-logistic model is preferred if available. Cell cultures for which no model can be fitted for a given drug cannot have the corresponding AUC calculated and as a result are excluded from further analysis involving the corresponding drug. In other cases, the fitted model is increasing instead of decreasing. For those, a plot is produced to let the user decide whether to keep them (*e.g.* linear models with little variation, suggesting resistance) or exclude them (*e.g.* completely abnormal behaviour, for instance due to bad data quality/measurement). If there are duplicates for a given cell culture, the data from all duplicates are fed to the algorithm, and they will be used together to fit the model for the cell culture. Once the best-fitting model is identified, its parameters are extracted when available: upper limit, slope/steepness, IC50, as well as the corresponding p-value and error for each of these parameters. Note that the IC50 value as the concentration at which the model has decreased by half its upper limit is only valid and can only be considered for the log-logistic model, which represents the expected evolution of a drug response curve. For the exponential decay and linear models, since there is no upper plateau to define a baseline, the IC50 cannot be

estimated. After extraction of the model parameters, datapoints are computed for the selected model to create a smooth curve, which will be used to calculate the AUC, on the log scale of the range of concentrations tested. This calculation of a smooth curve allowed, in addition to getting a more precise AUC value, to disregard the issue of cell cultures being tested with different concentrations since the curve is computed over the total range of tested concentrations, making AUCs comparable.

After curve-fitting comes the AUC calculation: from the smooth curves, the AUC is calculated using the *Bolstad2*[128] R package version 1.0-28. These AUCs are then normalized with the AUC of a constant function $f(x) = \alpha$ over the tested range of concentrations on the log scale, where α is the upper limit of the fitted model, which is a parameter of the curve that is computed at the curve-fitting stage.

From the input of viability rate of the cells at the different tested dosages of a given drug, the function produces several plot files: the raw data for all cell cultures in the considered drug; the raw data for the excluded cell cultures; all the smooth fitted curves and corresponding selected model type; and if requested as an argument of the function, a distinct plot for each sample with its fitted model and the model parameters. Simultaneously, the calculated AUCs values, the smooth curves datapoints for plotting, the parameters of the fitted models and the list of excluded cell cultures are also output as text files. Finally, for a smoother importing process in the steps that come after the AUCs calculations, the computed models and AUCs as well as the tables used to plot the graphs are also exported into *RData* files.

Once this method had been implemented, the AUC calculation pipeline was executed. Starting with the spectrophotometry measurements dataset, the percentage of a given cell culture duplicate population that survived a specific concentration of a given drug was calculated by dividing the corresponding spectrophotometry value by the measurement from DMSO exposure at the equivalent concentration. Then, for each drug a plot of the evolution of survival rates of the cell populations at the six concentrations were produced. This first visualization allowed for the identification of drugs that seemingly did not affect cell cultures generally and as a result were discarded from subsequent AUC calculation and analyses. This exclusion process can be illustrated with **Figure 5**: cultures survival decreases when exposed to higher doses of Fludarabine phosphate, though a few of the cultures react at lower doses than others, so the drug was kept for subsequent analyses as it may present good discrimination power between resistant and sensitive cell cultures; Pentostatin on the other hand did not appear to have a sensible effect on most if not all cell cultures at any concentration, and as a result it was not included in further calculations. Finally, the AUC computation method was applied to calculate

AUCs relative to each drug remaining. The resulting dataset, containing the AUC values for all cell cultures exposed to the drugs that were not excluded, will be hereafter called the “AUC data”.

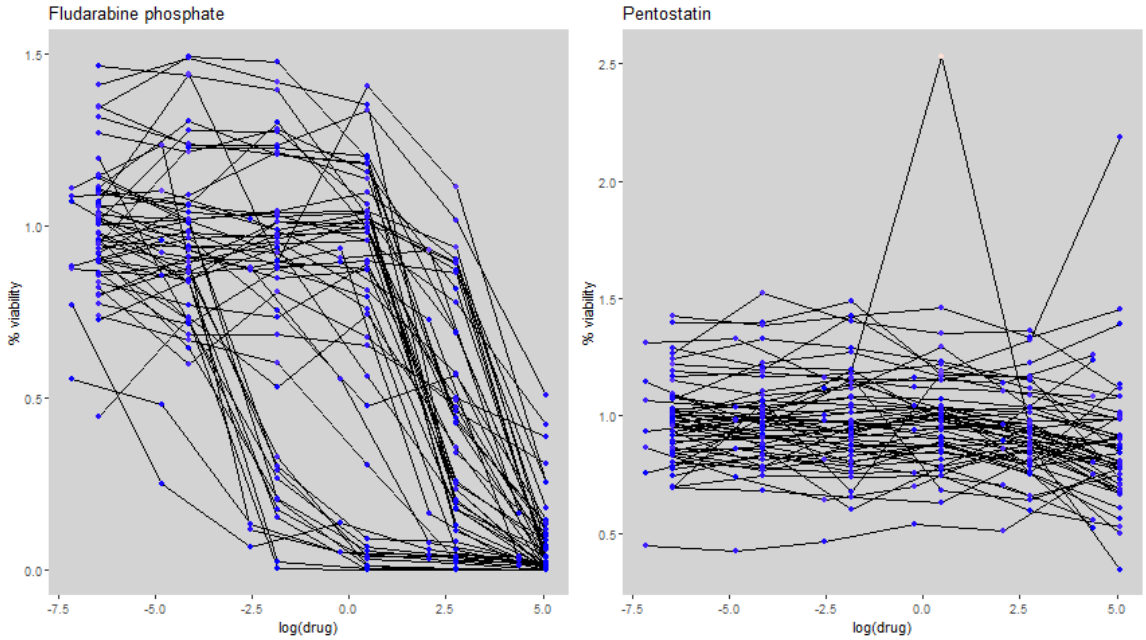


Figure 5: Example of cell cultures survival rates when exposed to Fludarabine phosphate (left) and Pentostatin (right). *Blue dots correspond to the percentage of cell population survival at the given drug concentration, averaged across all replicates for a given cell culture. Black lines represent evolution of the cell culture response at different concentrations by linking the blue dots of a given cell culture.*

Finally, for the validation of Glioblastoma cell culture models study AUC computation pipeline was executed using that data, and the corresponding drug response variables of AUC, IC50 and cells survival rate at 100 µM TMZ were extracted for all 51 cell cultures, into a “new TMZ response data” dataset.

4.3.2.2.3 Samples subsets

Finally, different subsets of cell cultures were used in the several analyses performed towards identification of predictive biomarkers of drug response. These subsets were defined as:

- “all 88 DASL cultures”, which included all samples present in the DASL microarray data
- “drug-exposed Glioblastoma cultures”, which is the subset of cell cultures from the “all 88 DASL cultures” list that correspond to Glioblastoma cell lines that were in the DASL data and in the drug exposure experiments
- “Primary Glioblastoma cultures” is the subset from “drug-exposed Glioblastoma cultures” that contained only primary Glioblastoma cell lines

- “MGMT methylated models validation cultures” corresponds to cell cultures used in the validation of Glioblastoma cell culture models study which have a methylated MGMT promoter
- “MGMT unmethylated models validation cultures” corresponds to cell cultures used in the validation of Glioblastoma cell culture models study which have an unmethylated MGMT promoter
- “MGMT unknown models validation cultures” corresponds to cell cultures used in the validation of Glioblastoma cell culture models study which for which MGMT promoter methylation status was unknown
- “MGMT methylated models validation transcriptomics cultures” corresponds to cell cultures in the validation of Glioblastoma cell culture models study which have DASL or RNA-Seq data available and have a methylated MGMT promoter
- “MGMT unmethylated models validation transcriptomics cultures” corresponds to cell cultures in the validation of Glioblastoma cell culture models study which have DASL or RNA-Seq data available and have a unmethylated MGMT promoter
- “MGMT unknown models validation transcriptomics cultures” corresponds to cell cultures in the validation of Glioblastoma cell culture models study which have DASL or RNA-Seq data available and for which MGMT promoter methylation status was unknown

4.3.2.2.4 Summary

Several datasets and subsets of genes and cell cultures have been identified and defined for use in data analyses towards identification of predictive biomarkers of drug response. **Table 7** summarizes them and lists their dimensions.

Table 7: Dimensions of all datasets and subsets of probesets and cell cultures relevant to the predictive biomarkers of drug response identification analyses

	Number of cell cultures	Number of features (type)
DASL raw data	88	29,377 (probesets)
"DASL normalized dataset"	88	26,823 (probesets)
"Unbiased Genes set"		11,007 (probesets)
"Cancer Genes set"		2,473 (probesets)
"DASL Protein-coding genes set"		17,007 (genes)
"RNA-Seq Protein-coding genes set"		16,431 (genes)
"initial IC50s dataset"	45	109 (drugs)
"bounded IC50s dataset"	45	109 (drugs)
Spectrophotometry measurements	108 (54 cell cultures x 2 duplicates per culture)	666 ((110 screened drugs + DMSO) x 6 concentrations)
"AUC data"	54	97 (drugs)
"all 88 DASL cultures"	88	
"drug-exposed Glioblastoma cultures"	45	
"Primary Glioblastoma cultures"	33	
"new TMZ response data"	51	3 (drug response variables)
"MGMT methylated models validationcultures"	19	
"MGMT unmethylated models validationcultures"	19	
"MGMT unknown models validationcultures"	13	
"MGMT methylated models validationtranscriptomics cultures" (DASL / RNA-Seq)	6 / 9	
"MGMT unmethylated models validationtranscriptomics cultures" (DASL / RNA-Seq)	6 / 8	
"MGMT unknown models validationtranscriptomics cultures" (DASL / RNA-Seq)	1 / 2	

4.3.2.3 Drugs Repurposing Project

The goal of collaboration on the Drugs Repurposing Project was to search for biomarkers that would be predictive of cell cultures response to the different drugs. To that end, in addition to the Enrichment Analysis Methods, two analysis approaches were defined and implemented to investigate covariations between gene expression profiles and drug response. The corresponding code can be found on the GitHub repository in the *Identification_of_predictive_biomarkers_of_drug_response/ EMCAnalysis_IoannisDrugs_runningAnalyses.R* script.

4.3.2.3.1 LASSO regressions

The first analysis approach relied on LASSO regression of the gene expression data using drug response data as a response variable. The LASSO fits generalized linear models to the data through penalized maximum likelihood, often leading to assignment of null coefficients to many of the variables in the model and leaving only a few as selected relevant features for the model[76], and can be used with the normalized EMC datasets. Furthermore, as the method had been previously used by the Berkeley collaborators, there was also interest in attempting to reproduce and compare results with this LASSO approach. On the other hand, as a linear regression, the method may not be appropriate for non-linear relationships between genes molecular profile and the response to drugs of the cultures.

In this analysis, the glmnet[129] R package version 4.1-1 was used for each drug to perform a LASSO regression on the DASL data, using the drug response data as a response variable. Since the LASSO regression algorithm rarely selects the same set of features twice, the robustness of the genes associated with a drug was increased by running 100 regressions and extracting genes that were in the results of at least 50 of these runs.

4.3.2.3.2 Weighted Gene Co-expression Network Analysis (WGCNA)

The second approach taken for this project was to use the WGCNA[130,131] R package. In this analysis, the genes from the DASL data were first clustered based on their expression profile. Then, the package computed the correlation between the expression of each cluster's eigengene, *i.e.* a hypothetical gene representative of the expression profile of the genes in the cluster, and external variables, such as drug response data. Unlike the LASSO method, this approach relies first on grouping genes via clustering before attempting to compare them to the response variable, *i.e.* the drug response, and defined clusters can be quite large (up to a few thousands of genes). As a consequence, the WGCNA offered a way to identify many more potential genes of interest than the handful that would be selected by LASSO, and using a very

different approach as well. Thus, both methods were implemented, in order to analyze data in different ways and increase robustness and reliability of any overlapping result.

Because of the large size of clusters, which in turn may lead to at least a few genes correlating with the drug response, only the stronger correlations with a p-value of 0.05 or less and a correlation estimate with an absolute value of 0.60 or more were selected for further investigation. In those cases, only the genes that were significantly associated with the drug response data were extracted from the cluster for enrichment analysis.

4.3.2.3.3 Analyses execution

For the drug repurposing project, for which the goal was to identify predictive biomarkers of drug response in Glioblastoma cell cultures, the samples analyzed were the drug-exposed Glioblastoma cultures.

The initial run of the analysis was done using the DASL data subsetted with the unbiased genes set and drug-exposed Glioblastoma cultures, and the bounded IC50s dataset as the drug response data.

The first step performed in that analysis was the investigation of the DASL data, which underwent a PCA to look into the presence of outliers and in the impact of cofactors such as batch effect or the Primary/Recurrent Glioblastoma diagnosis of the tumour. Following that PCA investigation, it came out that interestingly there did not seem to be a strong difference in gene expression profiles between Primary and Recurrent Glioblastoma samples. As a result, it was decided to analyze the data using both the drug-exposed Glioblastoma cultures and the Primary Glioblastoma cultures subsets parallelly and compare the results.

Then, both LASSO regressions and WGCNA were performed to analyze the unbiased genes set / drug-exposed Glioblastoma cultures DASL dataset and the unbiased genes set / Primary Glioblastoma cultures DASL dataset, with the correspondingly subsetted bounded IC50s dataset as response variable, to identify genes whose expression would suggest a predictive potential for response to each drug.

The IPA functional pathways investigation method was used to identify cellular processes associated with the results.

Although these results showed promise, the bounding approximation was thought to be too much of an issue due to the bias introduced by approximating the extreme outlier values to values comparable to the rest of the dataset. Following discussions and searches, the EMC team came up with the suggestion for the AUC alternative. Furthermore, an interest for an

approach focusing on cancer genes was also put forward. As a result of these exchanges, the cancer genes set was defined, and the AUC calculation pipeline was implemented and executed to compute AUC values for all cell cultures exposed to each drug.

For consistency however, the same pipeline as the initial analysis run was followed and as a consequence, LASSO and WGCNA methods were applied to a total of four datasets to identify gene expressions that could be associated with AUCs as response data:

- the cancer genes set / drug-exposed Glioblastoma cultures DASL dataset
- the cancer genes set / Primary Glioblastoma cultures DASL dataset
- the cancer genes set / drug-exposed Glioblastoma cultures DASL dataset
- the cancer genes set / Primary Glioblastoma cultures DASL dataset

The obtained results were then investigated for enriched functional pathways using the computational Gene Ontology enrichment function.

4.3.2.4 Berkeley LASSO Analysis review

In addition to performing a complete analysis of the EMC data for the Drugs Repurposing project, I also looked into the results of the LASSO analysis performed at the Berkeley University prior to my involvement in the project. That prior analysis partly inspired my own, namely for the normalization tools for the DASL data, using the approach described in 4.3.2.3.1 LASSO regressions to identify predictive biomarkers of drug response, and identification of genes of interest from the LASSO-selected genes through IPA software functional analysis. Thus, we were interested in comparing results of my own analysis with results from the Berkeley analysis to potential validate these results as well as assess their robustness, considering only an overview of the steps of the Berkeley analysis was communicated rather than its exact protocol, so it could not be reproduced exactly.

First, reproducing the identification of genes of interest in IPA based on the list of genes associated with a given drug from the Berkeley LASSO results was attempted. This was done using the process from subsection 4.3.1.1 IPA functional pathways investigation to determine links between LASSO-selected genes and the drug and identify functional pathways involved. In the resulting network, the genes of interest were defined as the nodes with a high connectivity relatively to others (connected to three or more nodes) in the network or involved in at least three functional pathways related to cancer, signaling, or neuronal activities.

Secondly, the results of the Drugs Repurposing Project analyses were compared to the Berkeley LASSO analysis results.

The list of drugs that the Berkeley LASSO analysis associated genes with were compared to the list of drugs identified as promising from the Drugs Repurposing Project results. For the drugs that were present in both sets of results, the associated list of genes (and the functional pathways they belong to) for each drug from both of the pipelines were compared, to determine whether the results overlapped, *i.e.* the same genes and/or functional pathways were identified for the same drugs by both pipelines, and validated the approach.

4.3.2.5 Validation of Glioblastoma cell culture models

This study was about validating EMC Glioblastoma cell cultures models, by demonstrating that these models react in a similar way to TMZ as the parental tumors they are derived from, hence validating them as appropriate models for Glioblastoma studies. The corresponding code can be found on the GitHub repository in the *Identification_of_predictive_biomarkers_of_drug_response/RNASeq_correlation_analysis.R* script.

4.3.2.5.1 Correlations of patients and cell cultures responses

A first part of the study was to perform correlation tests between response to TMZ of the cell cultures and response to TMZ of the parental tumours.

The response to TMZ of the parental tumours was approximated by the patients' Progression Free Survival (PFS) and Overall Survival (OS), since they were treated with TMZ and thus a shorter survival should correspond to a lower sensibility to the drug. For the cell cultures, the response to TMZ used was either the AUCs or IC50s from the new TMZ response data. Furthermore, since the MGMT promoter methylation status is a positive predictive biomarker of TMZ response, the correlations were calculated with the data from three subsets of cell cultures and corresponding patients: all cultures available in the new TMZ drug response data, the MGMT methylated models validation cultures and MGMT unmethylated models validation cultures.

The IC50s, PFS and OS data was log-transformed.

The normal distribution of each response variable with each considered subset was evaluated with a Shapiro-Wilk normality test. As a consequence, Pearson's method was used to test correlation only between response variables that were both normally distributed. Otherwise, correlation tests were performed using Spearman's method. **Table 8** indicates the correlation method used for each test.

	Responses tested			
Subset	**Log(PFS) x AUCs**	**Log(PFS) x Log(IC50s)**	**Log(OS) x AUCs**	**Log(OS) x Log(IC50s)**
Methylated (n = 23)	Pearson	Pearson	Spearman	Spearman
Unmethylated (n = 24)	Pearson	Pearson	Pearson	Pearson
All (n = 47)	Spearman	Spearman	Spearman	Spearman

Table 8: Correlation test method for each cell culture x parental tumour response design *in the models validation study.*

4.3.2.5.2 Predictive biomarkers identification

In the continuation of the Drug Repurposing Project objectives, identifying biomarkers predictive of TMZ response in the cell cultures was attempted.

For this purpose, normalized DASL dataset and GLIOTRAIN cell lines RNA-Seq data normalized with the *vst* method from the DESeq2 R package were filtered to keep only the respective the Protein-coding genes and cell cultures used in the validation of Glioblastoma cell culture models study. These subsets were analyzed to search for biomarkers, using the new TMZ response data to identify genes for which the expression profile correlates with drug response.

However, rather than LASSO or WGCNA approaches, the investigation of potential biomarkers was done by running correlation tests between the expression profile of each gene against each response to drug variable. A given gene expression dataset was investigated as follows:

- For each gene in the dataset, the expression profile was tested for normality with a Shapiro-Wilk normality test.
- A correlation test was performed between the gene expression and each of the drug response variables (AUCs, IC50s and cell survival rate at 100 µM TMZ). If both the gene expression and the drug response variable were normally distributed, Pearson's correlation was used; otherwise, Spearman's correlation was preferred.
- Once the correlations tests have been performed, the p-values obtained from the correlation tests between all genes expression and a given drug response variable were adjusted for multiple testing using the FDR method. Genes for which the adjusted p-values were below 0.05 were considered to significantly correlate with the corresponding drug response variable.
- The genes significantly correlated with all three response variables were then compared, and the ones correlated with at least two response variables were estimated

to be robust findings and compiled into a “signature” list of genes associated with the analyzed dataset.

- The signature genes were input into the function described in the subsection 4.3.1.2 Gene Ontology enrichment function, to determine the functional pathways involved with these genes.

These steps were applied to analyze both gene expression datasets (DASL and RNA-Seq), with three subsets of cell cultures (MGMT methylated, unmethylated, or indifferent) for which expression data was available.

Following the first run of the analysis, it turned out that for all the datasets, no gene passed the 0.05 adjusted p-value cutoff. As a consequence, the analysis was re-run without that step, to consider genes as correlated with a drug response variable if the corresponding correlation test had a p-value below 0.05.

In addition, following promising but unpublished results from an experiment performed by the EMC team to determine the efficacy of Omacetaxine mepesuccinate and Cytarabine hydrochloride on glioblastoma cell cultures, the same analysis was performed on newly produced response data for these drugs, to investigate potential predictive biomarkers of response in the same way as was done for TMZ.

4.3.3 GLIOTRAIN Data Analysis

The analysis performed on the data available for the project consisted mainly in a DEA between ST and LT survivors-derived samples in each dataset. This was done with the parental tumours RNA-Seq, focal events datasets, as well as the EMC DASL data.

Furthermore, to validate these results, the genes identified by DEA of a given dataset were investigated in other datasets, to determine whether they were at least following a similar trend. This validation was done using the RNA-Seq, focal events, EMC DASL, TCGA RNA-Seq and TCGA Copy Number Variations (CNV) datasets, all of which were further subsetted into an IT survivors dataset and a ST+LT survivors dataset. The ST+LT datasets from GLIOTRAIN and EMC would be used in a differential analysis to compare ST samples to LT samples, while the IT datasets would be used to validate the results of that comparison. The TCGA Glioblastoma datasets were obtained from the cBioPortal platform[132], by searching for “glioblastoma” and downloading the Glioblastoma (TCGA, Cell 2013) archive.

4.3.3.1 Preparation and Normalization

The first step for this was to prepare the data for analysis. The corresponding code can be found on the GitHub repository in the *GLIOTRAIN_Data_Analysis/GTAnalysis_Preprocessing.R* script.

For the RNA-Seq data, the ST and LT survivors samples were extracted from the parental tumours raw count dataset. In addition, to produce the validation datasets the parental tumours raw count data was normalized using the *vst* method from the DESeq2[133] R package version 1.30.1, using information about the MGMT promoter methylation status, institute of origin and sequencing batch information as cofactors. That normalized data was separated into two to produce a ST+LT samples normalized RNA-Seq data on one hand and an IT samples normalized RNA-Seq data on the other.

Regarding the focal events dataset, even though it contains numerical values, these numbers represent categorical data. As a consequence, the dataset was not normalized, and only separated between ST+LT and IT samples subsets.

The normalized EMC DASL data will all Glioblastoma samples was used for this analysis as well, and separated between ST+LT and IT samples subsets.

Similarly as with the GLIOTRAIN focal events data, the TCGA CNV data was not normalized. In addition, the already normalized TCGA RNA-Seq Z-scores data was used as well. Both datasets were also separated between ST+LT and IT samples subsets.

Table 9 summarizes the number of samples of each type, as well as the number of features for each dataset.

Dataset	ST samples	IT samples	LT samples	Number of features
GLIOTRAIN RNA-Seq	26	69	31	45,623
GLIOTRAIN focal events	26	67	32	49
EMC DASL	24	16	6	26,823
TCGA RNA-Seq	98	49	5	20,531
TCGA CNV	287	195	40	24,174

Table 9: Dimensions of the subsets used for the DEAs and validation of the results.

4.3.3.2 *Differential Expression Analysis*

DEA of the GLIOTRAIN RNA-Seq ST+LT data was performed using the DESeq2 R package with the MGMT promoter methylation status, institute of origin, sequencing batch information and the survival group as covariates of the model. The cutoff for significance was 0.05 for p-values adjusted for multiple testing (FDR). The corresponding code can be found on the GitHub repository in the *GLIOTRAIN_Data_Analysis/GTAnalsysi_DEA.R* script, under sections *II.1 Run analysis* and *II.2 Explore results*.

For the focal events data, considering the values are ordered categories, a logistic regression model was fitted to the data for each chromosomal region using the *polr* method of the MASS[134] R package version 7.3-53, accounting for MGMT promoter methylation status, institute of origin, sequencing batch information and the survival group as covariates of the model. A chromosomal region was considered to have a significantly different copy number profile between ST and LT samples when the corresponding coefficient of the model had an absolute value equal or above 1. Regions for which this was the case were screened to extract the genes that belong to it, by comparing start and end positions of the regions and genes definition from the ENSEMBL database.

Finally, the EMC DASL data analysis consisted of performing t-test to compare the expression profile of the ST and LT samples for all probesets present in the dataset. The cutoff for significance was 0.05 for p-values adjusted for multiple testing (FDR).

The genes identified through DEA of the RNA-Seq and focal events datasets were input into the Gene Ontology enrichment function to investigate their synergy and biological relevance to the context of Glioblastoma. Since only two genes were identified from the EMC DASL analysis, an enrichment analysis was not necessary. Their roles were investigated individually by first searching for them on PubMed, and more broadly Google, to identify their functions, pathways they are involved with, and potential implication in glioblastoma. Since this first approach was successful in providing an overview of these genes and their functions, further investigation with more specialized resources was not pursued.

4.3.3.3 *Results validation in other datasets*

In order to increase confidence in the genes identified in the DEA of the data as relevant to overall survival of the patients, their profile was checked in other datasets. The goal of this was not necessarily to see if these genes would be significantly different between ST and LT,

otherwise they would also appear as such in the corresponding DEA, but to be less stringent and determine whether the profile follows the same trend as observed in the DEA.
The corresponding code can be found on the GitHub repository in the *GLIOTRAIN_Data_Analysis/GTAnalsysi_DEA.R* script, under sections *I.2 Validation Functions* and *II.3 Compare to other datasets*.

To that end, the following steps were used for each DEA-identified gene, in each validating dataset:

1. the gene was searched in the validating dataset. In the case of CNV validating datasets, the chromosomal regions were checked to see if they covered the position of the gene
2. if the gene was found in the validating dataset, a test was performed to see if there was potentially a pattern of expression profile associated to patients' survival. To be less stringent than the DEA in order to see if the data even suggests a trend, a p-value cutoff of 0.2 was used to consider a gene interesting for this validation.
3. if the gene passed that test, a correlation test was performed to determine whether or not the pattern followed the same direction as identified in the DEA, *i.e.* if expression profile in the validating dataset co-variated with patients survival in the same way as in the dataset from which the was identified

Note that the patients' survival response variable used in the tests was different depending on which subsets of samples were compared. For the ST+LT validating datasets, samples were compared as groups, so the response was categorical: either ST or LT. For the IT validating datasets, the response variable used was the continuous overall survival value for each patient. Furthermore, each validating dataset contained different data types. Therefore, the test to determine the existence of a pattern and the correlation test to identify its direction is different for each validating dataset. **Table 10** summarizes which tests were used for each validating dataset.

Validating dataset	Pattern-identification test	Correlation test
GLIOTRAIN RNA-Seq ST + LT	t-test	Spearman
GLIOTRAIN RNA-Seq IT	Pearson correlation	Pearson
GLIOTRAIN focal events ST + LT	Kruskal-Wallis	Spearman
GLIOTRAIN focal events IT	Spearman correlation	Spearman
EMC DASL ST + LT	t-test	Spearman
EMC DASL IT	Pearson correlation	Pearson
TCGA RNA-Seq ST + LT	t-test	Spearman
TCGA RNA-Seq IT	Pearson correlation	Pearson
TCGA CNV ST + LT	Kruskal-Wallis	Spearman
TCGA CNV IT	Spearman correlation	Spearman

Table 10: Statistical tests used for DEA results validation in other datasets

5 Results

5.1 Glioblastoma Disease Map

A major part of the PhD was dedicated to defining a Glioblastoma Disease Map, *i.e.* a molecular interactions network representing driver alterations of signaling pathways underlying the disease. This was done based on literature detailing known Glioblastoma-specific pathway alterations.

5.1.1 Genetic Alterations Representation

My work towards the representation of Genetic Alterations in a Disease Map led to the definition of representation rules that fitted the needs to represent in the Glioblastoma Disease Map the mutations affecting Glioblastoma tumor cells that were regularly referenced in the literature. These rules could be categorized into three types, described in more details further below: chromosomal aberrations, mutations altering function or transcription rate, and representation of mutations that were mutually exclusive or on the contrary systematically co-occurring. In addition to using it to build the Genetic Alterations submap, I also introduced the resulting model to the Disease Map Community as a poster at the 4th Disease Map Community Meeting in 2019[135].

5.1.1.1 Chromosomal Aberrations

Amplification as well as Homozygous or Heterozygous Deletion of whole chromosomal segments are frequent events in cancer. To model them, chromosomes or loci are presented as hypothetical complexes, so that the genes they contain may be included in the model to highlight the importance of the mutation, and use the MeSH terms “Chromosome Duplication”, “Loss of Heterozygosity” or “Chromosome Deletion” represented as a phenotype that would catalyze transition of the loci into an annotated “Amplified”, “KO” or “Homozygous deletion” state respectively, as illustrated by **Figure 6**. Since increased and decreased expression is implicit from the variation of copy number, it is not explicitly represented. In addition, though the chromosomal region should be preferred wherever possible, amplification or deletion of single isolated genes is not excluded, in cases where the literature only references duplication or deletion of that one gene rather than its locus. In these cases, the mutation-inducing phenotype should be represented with MeSH terms “Gene Duplication”, “Loss of Heterozygosity” and “Gene Deletion”.

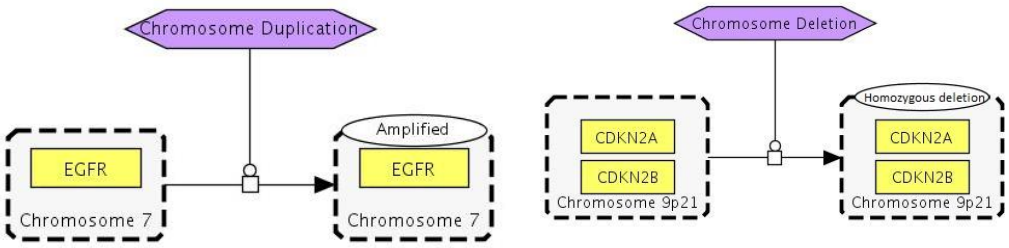


Figure 6: Examples of modelling a chromosomal amplification (left) and deletion (right). *Yellow boxes represent genes, dash-outlined grey boxes are chromosomes/loci as hypothetical complexes, purple items are phenotype nodes representing mutations. A full arrow represents a transition, and a circle-ended arrow represents catalysis.*

This representation is complementary and does not overlap with the representation of other mutations, which should still be modelled separately. Furthermore, since it is possible that a locus subject to copy number variations may contain genes that can also undergo mutations altering their function or transcription rate, these mutations should also be represented, with the chromosome or locus represented as a container rather than a hypothetical complex, and separately from the chromosomal aberration mutation, as exemplified in **Figure 7**.

The difference in representation of the locus comes from its role in the model and the Disease Map standards:

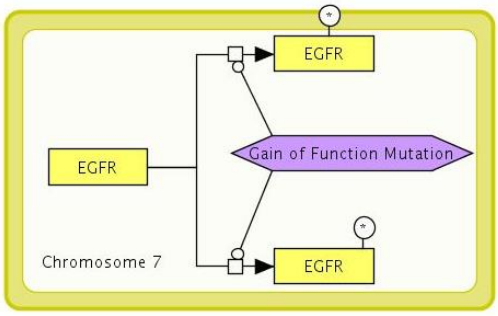


Figure 7: Representation of the mutation of a gene within the corresponding locus. *Thick yellow line boxes the chromosome as a container.*

- for duplications and deletions, the locus is the unit than undergoes transformation. This cannot be modelled with a container object, while the hypothetical complex allows not only to represent modification as needed, but also to contain genes relevant to the understanding of the importance of the mutation to the disease while making it explicit that this component is not a complex as strictly defined in the standards.
- for other gene mutations, the container is a more appropriate representation since Disease Map standards require that transformations occurring within a given biological entity be represented this way, while also strongly suggesting avoiding any transformation within a complex.

5.1.1.2 Mutations altering function or transcription rate

Another important issue to solve was how to meaningfully represent all known mutations for a given gene. Since mutations can be of different types (substitutions, indels, inversions...), occur

at different positions, lead to different amino-acid chains, *etc.*, there could be dozens or hundreds of mutations that would need to be modelled.

However, qualitatively speaking, these events generally result in one of four categories: increased or decreased gene transcription rate or increased or decreased protein activity. By reducing the modelling possibilities of mutations to these four outcomes, I was able to define a model that could faithfully represent the impact of mutations on the disease, while also avoiding representing each of them individually which could greatly overload the Disease Map. The idea is that all mutations leading to a similar outcome on the final protein are represented by the alteration of a single hypothetical site. If there are other mutations that can affect the same gene but leading to a different type of outcome, for instance if some mutations result in increased efficiency of the gene while other mutations lead to increased transcription of the gene, they are represented as two distinct hypothetical sites on the gene. Since the Disease Map is a qualitative model rather than a quantitative representation of mechanisms, the relative strength of the modification resulting from the different mutations can be overlooked, as only the qualitative result needs to be represented.

Thus, the mutations are represented by the transition of the gene from its normal state to a state where the mutation that occurred is represented by a '*' (labelled as "Don't care" in the CellDesigner software and meaning "any alteration") symbol on the hypothetical mutation site. That transition is catalyzed by a phenotype of either "Gain of Function Mutation" or "Loss of Function Mutation", as illustrated by **Figure 8**. All mutations that result in a similar outcome can be provided as notes for the hypothetical site they are associated with, hence the information is still available.

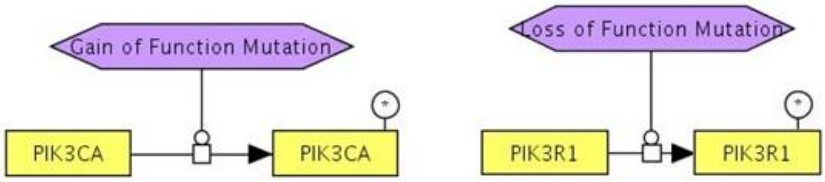


Figure 8: Example of how gain of function (left) and loss of function (right) mutations are represented in the model.

The actual impact of these mutations (*e.g.* lower transcription rate, lower binding potential, loss of activating phosphorylation site, *etc.*) must then be modelled appropriately in the Disease Map. While deletion of an entire gene is enough on its own to explain how that affects a pathway, it is not the case for Gain or Loss of Function mutations. As previously described, these can further be divided into two categories of outcome: increased (or decreased,

respectively) transcription rate of the gene, or increased (or decreased, respectively) activity of the mutated protein. The alteration of transcriptional rate can be represented by a direct catalysis or inhibition of the transcription of the normal gene by the mutated gene. As for modifications of the activity of the protein, it depends on the extent to which the mutated protein is characterized. If the literature provides a clear description of the different behavior, for instance if the mutated protein is constitutively active or doesn't require an intermediate to perform its role, the corresponding interactions with the mutated protein can be modelled. If no indication is provided, which is often the case, the only possible representation of the result of the mutation is to infer whether the mutation is a gain or loss of function, based on the gene and its role in the pathways, and the effect of the mutation is then represented accordingly as a positive or negative influence on the normal gene. In these cases, a note that indicates this interaction is inferred is added to the positive/negative influence interaction. These different influences of mutations on the Disease Map network are illustrated in **Figure 9**.

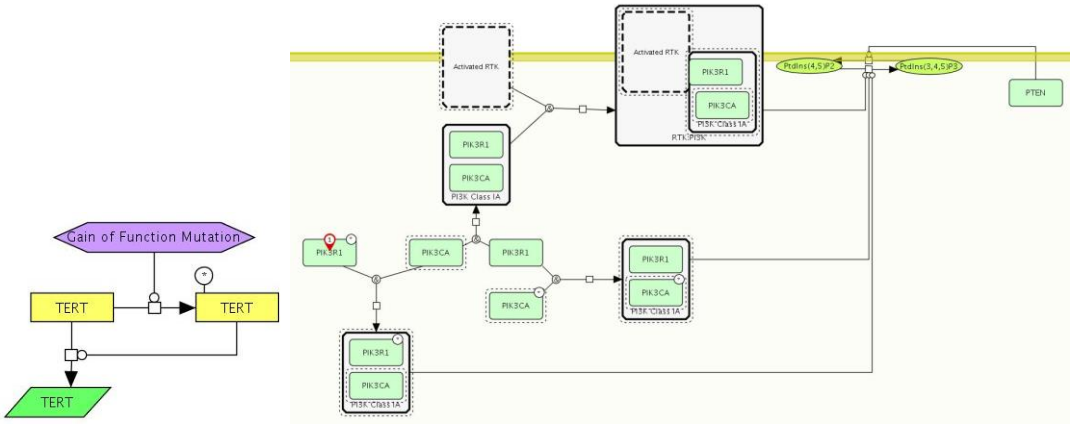


***Figure 9: Examples of representations of mutations impact in the Disease Map. Left**: Mutated TERT leads to increased transcription. Right: Mutated PIK3R1 and PI3KCA lead to increased efficiency of the PIK3R1:PIK3CA complex. Bright green node on the left represents mRNA; on the right, round-cornered light green nodes are proteins, thick black lines containing several nodes represent a complex, and round green nodes are small metabolites.*

5.1.1.3 Mutual Exclusion and Co-occurrence patterns

In some cases, the mutation of a single gene is enough to significantly affect the pathway to which the gene belongs. Then, due to selection pressure, mutations in other high-impact genes of the pathway are rare, and almost never observed simultaneously. On the other hand, when the mutations are not enough to disable or enable the pathway on their own, they can be found to be systematically co-occurring with more potent mutations of the pathway. Since several such patterns have been observed in Glioblastoma, they also needed to be represented in the model. The solution adopted was to model them with a direct positive or negative influence

between the products of the corresponding alterations, as illustrated in **Figure 10**.

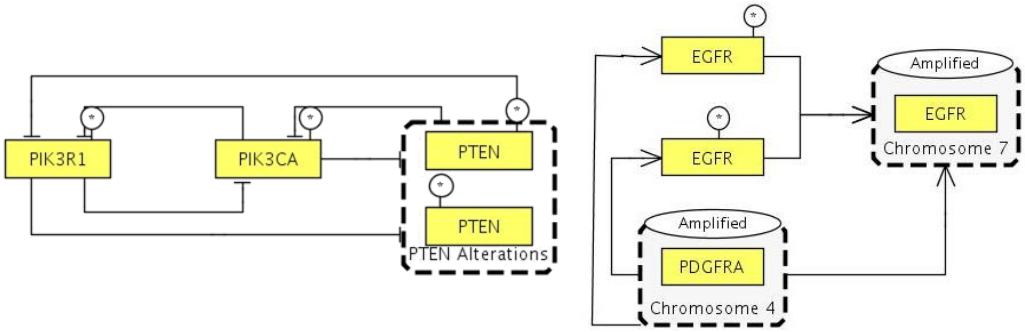


Figure 10: Example of mutually exclusive (left) and co-occuring (right) patterns between genetic alterations. *Perpendicular end of arrow represents negative influence of the starting node on the target node, whereas stick end of arrow represents a positive influence.*

5.1.2 Produced Disease Map

An initial screening of the literature identified the RTKs-PI3K-AKT/MAPK pathway, the RB pathway and TP53 pathway to be key drivers of Glioblastoma mechanisms. As a consequence, I built a molecular interaction network representation of each of these pathways as well as Glioblastoma-specific genetic alterations into independent submaps, which were combined once completed.

5.1.2.1 RTKs-PI3K-AKT/MAPK Submap

This first submap was the one which required most work. That is because while this pathway has a lot of different targets, and thus can activate or inhibit many pathways and cellular functions, its components that are typically altered in Glioblastoma (*EGFR*, *PDGFRA, PIK3CA, PTEN*…) are all upstream of the signaling cascade. That made it difficult to determine which targets were more likely to be affected in the context of the disease. Among the potentially interesting downstream effectors of the pathway, particular attention was given to the *FOXO* family of transcription factors, which have targets affecting apoptosis and the cell cycle, as well as *TSC2* and *GSK3B* which regulate cell growth processes, all of which are particularly relevant to tumor cells.

Beyond the high number of cellular functions this pathway could influence, in the context of investigating resistance mechanisms it is also interesting to note a relatively strong connectivity of the network: targeting a single gene for treatment may not prove to be efficient very long as there is usually at least one alternative path that allows to reach the same downstream targets.

To get into more details, the submap can be divided into several parts which overlap to some extent: the activation of Receptor Tyrosine Kinases and subsequent recruitment of the *PI3K*

complex, the transduction of the signal from *PI3K* to *AKT*, the activation of cell growth processes, inhibition of *FOXO* transcription factors, inhibition of apoptotic pathways, influence on the cell cycle, and finally the RAS/RAF/ERK cascade.

As mentioned earlier, the common starting point of the PI3K/AKT pathway is the activation of RTKs. In particular, *EGFR*, *VEGFR* and *PDGFRA* are well characterized, as they have been found to be upregulated in Glioblastoma. Upon activation of these receptors through binding with their respective ligand, the RTKs recruit the *PI3K* complex to the membrane, which triggers the release of the catalytic site of the *PIK3CA* subprotein.

Transduction of the signal from the activated RTK to *AKT* is done via the phosphatidinosytol 3-phosphate (PtdIns(3,4,5)P3) which is a phospholipid found in the membrane. The activated RTK:*PI3K* complex phosphorylates the PtdIns(4,5)P2 phospholipid to produce PtdIns(3,4,5)P3. This reaction can be reverted by the *PTEN* protein. The PtdIns(3,4,5)P3 recruits the *AKT* protein to the membrane, where it needs to be phosphorylated by both *PDK1* (which needs to be recruited to the membrane by a different PtdIns(3,4,5)P3), and the *mTORC2* complex to be fully activated, although some evidence suggest phosphorylation by only one of them may be enough to activate *AKT*, even if partially. Once fully activated, *AKT* is released from the membrane and can either remaining the cytoplasm or be transported to the nucleus. Inactivation of *AKT* was not extensively investigated, although it was found that the *PP2A* family and/or the *PHLPP2* protein may have a role in it.

Activated *AKT* promotes cell growth by phosphorylating the *TSC2* protein, leading to inhibition of the *TSC1:TSC2* complex which dephosphorylates the *RHEB*-associated GTP. Under a high enough concentration of *RHEB*:GTP, the *mTORC1* complex is activated and promotes cell growth through inactivation of *EIF-4EBP1* which is a cell growth inhibitor and activation with the help of *PDK1* of *RPS6KB1* which is involved in biosynthetic processes.

The *FOXO* family (*FOXO1, FOXO3, FOXO4* and *FOXO6*) of transcription factors is particularly interesting as targets of *AKT*, as they regulate transcription of genes involved in several different processes. Phosphorylation of the *FOXO* transcription factors by *AKT* (or by *SGK* which can phosphorylate them at the same sites) can happen both in the cytoplasm or in the nucleus, and leads to their sequestration by a *YWHA* (or *14-3-3*) protein in the cytoplasm. Otherwise, they positively regulate *BCL2L11, FASLG, TNFSF10, TRADD, BCL6* (which inhibits the apoptosis inhibitor *BCL2L1*), leading to promotion of apoptosis. They also positively regulate transcription of *CDKN1A, CDKN1B* and *RBL2* and repress transcription of *CCND1*

and *CCND2*, which leads to G1/S phase transition repression, partly through activation of the *RB1* pathway. Therefore, *AKT* mediated inactivation of the *FOXO* transcription factors inhibits paths towards apoptosis activation and cell cycle inhibition, which are key factors in tumor survival and development.

Activated *AKT* has other ways to inhibit apoptosis than just through the *FOXO* family as described above. It can phosphorylate *MDM2*, which promotes its transport to the nucleus where it can regulate and inhibit *TP53*-induced apoptosis. It also phosphorylates the *BAD* protein, which is then unable to inhibit the apoptosis inhibitor *BCL2L1*.

Similarly, the cell cycle is activated by *AKT* through inactivating phosphorylation of *GSK3B*. When active, *GSK3B* both inhibits the G1/S phase transition inhibitor *MYC*, and targets the *RB* pathway inhibitor *CCND1* for degradation.

Finally, the RAS/RAF/ERK cascade was added to this submap although it is relatively autonomous from PI3K/AKT, because there is some cross talk between the two. In particular, upon RTK activation, the *RAS*:GDP complex at the membrane is phosphorylated into an activated *RAS*:GTP complex, a phosphorylation that can be reverted by the *NF1* protein. The *RAS*:GTP complex is able to recruit and activate the *PI3K* complex, which can be seen as an alternative path for *PI3K* activation instead of *PI3K* being directly recruited and activated by the RTK.

Aside from that, once activated, a couple of *RAS*:GTP can associate to recruit a couple of *RAF* protein which in turn recruit a couple of *MEK1* or *MEK2* proteins. That large *RAS/RAF/MEK* complex then activates *MAPK1* or *MAPK3* through phosphorylation. *MAPK1/3* is involved in feedback loop regulations, as it is able to trigger disassembly of the *RAS/RAF/MEK* complex, phosphorylate the *SOS1* leading to its *YWHA*-mediated sequestration and thus preventing it from playing its role as a necessary intermediate for the RTK phosphorylation of *RAS*:GDP, and promote transcription of the *DUSP6* protein which can dephosphorylate *MAPK1/3* and thus inactivate it.

Overall, the downstream influence of the RTK/PI3K pathway is quite wide, though I was not able to capture all proteins and interactions relevant to Glioblastoma. As a consequence, priority was given to targets known to be involved in typical hallmarks of cancer such as cell growth, proliferation and apoptosis for instance. However, the core Glioblastoma-specific alterations of this pathways are represented. Indeed, among the more frequently characterized genetic alterations of Glioblastoma, there are

- upregulation *of EGFR, PDGFRA, MET* and *ERBB2* RTKs through either mutation or duplication, or they are mutated for increased efficiency. In particular, in IDH-wildtype Glioblastoma *EGFR* is almost always either upregulated or mutated for increased efficiency or sensitivity, and there is even a variant, named *EGFRvIII*, which is frequently observed and is constitutively active and thus doesn't require ligand binding to initiate signaling
- mutations of the *PIK3R1* subprotein of *PI3K* which would disable its repression on *PIK3CA* catalytic site, or mutations to *PIK3CA* to increase its efficacy or allow it to phosphorylate PtdIns(4,5)P2 without recruitment by the RTKs, or a loss or inactivation of the *PTEN* protein that reverts PtdIns(3,4,5)P3 into PtdIns(4,5)P2
- loss or inactivation of the *NF1* protein that reverts the activated *RAS*:GTP complex into the inactive *RAS*:GDP
- although relatively rare, *RAS* mutations for increased sensibility have also been observed

As a result, while the pathway presents many downstream effectors, the main driver mutations of this signaling pathway occur at its very beginning.

5.1.2.2 RB Submap

The Retinoblastoma submap was much simpler to assemble, since *RB* associates itself with *E2F* family proteins to inhibit transcription of their targets. Typical alterations of this pathway in Glioblastoma lie in either the inactivation of *RB* or the activation of its inhibitors.

RB1 has several upstream regulators. Upon DNA damage or exit from mitosis, it becomes hypophosphorylated, which is its activated state. Active Cyclin Dependent Kinases such as *CDK4, CDK6* and *CDK2* can deactivate it through phosphorylation. These CDKs are repressed by *CDKN1B* and *CDKN2A*. In addition, the *PP1* complex has been shown to facilitate *RB1* dephosphorylation and repress its rephosphorylation.

In its hypophosphorylated, active state, *RB1* is able to bind to *E2F* family transcription factors to inhibit their transcriptional activity but not their ability to bind to their targets. As a result, *RB1* actively inhibits transcription of these *E2F* targets since it prevents other transcription factors and transcriptase to bind to the DNA. Once the concentration of active *CDKs* is high enough, *RB1* hyperphosphorylated and releases the *E2F* transcription factors, thus allowing transcription of their targets. However, binding of *RB1* to *E2F1* specifically is slightly different

than with other *E2F* family members, as even when hyperphosphorylated, *RB1* is still able to bind to *E2F1* and partially inhibit it, specifically in regard to apoptosis-promoting *E2F1* targets.

Another, lesser function of *RB1* is that it can target for degradation the gene *SKP2*, which promotes *CDKN1B* degradation. *RB1* thus participates in a positive feedback loop, where it prevents degradation of an inhibitor of *RB1* inhibitors.

Modifications to the RB1 pathway in Glioblastoma typically occur through:

- deletion or disabling of *RB1*,
- amplification of *CDK4* and *CDK6*,
- deletion of the locus that contain the *CDKN2A*, *CDKN2B* and *ARF* genes

which all limit or prevent *RB1* to inhibit the cell cycle.

5.1.2.3 TP53 Submap

In Glioblastoma, the *TP53* pathway is usually altered through the *MDM2, MDM4* and *ARF* regulators of *TP53* transcriptional activities. However, although there are many publications listing targets of *TP53*, barely any look into it in the specific context of Glioblastoma, which led to the inclusion of *TP53* targets indiscriminately.

Beyond the *TP53* targets, a lot of attention was given to its regulation by *MDM2* and *MDM4*. Indeed, *TP53* can be bound by an *MDM2* or *MDM4* monomer, by and *MDM2* homodimer or by an *MDM2:MDM4* heterodimer. The bound *TP53* can remain sequestrated, but also transported to the cytoplasm and targeted for degradation by the proteasome.

Upon DNA damage, TP53 is released from its inhibitors by proteins such as *PRKDC*, *ABL1*, *CHEK1*, *CHEK2* and *ATM*, and is then able to initiate transcription of its targets.

Modifications to the TP53 pathway in Glioblastoma typically occur through

- deletion or disabling of *TP53*,
- amplification of *MDM2* and *MDM4*,
- deletion of *ARF*

which all limit or prevent *TP53* transcriptional activities towards apoptosis, cell cycle arrest, autophagy, DNA repair, cellular senescence, *etc.*

5.1.2.4 Genetic Alterations Submap

Finally, using the model defined before, the genetic alterations submap was relatively straightforward to build. Indeed, the existence of genetic or transcription rates abnormalities in

a specific disease is easier to identify and report than to characterize the effect of point mutations on a given protein and the proteins it interacts with.
Although this submap is not sufficient on its own since it's necessary to represent the abnormal behaviour of mutated proteins, it was important to model these alterations at the genetic level rather than just the mutated proteins as they are at the core of the disease mechanisms, but also because this submap displays the panel of mutations that may happen in a primary tumor, as well as all other mutations available to it to overcome treatment and develop resistance mechanisms.

As most of the well-characterized genetic alterations have been mentioned in the context of the other submaps, there are only a few notable ones that were not part of the three main pathways but should still be mentioned:

- The *MGMT* promoter is frequently methylated, leading to a repression of transcription. This is actually a positive predictive biomarker for response to treatment, as *MGMT* is involved in DNA repairing mechanisms that mitigate the effects of TMZ treatment.
- *TERT* is often upregulated

Finally, across these genetic alterations in Glioblastoma, patterns of mutual exclusivity of mutations or on the contrary frequent co-occurrence of mutations have been observed. These patterns are:

- *PDGFRA* amplification occurs rarely alone, and is typically present alongside *EGFR* amplification, upregulation or increased efficiency mutation. Furthermore, *EGFR* upregulation and increased efficiency mutation also tend to happen alongside *EGFR* amplification. This shows that amplification of the *EGFR* gene, regardless of other mutation, is one of the most important alterations in Glioblastoma, as it is present in the majority of IDH-wildtype Glioblastoma, and is not prevented by the emergence of alternative ways of increasing *EGFR*-mediated initiation of the PI3K pathway.
- Mutations of *PIK3R1, PIK3CA* and *PTEN* are mutually exclusive. This suggests that only one of them is enough to topple the balance of PtdIns(3,4,5)P3 concentration towards the pathways activation. It also means that targeting any one of these proteins for treatment may not be enough, as mutations to any of the other two may compensate the effect of treatment.
- *NF1* mutations are mutually exclusive with *EGFR* mutations. That suggests that in some cases the *RAS* protein is enough to trigger the whole pathway, and targeting only *RTKs* may result in the rise of activation of *AKT* via *RAS* instead.

- *RB1* mutations tend to occur alongside *NF1* mutations, *PTEN* mutations and *TP53* mutations. This suggests that the three pathways (PI3K/AKT, RB and TP53) are all required for Glioblastoma, and disruption of only one or two of them may not be enough for the tumor to survive and develop.
- Interestingly however, *EGFR* mutations seem to be mutually exclusive with RB1 mutations and *TP53* mutations, while frequently co-occurring with the deletion of the *CDKN2A/CDKN2B/ARF* locus. This somewhat contradicts the previous interpretation but may be explained by the cross-talk points between the three pathways identified downstream of *AKT* activation, which added to the disruption caused by the loss of the *TP53* regulator *ARF* and *RB* regulators *CDKN2A* and *CDKN2B* may be enough to disrupt the other two pathways.
- Deletion of the *CDKN2A/CDKN2B/ARF* locus is mutually exclusive with *CDK4* amplification and *RB1* alterations. This means deletion of that locus is enough to completely disrupt the *RB* pathway, and attempting to rehabilitate the *CDKN2A* and *CDKN2B* may be circumvented by the apparition of the other two types of mutations.
- Amplification of *CDK4* is mutually exclusive with *NF1* mutations, while also occurring alongside *MDM2* amplification at a high frequency. Co-occurrence of *CDK4* and *MDM2* amplification may be simply due to their loci being close to each other.
- *TP53* mutations are mutually exclusive with *MDM2* and *MDM4* amplification as well as *CDKN2A/CDKN2B/ARF* locus deletion, which makes sense as it means the pathway can be disrupted with alterations to *TP53* alone, or to its regulators, but alterations to both are unnecessary and constitute another potential way to circumvent treatment targeting one or the other.

5.1.2.5 Glioblastoma Disease Map

Once combined, the four submaps lead to the assembly of the Glioblastoma Disease Map represented in **Figure 11**.

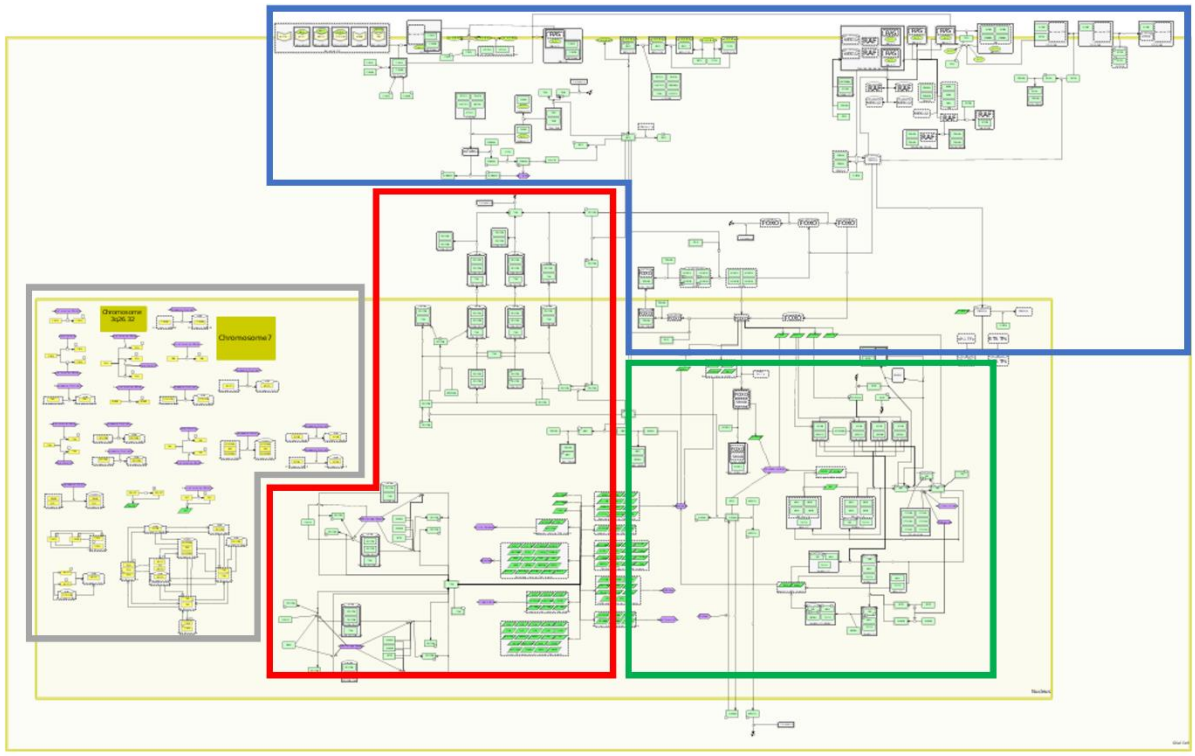


***Figure 11: Glioblastoma Disease Map assembled from the four submaps**: RTK/PI3K/AKT (blue), RB (green), TP53 (red) and Genetic Alterations (grey).*

Table 11 summarizes information about the Glioblastoma Disease Map and submaps:

	PI3K-AKT	RB	TP53	Genetic Alterations	Combined
Number of publications included	14	3	5	7	29
Number of unique entities	132	35	50	63	262
Number of modelled interactions	99	33	45	68	239

Table 11: Summary of number of publications, entities and interactions involved in the Glioblastoma Disease Map

The Glioblastoma Disease Map was uploaded to the MINERVA platform and is publicly available at

https://pathwaylab.elixir-luxembourg.org/minerva/index.xhtml?id=glioblastoma_map.

Through this overall network, several crosstalk points between the PI3K-AKT, RB and TP53 pathways can be identified. In particular, as illustrated in **Figure 12,**

- ***AKT*** plays a central role in the PI3K/AKT pathway, but also influences the TP53 pathway through phosphorylation of ***MDM2***, which then inhibits ***TP53***.
- The ***FOXO*** Transcription Factors family, inhibited by ***AKT***, can promote transcription of ***CCND1*** and ***CCND2***, inhibitors of the ***RB*** pathway, but also its activator ***CDKN1B.*** In addition, ***MDM2*** can lead to mono-ubiquitination or poly-ubiquitination of ***FOXO*** transcription factors, leading to their transit to the nucleus or degradation, respectively.
- ***TP53*** promotes transcription of ***PTEN***, which is an inhibitor of the ***PI3K*** pathway

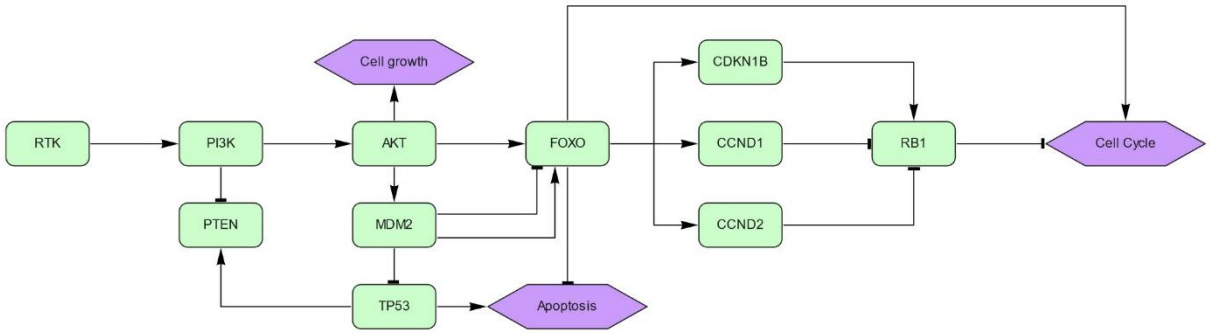


Figure 12: Overview of bridges between the PI3K/AKT, RB and TP53 pathways.

5.2 Data management methods and systems

Considerations around data management methods and solutions grew to become a significant part of work I did during the PhD. While this part did not develop directly towards addressing the question of Glioblastoma resistance mechanisms, it was nonetheless relevant to improve my awareness and understanding of limitations such considerations commonly impose on downstream analyses. This was achieved through work on three different topics: data management of clinical trial data at the Cancer Trials Ireland company; curation and storage of the GLIOTRAIN data into a tranSMART database; and migration of data from heterogeneous sources to the standardized OMOP Common Data Model.

5.2.1 Cancer Trials Ireland

At Cancer Trials Ireland, processing of clinical data from two different studies was undertaken, in order to investigate their completeness and validity.
Unfortunately, due to the confidentiality agreement between ITTM and CTI as well as the nature of the work performed there, the results, figures and tables produced during the secondment cannot be published and can only be described abstractly.

For the Glioma Biomarkers study, for which my role was to input biomarkers measurements data from scanned Case Report Forms into a database, a first version of the database was populated. In addition, the screening of the data allowed to identify a lot of mistakes and missing values in the CRFs. This outcome was not a surprise since the CTI team had already been aware and indicated that the initial data collection process for this particular study had not been up to their usual standards and they expected a lot of back and forth between themselves and the hospitals participating to the study to revise and obtain the data that could not be used from the CRFs. As a consequence, up to 60-70% of the data expected to be used in this study was processed into the database as a result of the secondment, the rest being data that could not be used or read from the CRFs and required further input from the source hospitals.

For the Breast Cancer trial, the objective of which was to produce summary statistics and reports for the trial data using the SAS software, I produced the requested summary tables characterizing the database, including demographics, statistics about certain biomarkers, number of visits, response to treatment, adverse events, among others. In addition, the code for creating these tables was generic and parametric so that it may be used for other studies in the future, akin to functions in other languages although the term is not quite accurate in the context of the SAS software. These “functions” were used in scripts to output characterization

tables and reports while also referencing any encountered issue that requires additional attention such as important information that is missing, outliers or unexpected data types in the database.

5.2.2 TranSMART and the GLIOTRAIN data

Another part of the work around data management took place in regard to the processing of the clinical, RNA-Seq raw counts, Whole Genome Sequencing (WGS) and Methylation data of the 154 GLIOTRAIN samples to load it to a tranSMART database available to the consortium. Indeed, it first needed to be formatted to format compatible with the database, which also required investigating and documenting any potential issue with the data as it was provided after sequencing, both in its format and distribution. To that end, mainly Principal Components Analysis as well as graphical visualization approaches of the data distribution were used to investigate the data before it was loaded onto the database with dedicated tranSMART loading tools.

5.2.2.1 Provided Data Characterization

Curation and management of the GLIOTRAIN data, in close collaboration with other members of the consortium, allowed me to implement data management good practices through pseudonymization of data, its characterization, documentation, and organization into a database.

To ensure data privacy, sensitive data was anonymized and a GLIOTRAIN samples labelling system was defined.

In addition, several noteworthy characteristics of the data were identified.

As demonstrated in **Figure 13**, the PCA analysis of the RNA-Seq data showed that there is a clear distinction between tumor samples and cell lines samples, which was expected but was still worth checking as it confirmed the relevance of the approach. Furthermore, a slight bias can be detected in the data. This appears to be linked to the institute that contributed each sample. Finally, two outliers clearly stand out from the rest. They were kept in the data but pointed out to the consortium members, to leave the choice to them whether to include these outliers in their own analyses and take appropriate measures.

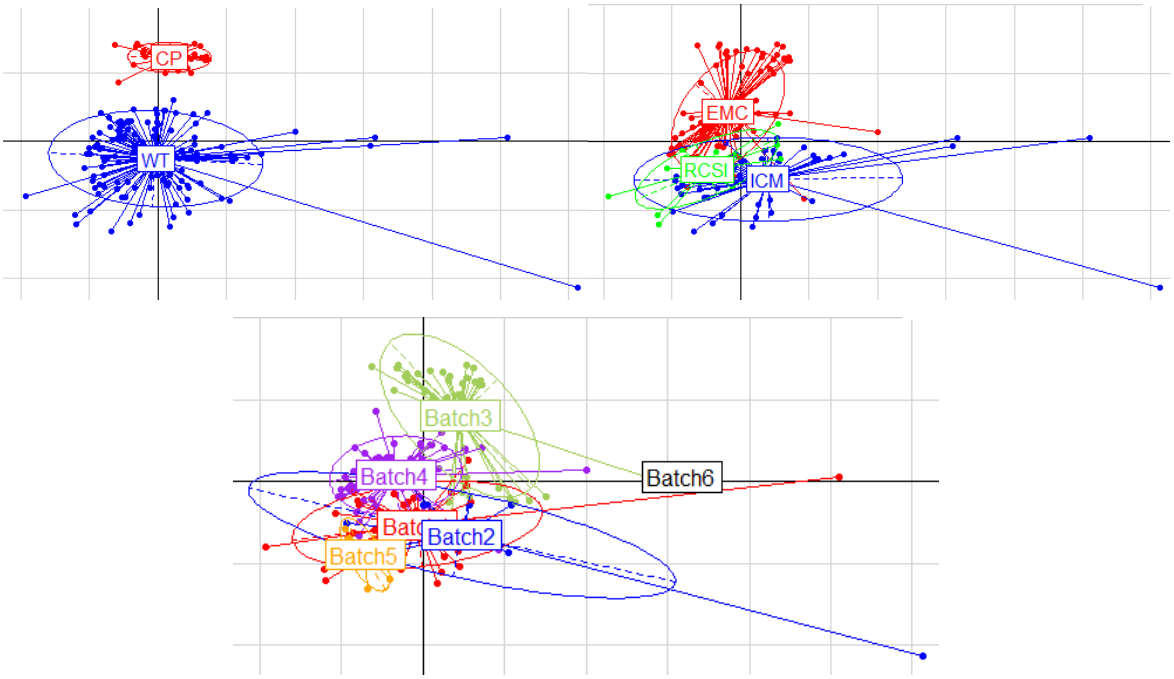


Figure 13: Projection of the 154 GLIOTRAIN RNA-Seq data samples on the first (X axis) and second (Y axis) components from a PCA*. Upper left: coloration by sample type, CP: Cell Pellets, WT: Whole Tissue. Upper right: coloration by source insitute. Bottom: coloration by batch.*

Concerning the WGS data, no effect from batch or source institute was detected. However, comparison of CNV profiles between cell lines and their corresponding parental tumour at VIB revealed that 7 out of the 26 cell lines presented a very different from the profile of their corresponding parental tumour. Indeed, in these cell cultures most of the copy number variations found in the corresponding parental tumour samples appeared to have been either lost or inverted from deletion to amplification and conversely. This issue could not be explained, and as a result these cell lines were flagged as “questionable”.

By aligning individual fragments of each chromosome for each patient, figures as displayed for Chromosome 9 as an example in **Figure 14** were obtained. From them it could be observed that many small segments were missing at various positions for any given chromosome of a patient. It was shown that this was due to the low-coverage nature of the sequencing, which does not allow for precise and extensive sequencing, leading ambiguity and dropping of poor quality segments in the VIB pipeline. But more importantly, there were also large regions that were systematically missing at the exact same positions for all samples. Discussion with the VIB team clarified that these were regions that had to be removed after sequencing due to their

low count number and poor mappability, and in most cases these regions appear to coincide with centromeres although sometimes much broader, as demonstrated by ***Table* 12**.

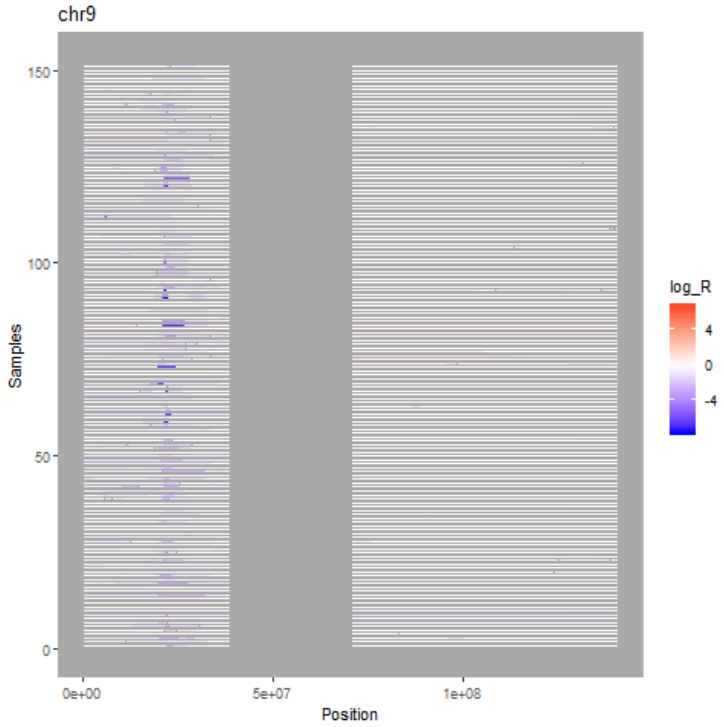


***Figure 14: Example of the WGS coverage visualization with chromosome 9 for the 151 WGS data samples**. Each line on top of the grey background represents an individual fragment of the chromosome for a given sample. X axis represents the position on the chromosome. Samples are piled on top of each other along the Y axis. Color of the fragment represents the amplitude of the focal event, both towards amplification (red) or deletion (blue).*

In the focal events dataset, two of the identified focal events, on cytobands 18q11.1 and 6q12, included one of the systematically missing regions of the chromosome as described above. However, since these cytobands are much larger than the missing regions, it is conceivable that focal events may be detected despite these missing regions, although this is a piece of information that should be kept in mind when using this dataset in an analysis, especially if the results involve these two cytobands.

Finally, it was noted that there were only two batches for the WGS data, while there were four for the RNA-Seq data. This was explained by the fact that the sequencers were able to handle up to 240 samples at a time for low-coverage WGS against 64 for RNA-Seq sequencing.

Chromosome	Missing region		Corresponding centromere[136]	
	start	end	start	end
chr1	12 825 000	13 825 000		
chr1	121 325 000	145 425 000	122 026 460	125 184 587
chr2	89 025 000	95 625 000	92 188 146	94 090 557
chr3	90 175 000	93 575 000	90 772 459	93 655 574
chr4	49 025 000	52 725 000	49 708 101	51 743 951
chr5	45 875 000	49 625 000	46 485 901	50 059 807
chr5	68 775 000	70 775 000		
chr6	57 675 000	62 025 000	58 553 889	59 829 934
chr7	57 625 000	63 425 000	58 169 654	60 828 234
chr8	43 275 000	47 625 000	44 033 745	45 877 265
chr9	38 775 000	71 025 000	43 236 168	45 518 558
chr10	38 425 000	42 875 000	39 686 683	41 593 521
chr11	50 175 000	55 075 000	51 078 349	54 425 074
chr12	34 275 000	38 475 000	34 769 408	37 185 252
chr16	35 075 000	46 575 000	36 311 159	38 280 682
chr17	22 125 000	25 325 000	22 813 680	26 885 980
chr18	14 775 000	18 575 000	15 460 900	20 861 206
chr19	24 325 000	28 375 000	24 498 981	27 190 874
chr20	26 125 000	29 875 000	26 436 233	30 038 348
chrX	58 275 000	62 075 000	58 605 580	62 412 542
chrX	154 913 804	154 925 000		

***Table 12: Position of systematically missing WGS data** and comparison with centromeres position.*

5.2.2.2 Extract, Transform, Load Process and Documentation

The results of these exploratory analyses brought light to some potential issues in the data, such as the source institute bias or the outliers in the RNA-Seq data, and the questionable cell lines. Despite that, it was decided to limit modifications to the data and provide it as close as possible to how it came out of the sequencing pipeline, to grant the consortium members the freedom to handle these issues as they see fit, which would not be possible if data was provided normalized or further transformed.

Instead, these observations were extensively documented and communicated to ensure that consortium members would be aware of them.

The data was then loaded to the tranSMART database with annotations allowing for an intuitive and easy to navigate tree structure in the graphical interface. With no issue identified in the database following upload, access to the database was provided to the consortium.

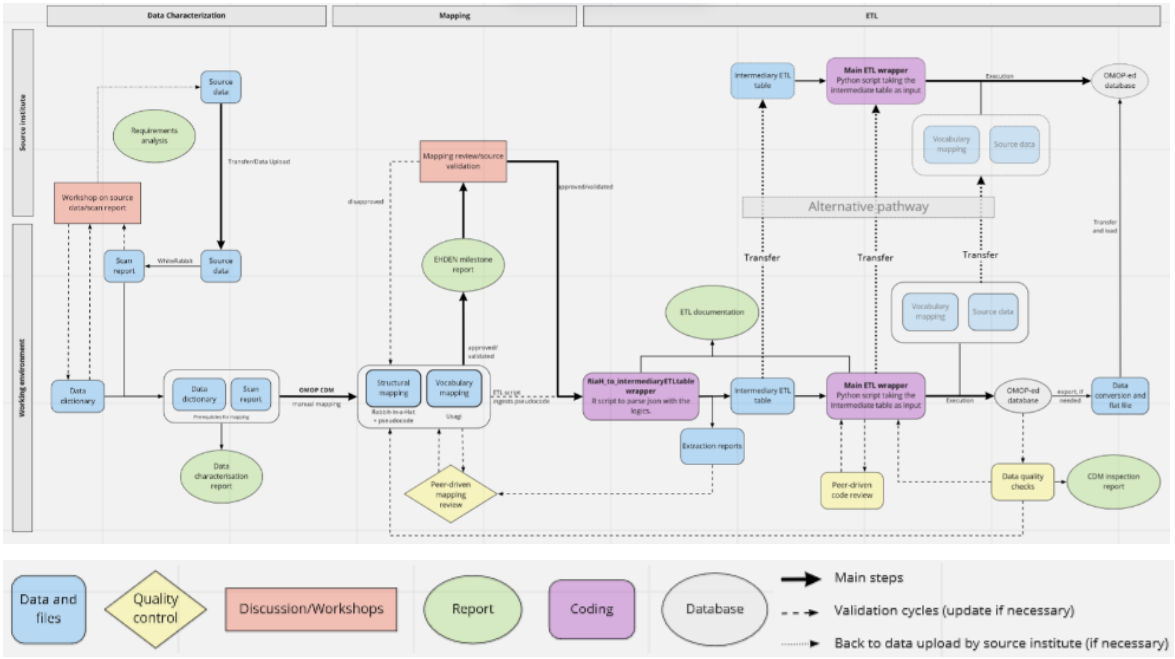
5.2.3 The OMOP CDM

As part of ITTM and in order to improve efficiency of mapping source datasets to the OMOP CDM format, I was involved in the definition of processes and tools to facilitate these activities. This work involved mainly formal development of a pipeline and software around OMOP mapping rather than active manipulation of data, but did require accounting for and documenting good practices for data management.

5.2.3.1 SOP writing

As a consequence of that reflection, I had a major role in defining the following pipeline for mapping project, which is represented by the ***Figure* 15** diagram.

Figure 15: Pipeline of the OMOP mapping process (top) and corresponding legend (bottom)



Below are general descriptions of the different processes represented in this pipeline, as well as others not directly involved in mapping but required for ITTM OMOP projects, and which solutions I found to improve and formalize them.

Requirements Analysis: The earliest stage of the project should focus on exchanges to understand exactly what is expected for this project in terms of activities (mapping, evaluation of mapping, ETL implementation…) and timelines, identify the main interlocutors both on the ITTM and source side, and get a broad overview of the data to map (size, storage infrastructure, terminology used…). For this a form to submit to customer to orient discussion was designed, to clearly express the information needed and obtain it faster.

Data Characterization: The first phase of the actual mapping pipeline consists in exploring and getting familiar with the source data to map. It requires extensive communication with someone from the source institute who knows the data well. It was important to define this phase apart from the downstream mapping activities, since it was realized from experience that starting to define mappings before having a clear understanding of the data often leads to misunderstandings, repeated modifications of mappings and loss of time which could be avoided by knowing the data well enough and having a general idea of how the mapping should be defined before even starting. To ensure this is carried out properly, in addition to the production of a Data Dictionary, *i.e.* a list of all fields in the source database with their description and possible values, which was already good practice at ITTM, the suggestion was put forward that an extensive report about the knowledge gathered during this step should be written and both should be validated by the source institute to confirm the data has been well characterized. A template of that report with guidelines to write it was also created.

Structural Mapping: Consists in defining where the data from the source database fits in the OMOP CDM, and whether any transformation (*e.g.* calculation of the value in the OMOP field from two different fields in the source data; changing the date format; using the value from field *X* or *Y* depending on the value in field *Z*…) is required to achieve it. This activity, carried out using the Rabbit-in-a-Hat software developed by the OHDSI community, is extremely time-consuming, however due to the nature of the work which is to define how the source data fits in the OMOP CDM and thus is highly specific to the structure of the source database, a solution for optimization of this step was not identified. Somewhat to the contrary even, through the definition and integration of *Machine-readable Syntax*, presented at length in the corresponding subsection below, the step of defining mappings in that syntax was introduced into this activity in addition to verbose descriptions. While the writing of both formats sensibly increases the workload needed during this process, both are needed since on one hand the verbose description allows for a clear, unambiguous and straightforward explanation of what is needed to transform the data from its source structure to the OMOP CDM, and on the other hand the mapping code syntax greatly reduces the workload needed to implement the ETL program that executes the mapping (see paragraph below), hence largely compensating for the work invested.

In addition, a script able to merge two different mapping files was implemented, so that several people may work parallelly to define mappings for different source fields and have their mappings combined once completed.

Semantic Mapping: The objective in this task is to map non-numeric values (*e.g.* free text, categories, ontology codes…) from the source data to the OMOP Standardized Vocabularies. This is done mainly through the use of the Usagi software and Athena search engine, both being resources developed by the OHDSI community, and here as well due to the nature of the work being completely dependent on the source data, I was not able to suggest efficiency-improving methods, except maybe for defining guidelines to prioritize the methods to use for the mapping:

a. if the source data uses codes from a standard ontology that is available in the OMOP Standardized Vocabularies (*e.g.* SNOMED, ICD-10, LOINC, RxNorm, MeSH), automatically generate the mappings by querying them directly from the Standardized Vocabularies
b. compile codes that are not part of available standard ontologies into a file input for the Usagi software and define mappings in Usagi
c. in cases when Usagi suggestions and search features are not enough to find satisfying mappings, use the Athena search engine for an advanced search
d. if still no appropriate mapping is found, contact the source institute to get insight on the code and suggestions on alternative terminologies which could help find better results
e. as last resort, map to the OMOP concept_id '0', corresponding to "No matching concept"

The output of this mapping process should be a CSV file containing the mapping between the source codes and OMOP concept_ids

ETL Writing: Once both Structural and Semantic mappings have been defined, a program implementing the execution of these mappings for all the data in the database is needed. The OHDSI community does not provide many resources for this step, and suggests that it should be entrusted to someone competent in ETL implementation, without more details[74]. This suggests they expect each ETL program to be tailored specifically to the source data. However, we developed a Python program with modularized features that enable re-using code with a minimal workload to adapt it to each project. This program is described in more details in the *ETL Software* subsection below. In addition to that program, and again for the sake of transparency and documentation, the writing of an ETL code report was suggested for integration into this process, to describe the code structure, dependencies and specificities of the program and provide an overview of it to the source institute. For that purpose, a template of that report was created, which would require only a few modifications since the program remains mostly unchanged across projects.

Peer Review: In order to ensure quality of the work, all project-specific mappings, ETL code and deliverables should be double-checked by a second person, and discussed until consensus is reached if the second opinion disagrees with what was reviewed. To smoothly integrate that process within the general pipeline, in particular for Structural and Semantic mappings which may take a long time before reaching completion, small bits of work, such as a group of about 10 field mappings or a section of a report, should be submitted for review once completed, so that review and subsequent discussion may take place in parallel to the progress of other parts of the work.

Timelines and Progress Tracking: Since the goal is to efficiently progress through the mapping processes, it was important to define a way to estimate timelines and keep track of the tasks that were completed or pending. To that end, two complementary approaches were defined:

- From the data dictionary created during the ***Data Characterization*** step a "mapping master file" is created, where all fields that require ***Structural Mapping*** and all values that require ***Semantic Mapping*** are listed. These lists have two purposes: firstly, estimate the total time both activities would take to achieve, since it was estimated from experience that on average it takes 30 minutes to complete structural mapping for one source field, and 10 minutes to complete semantic mapping of one source code (accounting for both the definition of the mapping and its review), so that timelines may be defined; secondly, in these lists next to each field name or source code, the status of the mapping is indicated and updated as needed with annotations "pending", "in progress", "for review" or "validated". A template was provided as well for this mapping master file with prepared columns which only need to be filled with project-specific data, and the automatic calculation of activities duration once the lists are populated. This particular attention to Structural and Semantic Mapping steps is important because they are the most time- and resource-consuming activities out of the entire pipeline.
- Generally speaking, OMOP projects should still fall into broader project management policy of ITTM, and thus be managed through the JIRA tool used at ITTM. But here again, in order to formalize and facilitate this process, and thanks to my work on defining all these activities that constitute and OMOP project pipeline, these processes were broken down from large categories like Data Characterization or Semantic Mapping all the way down to the level of elementary tasks, and as a result a list of all tasks expected

to be needed in JIRA was created, with guidelines on how to handle them in ambiguous cases.

Project Documentation: As mentioned above, several reports should be written over the course of the project. These reports were designed so that there would be extensive documentation about any given project, with two objectives in mind: transparency with the data owner, so that they would have a clear understanding of the state of the project and what was achieved and done to their data; documentation for any newcomer to the project so that onboarding them may be relatively straightforward and avoid any important information being omitted during the process. As mentioned, for each of these reports a template was created with guidelines for the sections that need to be filled.
In addition, for internal documentation templates of Confluence pages for a given project were also prepared. Rather than repeating information from the different reports and mapping files, these pages are for internal use and should contain general information about the project, expected timelines, primary contacts, location of the different mapping and report files in the ITTM infrastructure, and provide an overview of the status of the main tasks in JIRA.

From the definition and characterization of all these activities in the OMOP pipeline and subsequent tasks, each of the steps of this pipeline were described in detail as internal SOPs for the company.

5.2.3.2 ETL software

Through my work on the ITTM ETL software, the program was further developed and optimized, and possible improvements have been identified.

Refactoring led to cleaner, more readable and well-organized code. Replacing repeating code snippets with function calls and increasing the number of descriptive comments made the software easier to navigate and understand. Furthermore, two new modules were added to centralize functions dispersed in the code but related either to semantic mapping for the first module, or to the reporting of standard output, warning or error messages for the other.

Beyond refactoring the code, the modules were further developed to add new features, such as compiling and exporting ETL execution report data at the end of the program or import mapping definitions directly from the Rabbit-in-a-Hat mapping file for instance. In addition, a template was created for the project-specific code, which centralizes all instructions specific to a project such as the path to input files or mapping operations too complex to be represented with the machine-readable mapping syntax that require code implementation.

Finally, while the program was significantly improved from the work done, features that could be further optimized or added have also been identified. In particular, it was found that formula evaluation to interpret machine-readable mapping code was the most resource-consuming process during execution, and its optimization would greatly benefit performance.

5.2.3.3 Machine-readable mapping syntax

In order to optimize transition from OMOP mapping definitions to their implementation, I contributed heavily to the designing and implementation of a machine-readable syntax for defining mapping.

The formalization process of common mapping instructions led to the definition of several structures.

Firstly, the architecture of the mapping definition pipeline needed to be accounted for. In particular, the Rabbit-in-a-Hat software which is the tool used to define mappings is structured in such a way that information relevant to a given mapping may be scattered over multiple places. As a consequence, a given mapping should have a starting point in which the position of all pieces of information relevant to the mapping are provided by the mapper. Such starting points and guidelines to use them were defined and documented.

In addition, like in most programming languages assigning a value to a particular variable, *i.e.* to a field in an OMOP table, was handled with the straightforward “field = value”.

Finally, since scenarios where alternative mappings are needed depending on the data itself regularly happen in mapping projects, a structure similar to “IF” statements and blocks in programming languages with the subsequent alternative mappings was also devised.

Furthermore, 17 frequently used mapping operations were identified and assimilated to calling a function with a name explicit of its use.

Following definition of the machine-readable mapping syntax the corresponding code to interpret it was implemented and integrated to the ETL software, and both are currently used in ITTM OMOP mapping projects.

5.3 Data Analysis

5.3.1 Identification of predictive biomarkers of drug response

Research conducted for identification of predictive biomarkers of drug response was done in close collaboration with another GLIOTRAIN ESR at the Erasmus Medical Center. The goal for this work, which involved two different studies, was mainly to perform quantitative analyses of gene expression data in relation to drug response measurements to investigate genes that could potentially help predict effectiveness of drugs on Glioblastoma.

5.3.1.1 AUC computation

In order to perform these quantitative analyses, the need to compute the response variable from the raw data measurement of cell culture survival rate at different drug concentrations emerged. For this, it was decided to computationally fit curves to that survival data and use the Area Under the Curve (AUC) as response variable for the analyses. IC50 may also be derived from the curve. Indeed AUC and IC50 are commonly used as response variables for such quantitative analyses, and since the range of concentrations used for some of the drugs required extreme extrapolation for the IC50, using AUC which did not rely on such approximation appeared to be more reliable.

To facilitate and automatize curve-fitting and AUC computation for the predictive biomarkers of drug response analyses, a bioinformatics pipeline was developed which, for a given screened drug:

- takes as input a matrix of the survival rates of cell cultures (including duplicates) after exposure to different concentrations of the drug
- fits either a log-logistic, linear or exponential decay model to the datapoints of each sample, depending on which fitted model presents the smallest residuals
- produces several plot files where data is aggregated per cell culture:
    - the raw data lines
    - the raw data for “excluded samples”, for which curve-fitting was not successful
    - the smooth fitted curves and corresponding selected model type
    - optionally, one individual graph per cell culture where the parameters of the model are displayed directly on the plot
- exports as text files:
    - the smooth fitted curves data points for plotting
    - the parameters for the fitted models

    - the list of excluded samples
- *.RData* files that contain
    - the fitted models
    - the smooth fitted curves data points for plotting

Once the program was implemented, it was used to calculate AUCs and IC50s for both the drugs repurposing and cell culture models validation projects.

Furthermore, the program was also used for a different project I did not work directly on, which was about validating IDH-mutant Glioblastoma cell cultures models. This led to me signing as a co-author for a publication[137] on that study.

5.3.1.2 Drugs Repurposing Project

The goal of the Drugs Repurposing Project was to analyze microarray transcriptomics data of treatment-naïve cell cultures in comparison with response data of same cell cultures to 110 drugs, in order to identify genes for which the expression profile may be predictive of sensitivity to a given drug, as well as use the results to select the most interesting drugs candidates for treating Glioblastoma. First, a PCA was performed on the data to investigate its distribution, in particular in regard to external factors such as sequencing batch and type of sample. Then, data was analyzed both with Least Absolute Shrinkage and Selection Operator (LASSO) and Weighted Gene Co-expression Analysis (WGCNA) methods, in order to cover the analysis with different methodologies (*i.e.* regression and clustering). Furthermore, as other factors including the type of sample (Primary and Recurrent Glioblastoma) and the set of genes included in the analysis (cancer-focused or indifferent), several analyses were conducted with the different combinations of these parameters.

The analysis pipeline started with a characterization of the available data. Initial investigation of the DASL data, looking at the distribution of the mean and variance of the different probesets did not reveal any significant abnormality or outlier. A PCA also revealed no strong bias associated with batches as shown in **Figure 16**.

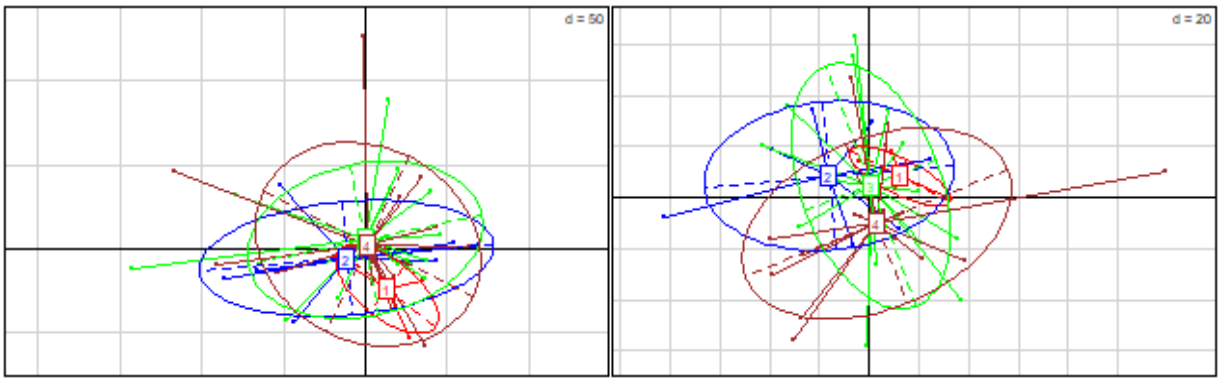


Figure 16: Samples from the DASL data projected on the first (X axis) and second (Y axis) Principal Components (left), and third (X axis) and fourth (Y axis) Principal Components (right) from a PCA. *Coloring by batch.*

The PCA also revealed no strong dissociation between Primary and Recurrent Glioblastoma samples, as illustrated by **Figure 17**, which is concerning since there are known differences between these types, and therefore there should be a noticeable difference in their expression profile.

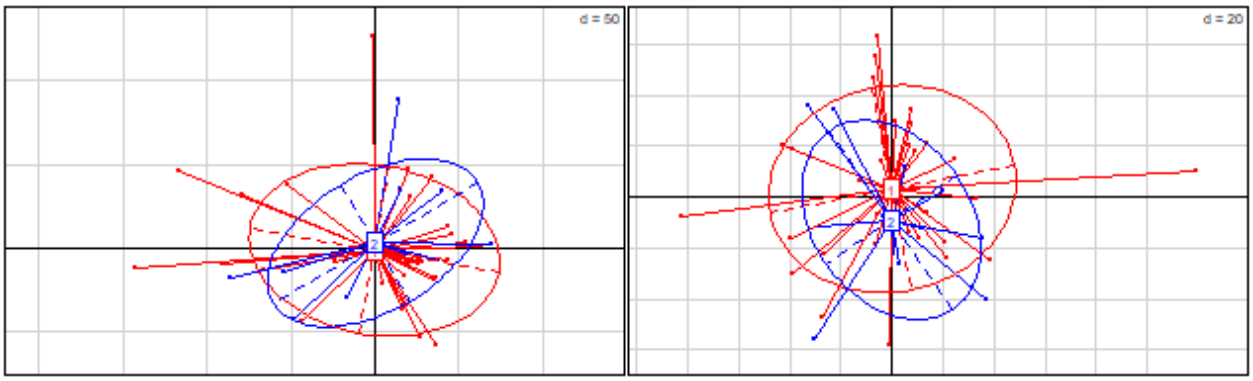


Figure 17: Primary (in red) and Recurrent (in blue) Glioblastoma samples from the DASL data projected on the first (X axis) and second (Y axis) Principal Components (left), and third (X axis) and fourth (Y axis) Principal Components (right) Principal Components from a PCA

In the first run of the analysis using the initial IC50s dataset, both the LASSO and WGCNA approaches produced interesting results.

Using LASSO, there were only a few, if any, genes associated to each drug. Although these genes were rarely directly connected, there were many cases where they would be distant by only one or degrees neighbours. Growing the networks by adding such neighbours also allowed for the emergence of functional pathways, and as a result I identified 3 drugs for which Glioblastoma-relevant pathways were associated with the LASSO-selected genes when including all Glioblastoma samples in the LASSO analysis, and two drugs had such a pattern when only Primary Glioblastoma samples were used. These results are described in ***Table*** **13** and ***Table*** **14**.

Table 13: Glioblastoma-relevant results from the initial run of the LASSO analysis including all Glioblastoma samples

Drug	LASSO-selected genes	Emerging Pathways
Pemetrexed	*HLA.DRB5, HSD17B1, ID2, OR52E8, PIGN, PVALB, RASSF5, RBM44, SLC26A11, STK17B, TAS2R43, TBC1D3B, TRPA1, WFIKKN2, WNT3, ZNF229, ZNF28*	Protein Kinase A Signaling, Sirtuin Signaling
Pralatrexate	*ADCYAP1, ARG2, CHRDL2, CNGA1, CSAG1, DOK6, ENTPD3, FCGR2B, FNDC9.1, GJD3, GPC3, HIST3H2A, HOPX, HPDL, HSPB3, IL1RN, KCNS3, LINGO2, LRRC26, MASTL, MCM3AP.AS1, N4BP3, NODAL, PLAC9, PRSS36, SLC6A12, XRN2*	Molecular Mechanisms of Cancer, Axonal Guidance Signaling, Glioblastoma Signaling
Sorafenib	*BMP6, CNGA1, FAM81A, GTF2H2B, GUCY1A1, KCNH7, KCNIP2, KCNQ3, MADCAM1, OLFM4, OMD, P2RY14, PCDH8, PCDHGB4, RPL23AP64, SMA5, ST18, UNC5C, ZFPM2*	Glucocorticoid Receptor Signaling, Neuroinflammation Signaling

Table 14: Glioblastoma-relevant results from the initial run of the LASSO analysis including Primary Glioblastoma samples

Drug	LASSO-selected genes	Emerging Pathways
Sorafenib	*ACVR1C, ADGRG7, ADH1B, CLIP1, DCHS2, DNAJA4, EMX2OS, GJB6, GRM3, GUCY1A1, SLC7A2, TMEM233, TNNT2, UBE4A, ZFPM2*	Synaptic Long Term Depression, Glucocorticoid Receptor Signaling, Gap Junction Signaling
Capecitabine	*CCL25, CRTC1, ENOX2, GLIDR, HBG2, LARGE1, OR8G2P, PLAA, PPIL2, RNF213, SRC, TMEM104, TMEM234, TMEM39B, TSC2, ZNF117*	14-3-3-mediated Signaling, AMPK Signaling, Glioblastoma Signaling

Conversely, the WGCNA approach produced clusters of genes much larger than the LASSO. This led to different results in the IPA enrichment investigation. Indeed, the genes in any of the cluster were numerous enough that connections between them could be established already with looking for common neighbours, though enrichment of the network with first-degree shared neighbour was still performed. Despite that, there were also many of the genes in the cluster which remained unconnected to the rest. In addition, since there were many more genes, the number of associated functional pathways were also higher than for the LASSO results. It is also important to note that contrary to the LASSO approach which selects a different set of genes for each drug, the WGCNA clusters genes based on their expression profiles before correlating them with drug response. As a result, a same gene cluster can be correlated to the response of several drugs, and associated functional pathways would be mostly the same since they emerge mainly from the initial gene cluster, although there may be additional pathway that would emerge due to the addition of the drug, its targets and any intermediate component

between them and the genes of the cluster. Finally, I had to focus on the stronger cluster/drug associations, otherwise there would be too many networks to investigate. Therefore, the threshold of investigation was arbitrarily set to an absolute correlation threshold of 0.6 or higher with a p-value of 0.05 or less. This to shortlist a more manageable seven clusters to investigate than 21 clusters if the absolute correlation threshold was set 0.5, or over 40 clusters if all clusters that had a correlation p-value of at most 0.05 with at least one drug had to be investigated. However, that threshold was only reached when only the Primary samples were included in the analysis and the Recurrent samples were left out. As a result, for WGCNA I focused on that subset. ***Table*** **15** provides an overview of the results.

Despite these promising results which seemed to identify cancer and/or neuron-related genes correlated with response to several drugs, I was not satisfied with the bounded solution for the IC50s data. Following acquisition of the spectrophotometry measurements of cell cultures survival at different concentrations of drugs and implementation and execution of the AUC computation, evaluation of the output of the AUC computation process for the drugs revealed several noteworthy results.

First of all, there were 13 drugs which clearly seemed to have no effect of Glioblastoma cell cultures. As illustrated by **Figure 18**, the cell survival rates did not vary with drug concentrations, and most cell cultures presented similar response profiles. This concerned the drugs Allopurinol, Altretamine, Anastrozole, Cabazitaxel, Capecitabine, Cyclophosphamide, Docetaxel, Ifosfamide, Lenalidomide, Letrozole, Pentostatin, Pomalidomide and Thalidomide, which as a consequence were removed from the rest of the analysis.

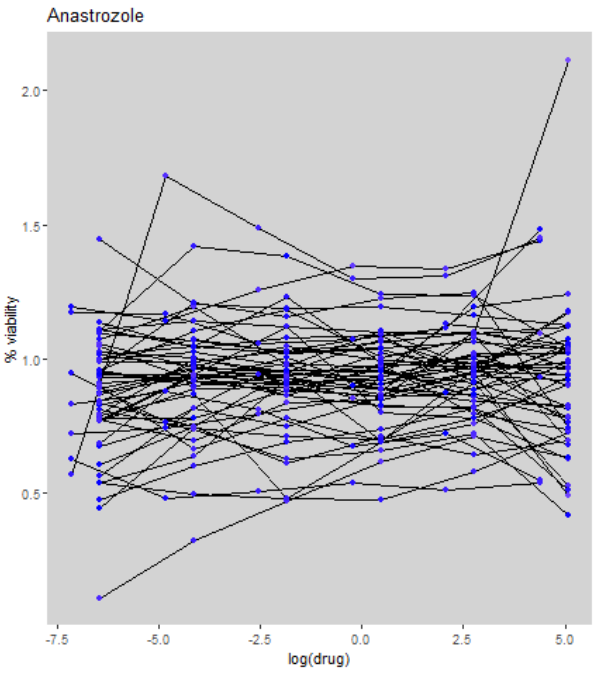


Figure 18: Drugs such as Anastrozole which did not seem to affect cell cultures viability were removed from the analysis. *Blue dots correspond to the percentage of cell population survival at the given drug concentration, averaged across all replicates for a given cell culture. Black lines represent evolution of the cell culture response at different concentrations by linking the blue dots of a given cell culture.*

Table 15: Clusters highly correlated with drugs (absolute coefficient of 0.6 or more, p-value of 0.05 or less) and Glioblastoma-relevant functional pathways linked to cluster genes involved with that correlation. *From the initial run of the WGCNA analysis including Primary Glioblastoma samples. For Cluster 1 and 2, several pathways were recurrently associated with the correlated drugs, and were thus listed along with the cluster name as "Recurrent pathways", which were referred to in the "Corresponding Canonical Pathways" column.*

Cluster and Recurrent Pathways	Correlated Drugs (number of genes)	Corresponding Canonical Pathways
Cluster 1 Recurrent pathways: • Axonal Guidance Signaling • Hepatic Fibrosis / Hepatic stellate Cell Activation • GP6 Signaling • Cardiac Hypertrophy Signaling • Role of Osteoblasts, osteoclasts and chondrocytes in Rheumatoid Arthritis • Osteoarthritis Pathway • Atherosclerosis Pathway	Enzalutamide (624)	Acute Phase Response, Glucocorticoid Receptor Signaling, IL-6 Signaling, Axonal Guidance, Glioma Invasiveness
	Bleomycin sulfate (666)	Recurrent Pathways, Glioma Invasiveness, ERK/MAPK Signaling
	Busulfan (728)	Axonal Guidance, Sirtuin Signaling
	Carboplatin (809)	Recurrent pathways
	Idarubicin hydrochloride (214)	Recurrent pathways, Molecular Mechanisms of Cancer
	Melphalan hydrochloride (666)	Recurrent pathways
Cluster 2 Recurrent pathways: • Synaptogenesis Signaling • Axonal Guidance Signaling • Opioid Signaling • Neuroinflammation Signaling • Huntington's Disease Signaling • Protein Kinase A Signaling • GABA Receptor Signaling • Synaptic Long-term Depression • CREB Signaling	Bendamustine hydrochloride (249)	Recurrent Pathways
	Oxaliplatin (293)	Recurrent Pathways
	Melphalan hydrochloride (288)	Recurrent Pathways
	Mercaptopurine (219)	Recurrent Pathways
	Lomustine (225)	Recurrent Pathways
Cluster 3	Temozolomide (67)	Hepatic Fibrosis / Hepatic stellate Cell Activation, mTOR Signaling
Cluster 4	Procarbazine hydrochloride (83)	Estrogen Receptor Signaling
Cluster 5	Omacetaxine mepesuccinate (48)	Protein Kinase A Signaling, Sirtuin Signaling
Cluster 6	Carboplatin (121)	Axonal Guidance Signaling, Osteoarthritis Pathway, Protein Kinase

		A Signaling, Synaptogenesis Signaling, Molecular Mechanisms of Cancer
Cluster 7	Afatinib (22)	Synaptogenesis Signaling, Huntington's Disease Signaling, Apoptosis Signaling

Moreover, the tested range of concentrations was likely not adapted for several of the screened drugs. Indeed, there were drugs for which the log-logistic shape of most the curves fitted with that model was clearly not captured within the tested range of concentrations, either because inflexion

- seems to happen before the lowest concentration tested (Carfilzomib, Dactinomycin, Gemcitabine, Ixabepilone, Romidepsin)
- reaches the lower plateau beyond the highest tested concentration (Amirolevilinic acid hydrochloride, Bendamustine hydrochloride, Busulfan, Carmustine, Dabrafenib mesylate, Dacarbazine, Floxuridine, Methoxsalen, Mercaptopurine, Procarbazine hydrochloride, Temozolomide, Vismodegib)
- takes place over a broader range on both sides of tested concentrations, with both early and late part of the log-logistic curve visibly occurring outside of the tested concentrations (Decitabine, Pemetrexed)

Although the model could still be fitted, the fact that this was the tendency for most of the fitted curves raised the potential issue that the model inference may be off for those drugs. In addition to that, there are also cases where the sigmoid shape is not even observable and leads to fitting either the linear model, when the data is monotonously decreasing, or the exponential decay model when the drug response data is likely reflecting the end part of a sigmoid shape. Furthermore, there were drugs for which there was a high number (>25%) of curves fitted to either the linear or exponential decay models rather than the expected log-logistic model for drug response. In some cases, it is likely due to a wrong tested range of concentration as described above, while in others it may be due to

- high sensitivity of the cell cultures to the drug for cases where exponential decay model was fitted for a few cultures while the others were fitted to the log-logistic model (Dasatinib, Temsirolimus and Topotecan hydrochloride)
- low sensitivity of the cell cultures to the drug for cases where a linear model with almost no slope was fitted for a few cultures while the others were fitted to the log-logistic model (Amirolevilinic acid hydrochloride, Arsenic trioxide, Bendamustine hydrochloride, Busulfan, Dexrazoxane, Temozolomide)

- unexplained response of the cell cultures to the drug or experimental errors (Floxuridine, Fluorouracil)

Finally, there were drugs for which a relatively important number (ten or more) of cell cultures were excluded from the final set of models, either because none of the models could fit the data, the fitted model was monotonously increasing (suggesting higher survival of cells with higher concentrations of the drug) instead of decreasing, or because the computed higher limit of the model, corresponding to the estimated percentage of cell population survival at null concentrations of the drug, was three times higher than the value for the control. These issues may be due to the drug not affecting the cell cultures at all, or come from inadequate concentrations testing ranges or poor data quality. While it would have been good if new experiments to produce new data for these drugs could have been performed, it was not possible to do so at that stage and the drugs had to be excluded from further analysis.

Conversely, there were also many drugs for which the curve-fitting process resulted in well-defined log-logistic curves, with very few samples excluded, and inflexion points of curves quite close to each other across all cell cultures. That was the case for 19 out of the 109 drugs, namely Afatinib, Amiodarone hydrochloride, Bortezomib, Bosutinib, Cabozantinib, Celecoxib, Crizotinib, Enzalutamide, Everolimus, Exemestane, Gefitinib, Imatinib, Mitotane, Nilotinib, Omacetaxine mepesuccinate, Plicamycin, Regorafenib, Sirolimus, Sorafenib, Sunitinib, Tamoxifen citrate, Thioguanine and Vemurafenib. These results suggest that the response (and by extension AUC and IC50 derived from it) to any of these drugs is highly similar for all cell cultures, and as a result it may not be interesting to look into them since it means that response would hardly be helpful to discriminate between cell cultures and their gene expression profiles.

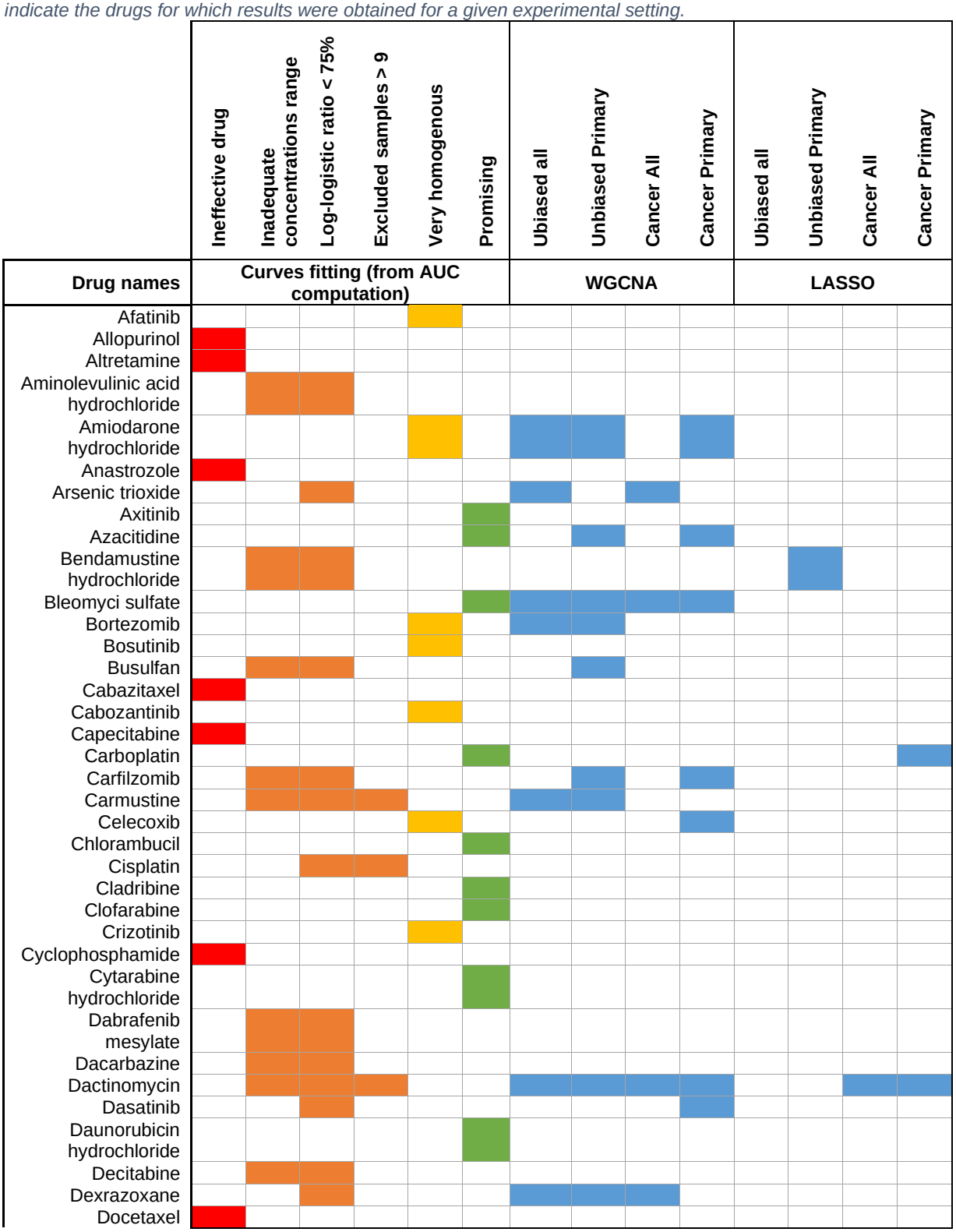
The drugs that haven’t been mentioned so far as presenting either notable issues in curves or well-defined log-logistic curves are drugs for which curve fitting was successful, but the models were not as well defined or were quite dispersed and heterogeneous, suggesting a more variable and not as consistent response to the drug. As such, these are the drugs which, at that point, show the most promise to be reliable enough to be used in an analysis while also having enough dispersion to enable identification of genes that have different expression profiles in cultures that are more responsive to a given drug than in cultures that are less responsive to that drug in a significative way. Those drugs are Axatinib, Azacitidine, Bleomycin sulfate, Carboplatin, Chlorambucil, Cladribine, Clofarabine, Cytarabine hydrochloride,

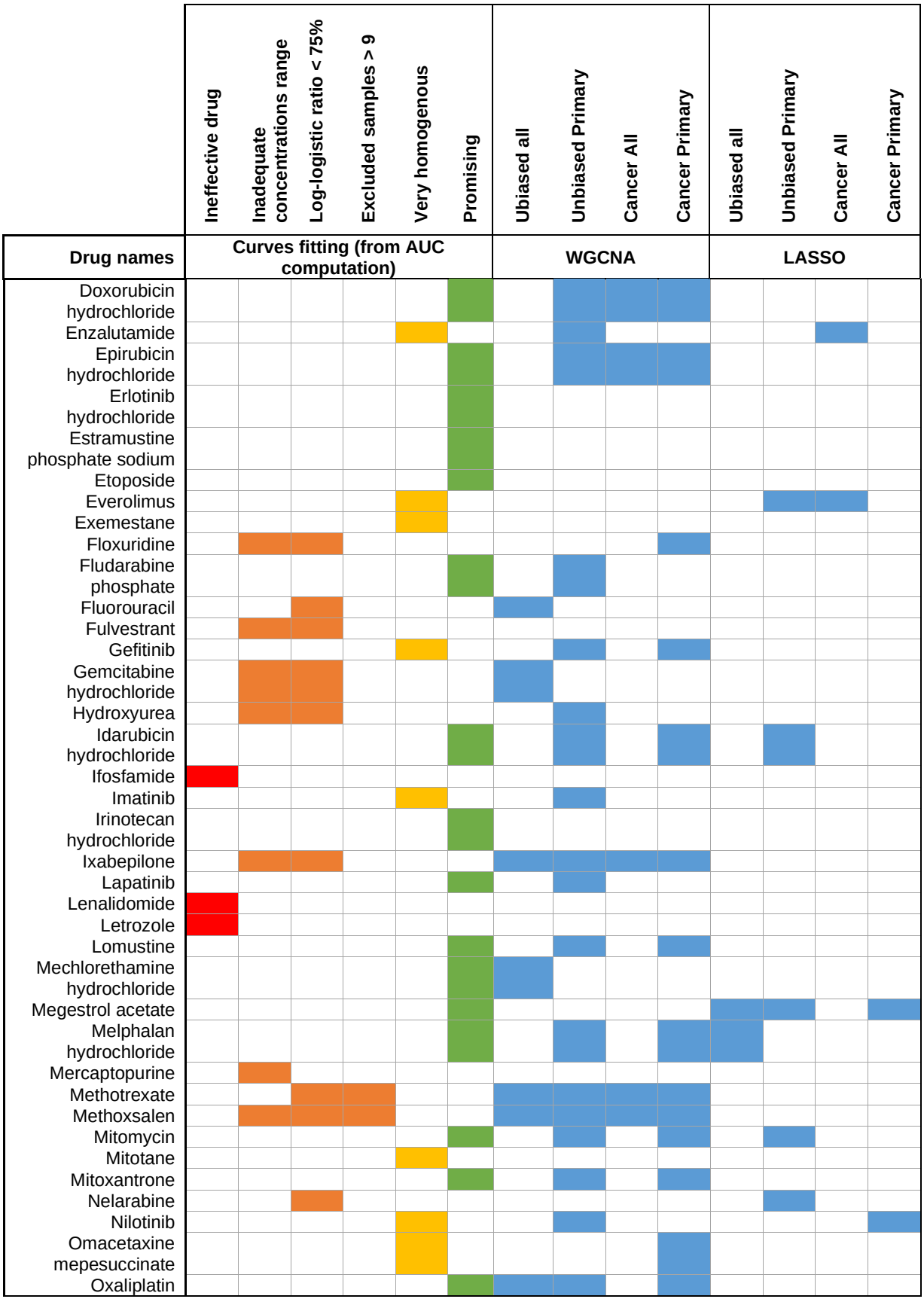
Daunorubicin hydrochloride, Doxorubicin hydrochloride, Epirubicin hydrochloride, Erlotinib hydrochloride, Estramustine phosphate sodium, Etoposide, Mitomycin, Fludarabine phosphate, Idarubicin hydrochloride, Irinotecan hydrochloride, Lapatinib, Lomustine, Mechlorethamine hydrochloride, Megestrol acetate, Melphalan hydrochloride, Mitoxantrone, Oxaliplatin, Pipobroman, Pazopanib hydrochloride, Ponatinib, Raloxifene, Teniposide, Thiotepa, Tretinoin, Uracil mustard, Valrubicin, Vandetanib and Verinostat.All these findings about curve-fitting results for the different drugs can be found in the ***Table*** **16** summary.

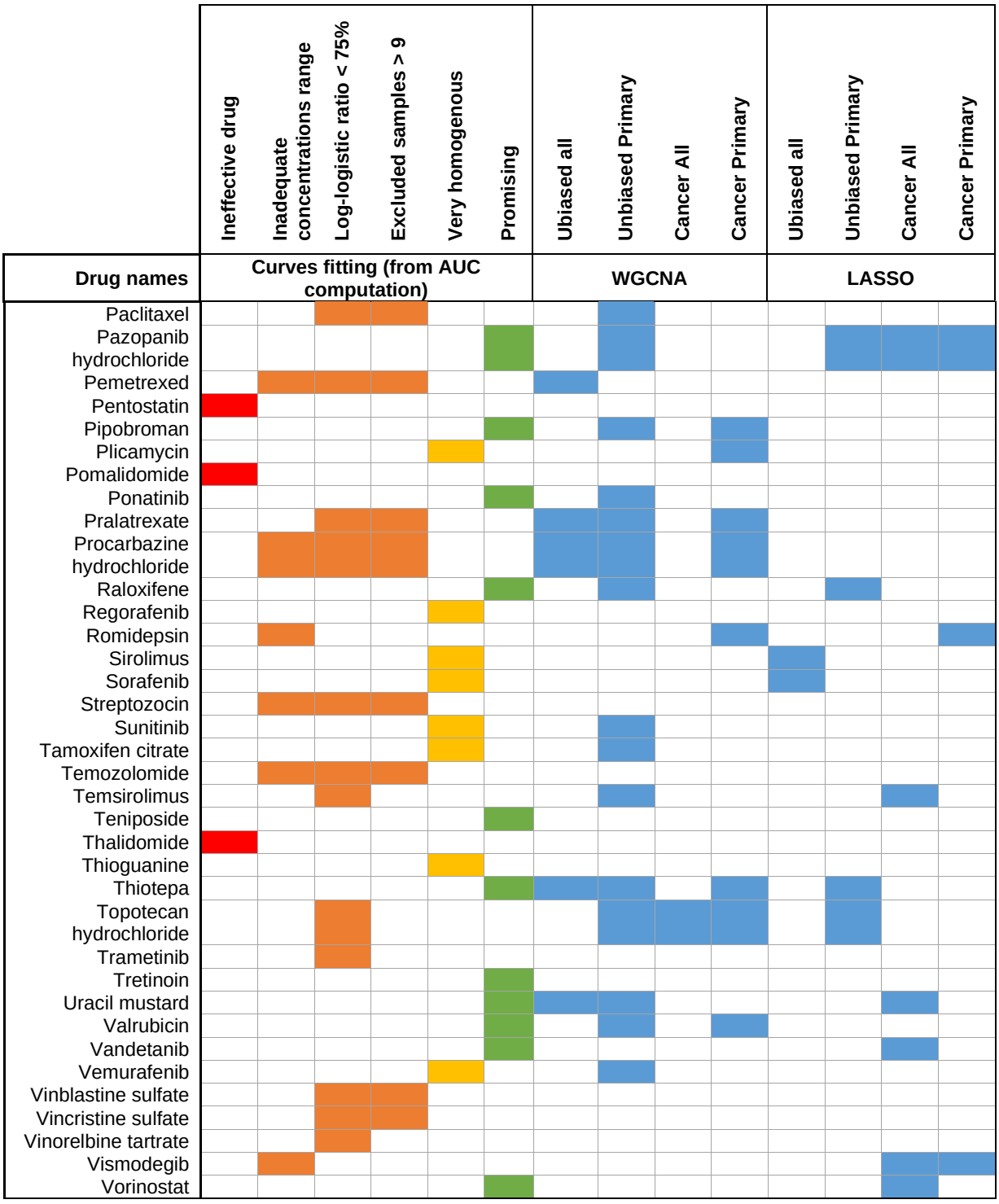
Still, all drugs except for the ones clearly unaffecting Glioblastoma cells were included in the analyses to see what would come from it, since the analysis relied on the assumption that difference in the cell cultures response to drug, captured through AUC which is influenced not by multiple other factors including slope steepness and final plateau of the curve, regardless of the shape of the curve, could be used as response variable for a multivariate analysis. While the typical log-logistic shape brings more confidence in the accuracy of the data used, we did not want to exclude any promising drug candidate on that sole criterion. To look for biomarkers for which the expression profile could be associated with drug response, the data was analyzed with both LASSO and WGCNA, each with two different subsets of samples (all Glioblastoma samples and only Primary Glioblastoma samples), and with the unbiased genes set and the cancer genes set. As a result, eight different sets of results were produced and needed to be screened, interpreted, and compared. Furthermore, the enrichment analysis method was switched from the initially used manual and time-consuming investigation in IPA relying on proprietary molecular interactions and pathways and offering limited control over network development parameters to the method described in the 4.3.1.2 Gene Ontology enrichment function subsection which provided a more reliable solution as it was automatized, provided statistical tests and p-values along with results, and relied on the Gene Ontology which is a more standard and widely used resource.. The significant enrichment analysis results for each of the eight analysis designs were compiled in **Table 22** and **Table 23** from the 8.2 Drugs Repurposing LASSO and WGCNA results annex.

Table **16** provides an overview of which approach(es) resulted in the identification of potentially interesting functional pathways, which would lead to the identification of potential predictive biomarkers, for each drug.

Table 16: Overview of the results emerging from AUC computation, LASSO and WGCNA analysis in the Drugs Repurposing Project. *Red cells indicate drugs that were excluded due to apparentlt no effect on cell cultures, orange cells indicate questionable relevance of fitted curves, yellow are for drugs for fitted curve models were very homogeneous, green cells indicate fitted models presented heterogeneous log-logistic profiles, and blue indicate the drugs for which results were obtained for a given experimental setting.*




Ineffective drug
Inadequate concentrations range
Log-logistic ratio < 75%
Excluded samples > 9
Very homogenous
Promising
Ubiased all
Unbiased Primary
Cancer All
Cancer Primary
Ubiased all
Unbiased Primary
Cancer All
Cancer Primary
Drug names
Curves fitting (from AUC computation)
WGCNA
LASSO
Doxorubicin hydrochloride
Enzalutamide
Epirubicin hydrochloride
Erlotinib hydrochloride
Estramustine phosphate sodium
Etoposide
Everolimus
Exemestane
Floxuridine
Fludarabine phosphate
Fluorouracil
Fulvestrant
Gefitinib
Gemcitabine hydrochloride
Hydroxyurea
Idarubicin hydrochloride
Ifosfamide
Imatinib
Irinotecan hydrochloride
Ixabepilone
Lapatinib
Lenalidomide
Letrozole
Lomustine
Mechlorethamine hydrochloride
Megestrol acetate
Melphalan hydrochloride
Mercaptopurine
Methotrexate
Methoxsalen
Mitomycin
Mitotane
Mitoxantrone
Nelarabine
Nilotinib
Omacetaxine mepesuccinate
Oxaliplatin



5.3.1.3 *Berkeley LASSO Analysis review*

Before the analyses described above, a different team at Berkeley University had performed a LASSO analysis and investigated functional pathways of their results using the IPA software. As a result, a comparison between their approach and was conducted in order to assess

robustness of results with similar yet unidentical approaches since exact protocol of the Berkeley analysis was not provided. The available information for this was the drugs and associated genes identified by the LASSO analysis, and which of those genes were identified as particularly interesting using the IPA software. The comparison was conducted in two stages: attempting to identify the same genes of interest from the list of genes associated with each drug in the Berkeley results; and comparing the Berkeley results with results of the Drugs Repurposing Project analyses described above.

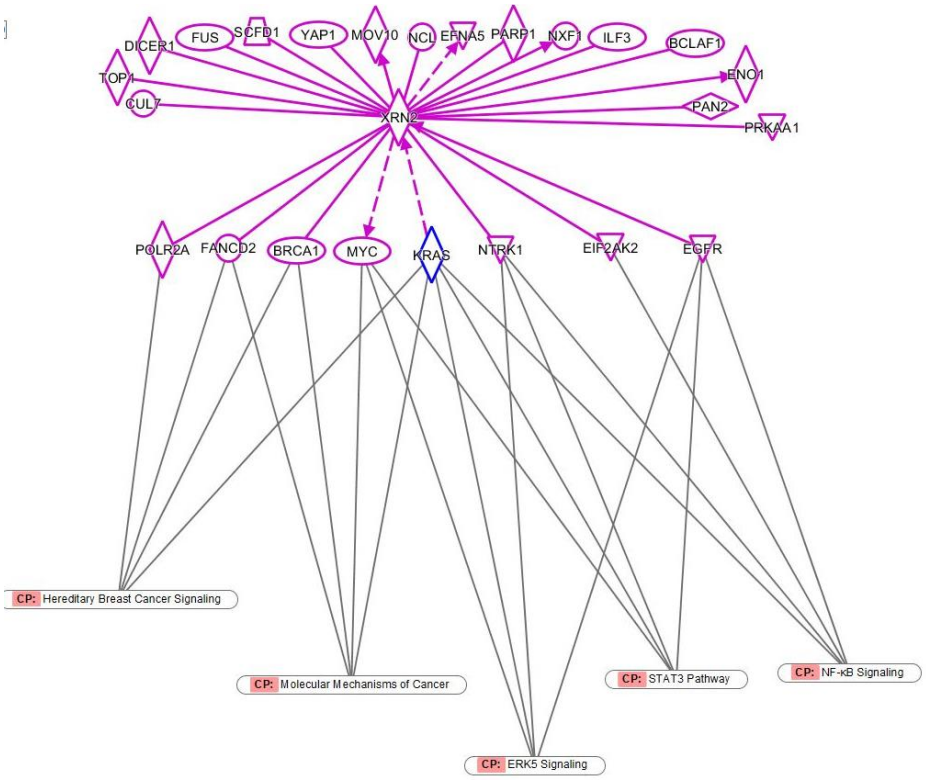
5.3.1.3.1 Investigation of the IPA enrichment analyses interpretations

The pathway enrichment analysis in IPA performed for the genes identified by the Berkeley LASSO analysis for each of the 10 shortlisted drugs led to the following results:

- For Bortezomib, Cytarabine Hydrochloride, Mitotane and Sorafenib, which all had only one or two genes associated by the LASSO analysis, no clear connection between the drugs and their respective LASSO-associated genes nor any functional characteristic emerged from the investigation.
- For Hydroxyurea as well, no clear biological function emerged from the network that formed to connect the drug and its four LASSO-associated genes, and that network itself was limited mostly to shared upstream miRNA regulators between the LASSO genes and the targets of the drug.
- Surprisingly the drug Pazopanib hydrochloride, for which the LASSO selected a large number of genes compared to the other drugs, did not output clear biological functions, as the LASSO candidates seemed very disconnected from each other and from the drug.
- Among the nine LASSO-selected genes for Dexrazoxane,
    - three (*MRM1, OLA1* and *ICAM5*) seem well connected to the drug
    - five (*RTBDN, ADGRL1, SPPL2B, ICAM5* and *SEMA6D*) are located at the cellular membrane, although they are involved in different processes (cell junctions, signal transduction, ligand binding...)
    - however, no functional pathway was found to be particularly well represented in the resulting network
- For Vemurafenib, several of the LASSO-selected genes were downstream of *BRAF* and *ARAF* which are inhibited by the drug, providing a biological explanation to these results. In addition, through the connections of *CTSG, TNNT2* and *DSP*, cellular adherens and cell-cell contact functions seem to emerge from the network, confirming the Berkeley results which had identified *CTSG and DSP* as genes of interest in their investigation.

- The analysis for Imatinib bore several noteworthy results:
    - The network generated by connecting Imatinib-associated LASSO genes with the targets of the drug was quite dense, however it was in a large part due to both sets of genes share common upstream miRNA regulators.
    - Among the connections that were not due to common miRNA regulators, several of *XRN2* direct neighbours associated with signaling pathways and cancer-related groups (see **Figure 20**), while SEPT4 direct neighbours strongly associated with neuronal pathologies-associated canonical pathways (see **Figure 19**)
    - Overall, the network displayed a high representation of Glioma/Glioblastoma associated pathways
    - These results are not completely in line with the ones from the Berkeley analysis, where *SLITRK1* was defined as gene of interest rather than *SEPT4*

Figure 20: XRN2, a LASSO-selected gene associated with Imatinib, is connected to genes belonging to signaling pathways and cancer-related pathways.



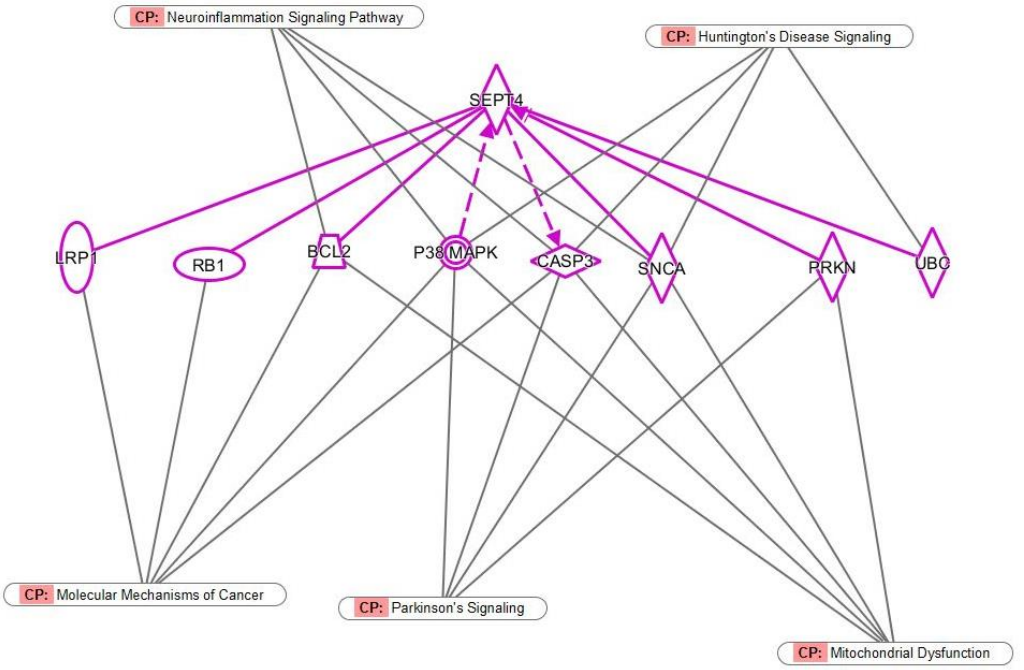


Figure 19: SEPT4, a LASSO-selected gene associated with Imatinib, is connected to genes belonging to neuronal pathologies-associated pathways.

- The Tretinoin results were particularly interesting: generating connections between the Tretinoin targets and the LASSO-selected genes added many genes to the network, yet there was a relatively short distance between the Tretinoin targets and the LASSO-selected genes, regardless of the path as illustrated by **Figure 21**. In addition, together the genes from that network revealed to be heavily involved with signaling pathways such as the *PI3K/AKT*, *14-3-3* mediated signaling, *ERBB, NF-kB*… pathways, which I encountered frequently in my research to build the Glioblastoma Disease Map. Although the Berkeley analysis did not recognize *DNAI2, SEMA3E* and *SLC39A1* as interesting following their run in IPA, overall my results are in line with theirs for this drug.

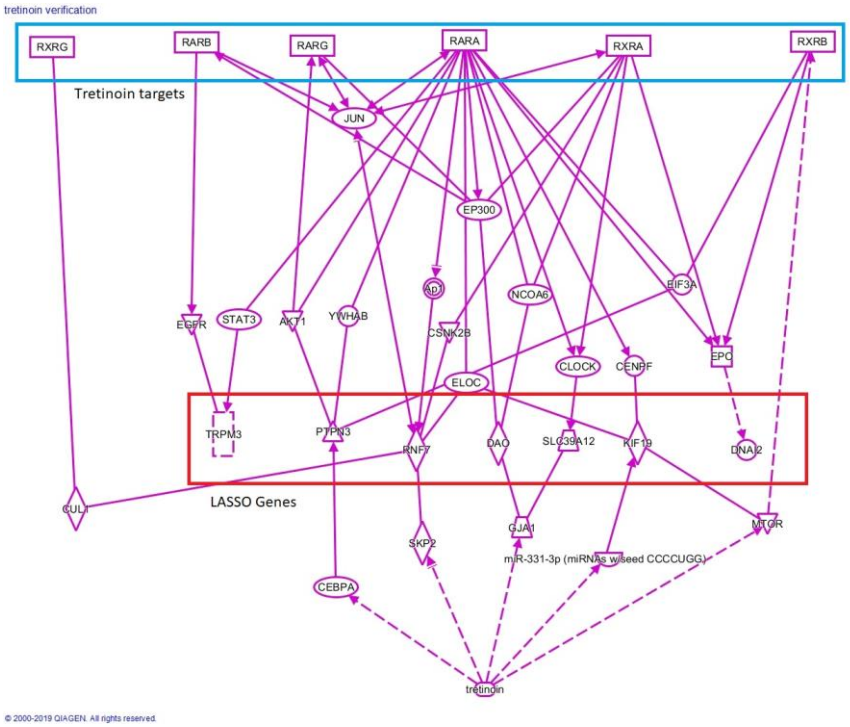


Figure 21: Network obtained by establishing links between Tretinoin targets (in the blue rectangle) and LASSO-selected genes (in the red rectangle) for Tretinoin

To conclude, this attempt at reproducing findings from IPA enrichment analysis starting from the same lists of genes ended with mixed results. Reproduction of results was achieved for only three (Tretinoin, Vemurafenib and Imatinib) out of the ten drugs investigated. This finding suggest that the limited information provided of the general steps of the procedure was not enough for robust reproduction of results, and highlighted the importance of documenting and communicating protocols. Considering that the three drugs for which results could be reproduced were also the only ones for which Glioblastoma-relevant pathways emerged, it could also be that those were the only ones for which associated genes presented biological

coherence, and completely unrelated genes for the other drugs leading to random network expansion in IPA may also be a factor in the divergence of results.

5.3.1.3.2 Comparison of computational analyses results

Out of the 36 drugs that the Berkeley LASSO analysis associated genes with, 8 came up in the results of my LASSO analyses, 18 in my WGCNA runs, and 16 could not be associated with any gene with LASSO and WGCNA analyses and were thus absent from my results. Of the 10 drugs shortlisted in the Berkeley analysis for further IPA investigation, 2 were found in my LASSO analysis, 5 in my WGCNA analyses, and 4 were absent from my analyses. **Table 17** details the overlap between the drugs highlighted by the Berkeley analysis and mine.

Table 17: Overlap between drugs identified in the Berkeley LASSO analysis and my analyses. *Drug names in red are the 10 drugs shortlisted at Berkeley for IPA investigation. Drug names in bold were identified in both my LASSO and WGCNA analysis*

	Drugs identified in the Berkeley LASSO analysis
Identified in my LASSO analyses	Bendamustine, hydrochloride, Carboplatin, **Enzalutamide**, **Melphalan hydrochloride**, **Pazopanib hydrochloride**, Sirolimus, Sorafenib, **Uracil mustard**
Identified in my WGCNA analyses	Arsenic trioxide, Azacitidine, Bleomycin sulfate, Bortezomib, Busulfan, Carmustin, Dexrazoxane, **Enzalutamide**, Fluorouracil, Hydroxyurea, Imatinib, Lomustine, **Melphalan hydrochloride**, Methotrexate, Oxaliplatin, **Pazopanib hydrochloride**, Pemetrexed, **Uracil mustard**
Absent from my results	Allopurinol, Cisplatin, Cytarabine hydrochloride, Dacarbazine, Dasatinib, Decitabine, Estramustine phosphate sodium, Erlotinib hydrochloride, Floxuridine, Fulvestrant, Mitotane, Pentostatin, Streptozocin, Trametinib, Tretinoin, Vemurafenib

Finally, out of all the genes that were associated to each of the 10 shortlisted drugs in the Berkeley LASSO analysis, the only overlap was the *OR2L13* gene which was associated to Imatinib and also came up in my results for this drug. While this does not suggest complete disagreement between the two analyses since the important information would rather be whether the pathways and functions the genes associated with each drug in the Berkely analysis overlap with the pathways identified from my results, it shows that at least there is likely no one or two genes (except maybe for *OR2L13* in regard to Imatinib) for which the expression profile is outstanding enough relatively to a drug that it would be robustly selected across analyses.

However, there was also no real overlap on the functional pathways side either. Indeed, from investigations in IPA of the Berkeley LASSO-selected genes for the 10 shortlisted drugs, the results of which were described in the 5.3.1.3.1 Investigation of the IPA enrichment analyses interpretations subsection above, interesting patterns and functional pathways were only found

for three of the drugs (Tretinoin, Vemurafenib and Imatinib), and none of them strongly align with results of the analyses I conducted.

This suggests again failure to reproduce the Berkeley LASSO analysis results without access to their protocol. In addition, since results of multiple LASSO runs on the same data may bear different results, if a similar bootstrap approach as my own implementation was not used at Berkeley, there is even less chances to obtain matching results. But more importantly, while obtaining aligned results through similar but different pipelines would have strengthened confidence in them, these findings raise the question of robustness and adequacy of analyses. Beyond the computational protocol, the issue of the data in particular should also be pointed out, since the results of the Drugs Repurposing Project analyses were obtained used AUC as response variable, while the Berkeley LASSO analysis relied on the heavily approximated initial IC50s dataset, which likely also introduced bias in their analysis, although it was interesting to investigate whether that would lead to completely different results or not. Hence, divergence of findings is not too surprising either.

5.3.1.4 Validation of Glioblastoma cell culture models

For this study the goal was to determine whether the Glioblastoma cell cultures from EMC responded similarly to TMZ exposure as the tumours they originated from. For this the correlations between cell cultures response (AUCs, IC50s) and the patients response represented by their OS and PFS were calculated, with the covariate of MGMT promoter methylation status accounted for, since it is known to be a positive prognostic biomarker to TMZ response. The correlation results are detailed in **Table 18**.

	Correlation (p-value)			
MGMT status	**PFS x AUCs**	**PFS x IC50s**	**OS x AUCs**	**OS x IC50s**
Methylated (n = 23)	-0.136 (0.54)	0.0619 (0.78)	-0.482 (0.021)	-0.0652 (0.77)
Unmethylated (n = 24)	-0.188 (0.38)	0.0752 (0.727)	-0.199 (0.35)	0.0235 (0.91)
All (n = 47)	-0.260 (0.067)	-0.0340 (0.82)	-0.365 (0.0092)	-0.0515 (0.72)

Table 18: Correlations between cell cultures response to TMZ and patients' survival

From these correlations it appears that only AUCs and patients' OS are correlated, both when considering all cell cultures together and when only looking at cell cultures that have a methylated MGMT promoter status. For cultures that have an unmethylated MGMT promoter, correlation is not observed. There also seems to be no correlation when using either PFS and IC50s as representation of patients' and cultures' response to TMZ, respectively.

In addition to these correlations, any potential biomarkers predictive of TMZ response was searched in the DASL and GLIOTRAIN RNA-seq transcriptomics datasets. That search was conducted in both the DASL and RNA-Seq data, while also considering all cell cultures, only methylated MGMT promoter cell cultures and only unmethylated MGMT promoter cell cultures. The significant enrichment analysis results obtained from the list of genes found to be correlated with TMZ response without multiple testing adjustment of p-values are listed in ***Table* 19**. While each dataset/samples set combination seem to present a theme slightly more prevalent than others, such as immune response for the DASL/MGMT-methylated samples combination, cellular response to signaling for the DASL/MGMT-unmethylated samples combination, or interestingly cilium assembly for the RNA-Seq/all samples combination which is in line with results from subsection 5.3.2 GLIOTRAIN Data Analysis, they all also present a variety of other unrelated pathways. This suggests a high diversity and limited functional coherence in the genes found, which supports the finding that these genes would not be considered significantly correlated if p-values were adjusted for multiple testing, and that their correlation with TMZ response was purely coincidental. This is even further supported by the absence of MGMT in any of the correlate genes list, despite its well-known role in mitigating effects of TMZ.

Similarly, although that was not directly to demonstrate cell cultures representativity of their parental tumours, I used that same approach to evaluate the presence of biomarkers predictive of Cytarabine hydrochloride and Omacetaxine mepesuccinate response on request of the EMC team. But since there was no reason to believe that MGMT promoter methylation status would impact response to those drugs which have different mode of actions than TMZ, that cofactor was ignored here. The corresponding results are presented in ***Table* 20**. Here as well, absence of significant correlation when adjusting for multiple testing, high heterogeneity of pathways associated with genes correlated when not adjusting for multiple testing, and the disconnect between these pathways and the known mode of action of these drugs suggest that the correlations identified were only coincidental.

Table 19: Top 20 results from enrichment analyses on recurrent genes correlated to TMZ reponses for all, only MGMT methylated and only MGMT unmethylated cell cultures

RNA-Seq	DASL
All samples	
• canonical glycolysis • mitral valve formation • inner dynein arm assembly • protein localization to cilium • epithelial cilium movement involved in determination of left/right asymmetry • purine ribonucleotide biosynthetic process • negative regulation of oxidative stress-induced intrinsic apoptotic signaling pathway • nucleoside biosynthetic process • polyphosphate metabolic process • actomyosin contractile ring assembly • intraciliary transport involved in cilium assembly • positive regulation of bleb assembly • termination of mitochondrial transcription • negative regulation of ribosomal DNA heterochromatin assembly • purine ribonucleoside triphosphate biosynthetic process • regulation of mitochondrial membrane potential • spindle assembly • non-motile cilium assembly • regulation of cytokinesis • gluconeogenesis	• extracellular matrix organization • positive regulation of prostaglandin biosynthetic process • membrane protein ectodomain proteolysis • embryo implantation • spontaneous neurotransmitter secretion • positive regulation of cell growth • response to hypoxia • negative regulation of ubiquitin protein ligase activity • neutrophil chemotaxis • heterotypic cell-cell adhesion • positive regulation of synaptic transmission, glutamatergic • membrane to membrane docking • establishment of endothelial barrier • positive regulation of epidermal growth factor-activated receptor activity • regulated exocytosis • cellular response to tumor necrosis factor • neutrophil aggregation • negative regulation of gap junction assembly • negative regulation of protein neddylation • cell surface pattern recognition receptor signaling pathway
Methylated MGMT promoter samples	
• DNA replication initiation • G1/S transition of mitotic cell cycle • positive regulation of DNA-dependent DNA replication • nuclear DNA replication • ciliary basal body-plasma membrane docking • negative regulation of G protein-coupled receptor internalization • negative regulation of calcium ion import into sarcoplasmic reticulum • positive regulation of polyamine transmembrane transport • negative regulation of ATPase-coupled calcium transmembrane transporter activity • chorion development • regulation of mitotic cell cycle phase transition • regulation of centrosome duplication	• extracellular matrix organization • glomerular mesangial cell development • regulation of short-term neuronal synaptic plasticity • synaptic vesicle maturation • positive regulation of regulatory T cell differentiation • inflammatory response • branching involved in blood vessel morphogenesis • negative regulation of guanylate cyclase activity • negative regulation of gap junction assembly • recognition of apoptotic cell • negative regulation of cytokine secretion • lymph vessel development

• putrescine biosynthetic process from ornithine • thyroid hormone transport • centriole replication • double-strand break repair via break-induced replication • cell division • negative regulation by host of symbiont molecular function • regulation of CD40 signaling pathway • maintenance of lens transparency	• phospholipase C-activating G protein-coupled receptor signaling pathway • blood vessel remodeling • negative regulation of viral entry into host cell • angiogenesis • positive regulation of myeloid leukocyte differentiation • opioid receptor signaling pathway • regulation of T cell tolerance induction • T-helper 1 cell differentiation
Unmethylated MGMT promoter samples	
• regulation of transcription by RNA polymerase II • canonical glycolysis • fever generation • actin filament severing • polyphosphate catabolic process • regulation of microvillus length • intestinal D-glucose absorption • glyceraldehyde-3-phosphate biosynthetic process • terminal web assembly • septin ring organization • actomyosin contractile ring assembly • modification of postsynaptic actin cytoskeleton • cytoplasmic microtubule organization • Wnt signaling pathway, planar cell polarity pathway • positive regulation of transcription by RNA polymerase I • forebrain dorsal/ventral pattern formation • gluconeogenesis • mitochondrial translational termination • regulation of behavior • fructose 1,6-bisphosphate metabolic process	• detection of chemical stimulus involved in sensory perception of smell • G protein-coupled receptor signaling pathway • detection of chemical stimulus involved in sensory perception of sour taste • positive regulation of timing of catagen • hydrogen peroxide biosynthetic process • keratinization • secretory granule organization • fructose catabolic process to hydroxyacetone phosphate and glyceraldehyde-3-phosphate • adenylate cyclase-modulating G protein-coupled receptor signaling pathway • cellular response to cadmium ion • sodium ion transmembrane transport • inner ear auditory receptor cell differentiation • binding of sperm to zona pellucida • cellular response to platelet-derived growth factor stimulus • regulation of circadian sleep/wake cycle • protein localization to synapse • regulation of oxidative stress-induced cell death • regulation of multicellular organism growth • regulation of MDA-5 signaling pathway • antimicrobial humoral immune response mediated by antimicrobial peptide

Table 20: Top 20 results from enrichment analyses on recurrent genes correlated to Cytarabine and Omacetaxin response

Cytarabine hydrochloride	Omacetaxine mepesuccinate
• antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent • cellular response to potassium ion • cellular response to muramyl dipeptide • positive regulation of neuroblast proliferation • response to pheromone • poly-N-acetyllactosamine biosynthetic process • cerebral cortex radial glia guided migration • positive regulation of myelination • trachea morphogenesis • regulation of growth hormone activity • negative regulation of DNA recombination at telomere • cuticle development • glycerophosphate shuttle • O-glycan processing • fungiform papilla development • negative regulation of G2/M transition of mitotic cell cycle • synaptic vesicle maturation • hematopoietic stem cell differentiation • regulation of transcription elongation from RNA polymerase II promoter • homophilic cell adhesion via plasma membrane adhesion molecules	• cellular response to fibroblast growth factor stimulus • coronary artery morphogenesis • cholesterol catabolic process • positive regulation of RNA polymerase II transcription preinitiation complex assembly • regulation of receptor catabolic process • DNA replication initiation • mesenchymal cell proliferation • positive regulation of membrane protein ectodomain proteolysis • cell cycle G2/M phase transition • telomere maintenance • inositol phosphate dephosphorylation • regulation of generation of precursor metabolites and energy • myoblast differentiation involved in skeletal muscle regeneration • negative regulation of mitophagy in response to mitochondrial depolarization • pentacyclic triterpenoid metabolic process • L-ornithine import across plasma membrane • ncRNA deadenylation • regulation of cellular response to very-low-density lipoprotein particle stimulus • phospholipase C-activating adrenergic receptor signaling pathway • positive regulation of heparan sulfate proteoglycan binding

5.3.2 GLIOTRAIN Data Analysis

This analysis aimed at investigating potential resistance mechanism through Differential Expression Analysis (DEA) of individual datasets (GLIOTRAIN RNA-Seq and focal events data as well as the EMC DASL microarray datasets) between short-term and long-term survivors samples. This comparison between short-term and long-term samples was meant to highlight genes for which the molecular profile was particularly different between the two groups and thus, likely playing a key role in patient overall survival. Furthermore, comparison of the DEA results in other datasets and using the intermediate-term survivors for validation was also implemented.

Following the DEA of the RNA-Seq, 310 genes were identified as differentially expressed in the RNA-Seq data between long-term and short-term survivors. They are listed in **Table 24** from the 8.3 Results from the RNA-Seq DEA annex. The Gene Set Enrichment Analysis revealed that cell-motility-associated pathways were overrepresented among these genes, suggesting that aggressivity and resistance of the tumor to treatment leading to recurrence of Glioblastoma is strongly linked to the ability of the cancerous cells to propagate within the brain before surgery.

From the analysis of the EMC DASL data suggested a difference in expression of only the *SPRY4* and *PRKAR1B* proteins between ST and LT survivors. *SPRY4* is a MAPK signaling pathway inhibitor positioned upstream of *RAS* activation and as such could influence the PI3K/AKT pathway, while *PRKAR1B* is part of the *PKA* complex, which is involved in many phosphorylation processes and has been observed to be involved in cancer. Interestingly neither of these proteins has been strongly associated with Glioblastoma yet, as PubMed searches combining “Glioblastoma” and either of them as keywords bears limited and mostly off-topic results.

In regard to the focal events analysis, no chromosomal region, and by extension no gene, was identified as significantly differentially expressed between LT and ST survivors.

Unfortunately, the results obtained from analysis of the RNA-Seq, WGS and DASL data could not be reproduced in other datasets. Even disregarding the p-value, the trends (*i.e.* over- or under-expression of the gene between ST and LT survivors) were not consistently reproduced, which suggest that at least one of the steps of the whole pipeline is inadequate, whether at the level of data normalization or analysis which would lead to incorrect results, or the method of validation in other datasets itself is inappropriate. Unfortunately, I did not have the opportunity to thoroughly investigate the matter. While the possibility of a mistake which would invalidate the DEA results cannot be ruled out, it is also true that these results seem promising and biologically relevant to Glioblastoma.

6 Discussion and perspectives

6.1 Glioblastoma Disease Map

6.1.1 Genetic Alterations Representation

Over the course of the Glioblastoma Disease Map building, the need arose to develop a way to represent genetic alterations in the Disease Map. Indeed, many such alterations have been well characterized to be critical to Glioblastoma tumour development and survival, and thus clearly representing them would be required to study and uncover potential resistance mechanisms.

The model I defined for representing genetic alterations was designed to fit the needs for the Glioblastoma Disease Map. It allows for the accurate translation of most frequently observed mutations in Glioblastoma into the Disease Map format, without overloading the network because of the high number of reported mutations while still providing a way to reference all these mutations as notes for the model. I think the suggested representations and particularly the aggregation of mutations by outcome provides a useful tool and solid basis for the community to integrate genetic alterations into Disease Maps. In addition, beyond the defined guidelines the model could be further developed to also support some of the more complex alterations such as histones modifications and other epigenetic events, which were not covered in this work. Indeed, while the need for them did not arise in the context of core Glioblastoma driver mutations, they may play a role in the determination of the effectors expressed downstream of the three characterized pathways, and these modifications can also be relevant to many other diseases[138,139] and thus be used for the potential corresponding disease maps to make them more accurate and reliable resources for investigating these diseases.

It is also noteworthy that the developed model representation for chromosomal duplication focuses on intrachromosomal duplication. However, genes are often also amplified through extrachromosomal DNA. An explicit way to differentiate between intra- and extrachromosomal amplification of a single gene was not defined in the proposed model and should certainly be part of any future developments for representation of genetic alterations in the Disease Map standards.

In hindsight, this shortcoming likely comes from the fact that while the proposed genetic alterations model was developed with the help and feedback of Disease Map community members to ensure alignment with Disease Maps standards, insight from experts on genetic alterations themselves was not sought out to validate the accuracy and completeness of the

model. As a consequence, this work and any future development on it would greatly benefit from submission and validation by biologists extensively knowledgeable about genetic alterations in cancer.

6.1.2 Produced Disease Map

Building a Glioblastoma Disease Map constituted a large part of my PhD research. Through identification of frequently altered pathways in Glioblastoma and modelling them, I created the foundations for the representation of Glioblastoma molecular mechanisms, where the most common mutations in Glioblastoma and their impact on functional pathways have been modelled.

An important challenge in the assembly of the network was the difficulty in finding Glioblastoma-specific characterization of molecular interactions.
Indeed, although there were quite a few Glioblastoma-related mutations identified in the literature, their effect on the gene or protein was rarely investigated as well. The presence of the mutation was noted, but how that impacted the corresponding pathway was not clear, and therefore in many cases that impact had to be modelled as a supposition based on the role of the normal protein in its pathway.
Similarly, articles describing either a section or an entire pathway were often describing it on either a general, non-altered level where the cascade may be characterized in detail, or on a scale broader than specifically Glioblastoma such as “in disease” or “in cancer” where alterations of the pathway in several diseases are mentioned and sometimes compared but rarely in depth. Meanwhile, information about the workings of these pathways in Glioblastoma was scarce, and there was no information about Glioblastoma-specific downstream targets of *AKT* or transcription factors. As a consequence, all of these targets had to be considered to be potentially equally affected by the Glioblastoma-specific alterations of their upstream regulators. A significant improvement for the Disease Map, and understanding of Glioblastoma in general, would thus be to investigate Glioblastoma cells expression profiles to determine whether indeed all targets are affected similarly, or whether only a few specific ones are involved in the context of Glioblastoma.

It is important to note that I consider this Glioblastoma Disease Map to be a first version of the network, which would benefit from further development and expansion, and in greater details. While the network does present the fundamental mechanisms of IDH-wildtype Glioblastoma, the interactions included involve either the start of a signaling cascade for the RTK/RAS/PI3K/AKT pathway, or regulation of transcription factors for the RB and TP53

pathways, but their downstream effectors were only partly investigated and there are likely several interesting and important subsequent pathways that were not identified and explored. As a consequence, the limited size and coverage of the Glioblastoma Disease Map compromise it's biological accuracy and relevance, as well as its use for an integrative analysis, to guide investigation of whole-genome and whole-transcriptome datasets which cover thousands of genes that are not represented in the network.

As a consequence, the next step for the development of the Glioblastoma Disease Map should be to expand and possibly refine the network.

One of the methods to achieve this would be to further explore literature to extensively characterize the targets of the *mTORC1* complex*, FOXO, E2F* and *TP53* transcription factors, but also of *AKT* and *RB* themselves. Other pathways as well, such as the Notch[140–143] or Epithelial-to-Mesenchymal Transition[144–147] pathways have been mentioned in the literature as potentially relevant to Glioblastoma, although not to the same extent as the three pathways characterized in this work.

Moreover, computational methods to enrich the network based on the GLIOTRAIN quantitative data may also be considered in further completing the map, as well as in identifying Glioblastoma-specific downstream effectors of the pathways.

An alternative would be to consider genes of interest identified from quantitative analyses of Glioblastoma-related data, such as the DEA of GLIOTRAIN data performed in this work, and extract the corresponding interactions from other curated networks. Several such projects have been considered and ultimately the Atlas of Cancer Signaling Network (ACSN) was found to be the best choice. The ACSN is a network following Disease Map standard, and focusing on cancer more broadly rather than Glioblastoma. It represents 9,692 entities involved in 8,137 interactions, and was manually curated based on 4,532 publications[148], providing a very thorough and high-quality Disease Map. As such, the ACSN would be a great resource to isolate and extract Glioblastoma-specific interactions to complete the Glioblastoma Disease Map. Taking this idea further, it could even be used in the integrative analysis instead of the Glioblastoma Disease Map, and from the results of that analysis extract the corresponding subnetworks from the ACSN and integrate them into the Glioblastoma Disease Map.

Beyond these suggestions to pursue development and improvement of the Disease Map, several shortcomings in the approach used during the PhD to build it should be pointed out.

Indeed, a major issue in the approach taken was to start with functional pathways rather than genetic alterations which would have been more logical and provided better grounded insight

on which pathways to focus on integrating in the network, although the choice of pathways was by no means completely arbitrary.

Furthermore, as mentioned above, literature about downstream impact of pathways alterations in Glioblastoma was limited, and a large part of the map was build based on literature about broader conditions such as cancer, or even healthy tissue. As a result, the accuracy, relevance and representativity of the network for Glioblastoma may be questioned. To validate the produced Disease Map, projection of quantitative data onto it would have helped confirm or contradict relevance of the overall network.

In addition, rather than building the Disease Map from scratch which was extremely time-consuming, it would have made more sense to start from an existing resource such as the ACSN or the SIGNOR network and modify it towards Glioblastoma specificity. The suggestion to extract subnetworks of interest from the ACSN mentioned above came from that realization in order to make up for that mistake and complete the Glioblastoma map with information from that resource.

Finally, while Disease Maps are supposed to be a community-driven endeavor, I failed to actively seek ought insight and validation of the model by my peers of the GLIOTRAIN project and other Glioblastoma experts, and performed the work myself and basing most of the work on articles alone. It is likely that had that not been the case, the shortcomings mentioned so far could have been greatly mitigated, and the Glioblastoma Disease Map itself would likely be more complete, reliable, and usable for quantitative analysis.

Nevertheless, I would argue that the produced Glioblastoma Disease Map is still a solid foundation for investigation of Glioblastoma resistance mechanisms. Indeed, through qualitative analysis of the network I could confirm that the RTK/RAS/PI3K/AKT, RB and TP53 pathways really are at the center of Glioblastoma tumors, and several cross-talks between them can already be highlighted at the protein-protein interactions level. In addition, mutual exclusion and co-occurrence patterns in frequently observed genetic alterations suggest that targeting any one of these alterations alone for treatment would likely not be enough, since even in already well-known Glioblastoma alterations there are paths towards resistance mechanisms, *i.e.* the rise of mutations leading to the same outcome as the one that was prevented by treatment.

However, through study of these paths on the disease map, some preliminary countermeasures can be devised. For instance, it seems clear that *EGFR* mutations and *CDKN2A/CDKN2B/ARF* locus homozygous deletion can disrupt all three pathways, with just these two nodes of the network, while the alternative requires many more mutations to reach

the same level of disruption. As a result, it may be worth investigating whether, for patients that present both *EGFR* alterations and *ARF* locus homozygous deletion, a treatment that both repress *EGFR* signaling and remediates to *ARF* locus homozygous deletion could significantly reduce tumor development and improve the patients' survival, since the tumor would then need the simultaneous apparition of many mutations to overcome treatment effects. In addition, since *RB1* mutations seem to be at the center of the alternative pathway, it could be worth considering to also target them in the treatment. Of course, these are speculations based on observation of a network, and I lack the medical knowledge to assess the availability of such treatments and feasibility of such a study. For validation of these hypotheses, reaching out to expert biologists and/or pharmacologists in these domains, for instance to my collaborators at EMC, would have been the final step of this network investigation. Unfortunately, these interpretations came late during the PhD and could only be briefly mentioned to them. Nevertheless, these observations and interpretations constitute an important part of my research towards understanding resistance mechanisms of Glioblastoma, and may benefit future studies.

6.2 Data Management Methods and Systems

The work conducted on different data management systems and frameworks led to the understanding and implementation of good practices for data handling all throughout the PhD. The issues encountered while compiling CTI data highlighted that there is always a probability of human mistakes in the collection of the data, which should therefore be carefully screened before integrating it in any analysis.

This principle was applied in curation of the GLIOTRAIN data for upload on the tranSMART database, through which the FAIR data principles were fully pursued. Issues with the GLIOTRAIN data were identified and documented, and the complexity of this process, to obtain relevant information about data sequencing pipeline, emphasized the importance of extensive documentation about transformations operated on the data in its use for an analysis and to increase confidence in its results. These observations were further confirmed during the analyses for identification of predictive biomarkers of drug response, and in particular through the comparison of the early Berkeley LASSO analysis results to my own which showed little overlap, likely due to divergence of methodologies stemming from absence of documentation of the Berkeley analysis, as is discussed in the corresponding subsection 6.3.1.2 Berkeley LASSO Analysis review below.

Finally, both the tranSMART database and OMOP projects work brought considerations about appropriate and useful formatting of the data. Indeed, while data may be provided in a certain format, that format may not be adapted to the database and require transformations. On the other hand, how this data is expected to be formatted for analysis should also be taken into consideration as it may lead to misunderstanding later on. Such issues were encountered for the GLIOTRAIN data where members of the consortium expected to get actual FASTQ sequences instead of data derived from it, for OMOP projects where data owners were unfamiliar with the OMOP CDM and surprised of the mapping results, but also for the analyses for identification of predictive biomarkers of drug response in which the initial IC50s dataset was inappropriate for actual quantitative analysis. As a consequence, documentation and communication upstream of the curation of data, to manage expectations and clarify what options are available and should be chosen was put forward in the OMOP mapping projects SOPs and its importance was communicated to members of the GLIOTRAIN consortium.

It should be mentioned that these data management considerations were initially not planned to take such a large role in the PhD work but rather grew from what was initially supposed to

be minor support to the GLIOTRAIN consortium infrastructure and ITTM activities. As a result, the work performed there was not as well-structured and defined as research at it should have been: while literature was still searched for FAIR principles, appropriate normalization methods or common mapping practices in the OMOP community, the screening was not as in-depth as it should have been for research standards. Furthermore, had it been better planned it could have been defined and carried out to be better integrated within the thesis work, towards goals of confirming hypotheses or benchmarking, with clear publishable results.

Beyond the application of FAIR data principles and mindful handling of data, the work carried out around Data Management systems and framework bore several resources which will continue to be used beyond the end of this PhD:

- the GLIOTRAIN tranSMART database will remain available to the GLIOTRAIN consortium members for at least five years
- the SOPs, Machine-readable mapping syntax and ETL software implemented are used and will be further refined at ITTM in OMOP mapping projects

6.3 Data Analysis

6.3.1 Identification of predictive biomarkers of drug response

6.3.1.1 Drugs Repurposing Project

Thanks to the LASSO and WGCNA analyses of the DASL and AUC data from EMC, I was able to generate a lot of potential associations between gene expression profile and response to the screened drugs. However, that also meant spending a lot of time going through all of them to see emerging patterns and determine which were the more interesting results.

First of all, a challenge in the interpretation of these results was that in some cases, the functional pathways resulting from a given analysis were very specific, while in other cases they could span over a much broader range of processes and biological functions (*e.g.* pathways related to cell cycle, DNA repair, cell differentiation, immune response, PI3K, MAPK or apoptotic signaling pathways…), making it difficult to interpret them and assess their relevance and reliability. Although such loosely associated results appeared in results from all types of analysis run combinations, they seemed to occur more frequently when using either the WGCNA approach or the unbiased set of genes, and as a result were particularly more frequent in analyses that combined both. That can be explained by the fact that WGCNA tends to generate clusters of genes much larger than LASSO-selected groups of genes, and the unbiased set of genes is by definition larger and includes all types of genes, regardless of function. Furthermore, it seems like these broader results come up more frequently for drugs for which curve-fitting resulted in a large proportion (>25%) of non-log-logistic models. This does not seem to be correlated with method or genes subset, and likely stems from a higher heterogeneity in the AUC responses, leading to a broader range of genes for which expression may appear connected to drug response. The sample subset used did not seem to impact the proportion of specific or broad range of functional pathways in the result of the run.

Another relevant outlook on this analysis would be the general comparison of the results based on the parameters of the analysis they were generated from, *i.e.* whether the curve-fitting process produced well-defined models or not, which method (WGCNA or LASSO) was applied, with which DASL data subset of genes (unbiased or cancer genes) and which sample of cell cultures (all Glioblastoma or only Primary Glioblastoma cell cultures) as input. The quality of the curve models produced to compute AUCs does not appear to have an important impact on the results of analysis. Whether the models produced were well defined and heterogenous enough to suggest variable response, very homogenous leading me to

initially expect limited discriminatory power, or even in cases where the range of tested concentrations seemed inadequate to the drug, results that were similar in relevance to Glioblastoma were obtained. It is worth mentioning that drugs presenting well-defined, heterogenous models led to the identification of a lot more results than with drugs with homogenous models, which themselves still produced more results than drugs screened over an unsuitable range of concentrations. The results for drugs for which models presented an important proportion of non-log-logistic models however seemed to be more variable, a little less relevant to Glioblastoma context, especially when it was cumulated with a large number of cell cultures excluded (ten or more, only happened for drugs with high non-log-logistic models proportion) from the analysis, and a tested range of concentrations seemingly inadequate to the drug.

Concerning the method used, WGCNA typically found many more genes correlated with AUC drug response than LASSO. However, as mentioned earlier this large number of genes which tends to be associated to a broad range of functional processes, and also can end up being correlated to several drugs due to a different subset of these genes, but lead to similar and thus redundant functional pathways. As a consequence, it may be worth it to run the WGCNA analyses again but defining parameters to limit the size of genes clusters, and potentially get more specific results. In addition, when both LASSO and WGCNA produce results for a given drug, they usually hardly overlap and thus do not validate each other. Despite my best efforts to ensure appropriate settings for the analyses, this could be due to different reasons such as inappropriate normalization of the input data, influence of external factors that were not accounted for in the models, non-linear relationship between gene expression and response variable, or failing to adjust p-values of correlation tests between WGCNA clusters and drug response for multiple-testing, which would have made the WGCNA approach more stringent. Unfortunately, I was unable to extensively investigate this discrepancy, and since although not overlapping the results from both analyses still seemed relevant to Glioblastoma, the decision was made to pursue interpretation of the results and highlight that any interesting and promising result should be taken with a grain of salt and demands validation by experimentation or further analysis, since it was not clear which of the two analyses methods may be inappropriately applied.

As for the subset of genes, the unbiased approach typically resulted in more drug/genes association found and a more diverse range of functions that may be associated, than the cancer genes approach for which identified functional pathways tended to be very redundant (cell cycle, DNA repair, immune response) yet still overlapping with results from the unbiased

approach. As such, the list of cancer genes defined may need to be expanded, but appears to be relatively accurate since the emerging pathways can be associated with cancer hallmarks. In addition, both approaches allow to look at the data under complementary angles, and I cannot find a reason to prefer one over the other.

For the samples of Glioblastoma cell cultures included in the analysis, it is worth noting that for all other parameters of the analysis run equal, the results obtained when including all or only Primary Glioblastoma samples tend to be similar but not identical, although there are also cases where they appear to be completely different. Analyses using only Primary Glioblastoma samples usually produce more gene-to-drug association results (except for analysis using LASSO and the cancer genes subset, where there were more results when including all samples), and for a few drugs only one analysis run with Primary Glioblastoma samples produced a result. As a result, I find it difficult to determine which subset would be more interesting to focus on: on one hand the Primary Glioblastoma samples produce more results and should present more homogenous expression profiles than when analyzed alongside Recurrent Glioblastoma samples, while on the other hand these additional cell lines may introduce some variability helpful to discriminate between responsive and non-responsive cell cultures while also increasing the sample size. Furthermore, since results from either subset are not similar enough to disregard one or the other, it could be that both should still be investigated.

Overall, besides identifying the fact that drugs for which experimental response data was likely inadequate to properly fully capture drug response of the cell cultures lead to less reliable results which should be more carefully scrutinized, and the possible caveat of inappropriate setting of one or both analysis methods, comparison of the results obtained from different parameters of analyses runs do not really allow to identify a set of parameters more robust or reliable than others. However, the genes identified through the different experimental set-ups and the associated functional pathways emerging from them seemed to align with expectations for Glioblastoma, such as pathways related to cell cycle, apoptosis, immune response, neuronal morphogenesis, *etc.* As a consequence, although doubt was raised regarding the correct parametrization of the analyses methods, the results would suggest the general pipeline was not completely misguided.

Another issue which would have been relevant to explore was the relevance and validation of clusters identified through the WGCNA method. Indeed, while the method clusters genes based on expression profile, it may be interesting to then determine whether the genes associated within a given clustered are actually related, or whether the co-expression relationship found

may only be coincidental. This could be done for instance by consulting molecular network and pathways databases to see if the genes belong to the same or connected pathways, and we could even consider using the ACSN as well for that purpose.

Finally, looking at the actual results to identify which of the drugs may be more appropriate to treat Glioblastoma and investigate potential predictive biomarkers, with a very superficial understanding of the mode of action of the drugs I would have shortlisted

- Epirubicin hydrochloride for which the results were specifically associated with cell cycle processes,
- Oxaliplatin which got results from three different WGCNA runs with somewhat diverse results, but neuron development-associated pathways were consistently present
- Pazopanib hydrochloride mostly selected by LASSO runs with hits such as cell cycle, DNA repair, neuron remodeling and regulation of transcription
- Bleomycin sulfate, which is a cytotoxic antibiotic, and one of the runs associated genes from susceptibility to cytotoxicity pathways to it, so I would focus on these genes in particular
- Topotecan hydrochloride has interesting results, but since the tested range of concentrations seems inappropriate, I would suggest measuring new datapoints and re-run the analysis to make sure
- Vandetanib is a drug that inhibits EGFR-mediated survival, so the unique result of Insulin-like growth factor receptor pathway, another Receptor Tyrosine Kinase, is interesting
- Dasatinib which supposedly inhibits proliferation, adhesion, migration and invasion, and is associated with neuron genesis pathways in the analyses results

Of course, that selection should be done along with the EMC team, who is more knowledgeable about the biological and pharmacological workings of the different drugs. However, while I was finishing the analysis and compiling results, they switched their focus to the study for validating cell cultures models and requested my help on it, postponing further investigation of my results to a later date which did not come before the end of my PhD. Hence, to conclude this analysis it would be worthwhile to have these results extensively investigated, and validated by testing the corresponding drugs on cell cultures presenting an expression profile of the potential predictive biomarker suggesting either sensibility or resistance to the drug.

6.3.1.2 Berkeley LASSO Analysis review

In the investigations of the results from the Berkeley University LASSO analysis of the data for the drugs repurposing project, I started by trying to reproduce in the IPA software the identification of genes of interests among the ones that were associated to the 10 shortlisted drugs. This was met with mixed results: while my findings aligned with Berkeley's in regard to Tretinoin, Vemurafenib and Imatinib, which were the drugs that seemed most promising for Glioblastoma treatment based on the findings from the LASSO analysis, I could not reproduce their interpretations of genes of interest for the other seven drugs. Nevertheless, the three drugs for which our results did concur were encouraging prospects for new Glioblastoma treatment, and I was hoping that my own analysis from the Drugs Repurposing project would confirm these drugs as good candidates for Glioblastoma treatment.

However, following execution of the LASSO and WGCNA analyses pipeline, it came out that my results were not in line with the ones from the Berkeley LASSO analysis. While there is an overlap in the list of drugs that were found to have potential predictive biomarkers, said biomarkers do not correspond to the same functional pathways and processes that were found at Berkeley. Furthermore, the drugs that I would find more promising for Glioblastoma treatment based on the results of my analysis were different than the ones shortlisted in the Berkeley analysis.

As a consequence, we can consider that the Berkeley LASSO analysis and my own lead to different results and do not validate each other. However, it should be highlighted that since I only had a very general understanding of the pipeline that was used in the Berkeley analysis, without a clear documentation on the parameters and reasoning behind the different steps, our methodologies were certainly different, although the extent of the differences could not be ascertained. Among the most fundamental divergences that I can think of are that the steps used to explore LASSO-selected genes in IPA software were likely different since I only got a general overview of the pipeline of the Berkeley analysis rather than a detailed description of the steps and parameters used despite requesting it. But even more significantly, I was told that the Berkeley analysis made use of the initial IC50s dataset, which contains a mix of numerical and categorical values, by approximating the categorical values using an undisclosed methodology, though the approximation rules applied were not disclosed. This is an important difference between the two pipelines, which likely explain in large part differences in our results, and as was stated previously, I believe using that dataset was bad practice

because of the important extrapolations used for calculating IC50s and inconsistencies it presents and as a consequence I feel more confident in my results.

6.3.1.3 Validation of Glioblastoma cell culture models

Correlation tests between cell cultures' and patients' response to TMZ suggest that although only partially, the cell cultures do reflect how their parental tumours respond to TMZ treatment. In particular, a lower AUC of the cell culture, which suggests that the culture was more responsive, does correlate with a longer patient survival time, regardless of MGMT status. This is in line with the goal of the study to demonstrate that the cell cultures models are representative of the parental tumour they are derived from, using TMZ response as a measurement of this similarity. The results from the methylated and unmethylated MGMT promoter groups also make sense biologically with the hypothesis: when methylated, MGMT does not mitigate the damages done by TMZ treatment, so the corresponding cell cultures should be responsive to it with the same relative amplitude as their parental tumours. On the other hand, an unmethylated MGMT promoter leads to a considerably lower impact of TMZ treatment on the cancer, and thus other factors are at play in the toxicity of TMZ for both the cultures and OS for the patients.

The fact that only AUCs and OS are correlated however raises questions regarding both IC50s and PFS. Indeed, if the cell cultures do behave similarly to the parental tumours they are derived from when exposed to TMZ like the OS x AUCs correlation would suggest, it is puzzling that the pattern was not observed with OS x IC50s correlation, since IC50s are also a measurement of drug response. Unfortunately I could not find an explanation for this difference, although a non-linear relationship between them is a lead that would be worth investigating. As for the PFS, which should be a representation of patient response and thus we could expect it to be correlated with AUCs same as the OS (and the p-value was closer to 0.05 than any of the other non-significant tests), it may be interesting to look into the reasons why both OS and PFS are routinely collected, since it could provide a lead on what the differences between the two are, and on whether other factors may need to be included for a model analysis to be predictive of PFS.

As far as finding predictive biomarkers to drug response, be it for TMZ, Cytarabine and Omacetaxine, results were not conclusive in both the RNA-Seq and DASL datasets. The functional pathways associated spanned a very diverse range of processes, with close to no relation to the known mode of action of each drug. This was an unfortunate but not unexpected outcome, since the lists of genes investigated for each drug were already questionable, in the

sense that the genes were found correlated to drug response only if multiple-testing bias was ignored. When accounting for it and applying a correction, no gene passed the 0.05 q-value threshold to be considered significant. Thus, it is likely that correlations found between genes and any of the drug were purely coincidental. Another issue was the fact that expression of MGMT was absent from the list of identified genes, which is puzzling since as mentioned before it should play an important role in discriminating cell that are responsive or not to TMZ when considering either all cultures or only the unmethylated MGMT subset.

These observations could have several explanations. First the number of cell cultures for which RNA-Seq or DASL data was available was relatively small, and even smaller when further stratified by MGMT promoter methylation status. This leads to smaller power of the analysis, and finding significant correlations becomes less likely. Another possible explanation is that there could be other cofactors at play, such as age or sex, which would require further stratifications of the cell lines for significant patterns to emerge. Finally, it could also be that such patterns are non-linear and could only be detected through multivariate analyses or non-linear regression models to be detected rather than with correlation tests for each single gene. This would also explain why MGMT was not identified in the results, if predictive potential is dependent on a non-linear relation to drug response.

As a consequence, the next steps for this analysis should be to test for potential covariates other than MGMT promoter methylation status to identify any relevant one, and perform a multivariate analysis, using for instance the same LASSO or WGCNA pipeline as in the Drugs Repurposing Project to attempt detection of predictive biomarkers.

6.3.2 GLIOTRAIN Data Analysis

The Differential Expression Analysis of GLIOTRAIN data yielded mixed results. 310 genes were identified from the RNA-Seq dataset as significantly discriminating between short-term and long-term survivors, and 2 from the EMC DASL data.

However, a similar trend in the profile of these genes in other datasets, including TCGA dataset on Glioblastoma, could not be detected. This could have several causes, at any step of the pipeline, which unfortunately made it difficult to investigate and identify a cause more likely than others. Indeed, the issue might be that the DEA results themselves are invalid, either because the processing, normalization or subsetting methods of any of the datasets was inadequate, or the DEA itself was wrongly set up. Such a situation could lead to falsely identify patterns that are actually not present in the data and thus could not be reproduced in other datasets. Alternatively, the issue could also come from using inadequate statistical tests to see if the

genes found from the DEA presented similar behaviour in other datasets. Although the pipeline was carefully reviewed, the issue could not be solved before the end of the PhD. While every effort was made to ensure the methods applied were adapted to each data type, this situation suggests that mistakes were still made, and again highlight the importance of reproducibility and independent validation to ensure that published findings are reliable, or at least point out potential issues with them.

Despite the uncertainty on the exact source of the validation issue, these identified genes may still be relevant to Glioblastoma as they appear to be in line with known features of the disease. The 310 genes seem to be heavily involved in cell motility pathways, which may explain at least partly the difference between the two groups of patients: higher diffusion abilities lead to more of the tumour cells escaping from resection at first surgery, which means more tumour cells able to proliferate as well as to develop resistances to treatments, thus resulting in faster recurrence of the cancer and degradation of the patient's health. While this does not deal with the direct molecular mechanisms through which treatment resistance emerges, it represents an important finding which may be used for prognostic evaluation of patients. As such, it would be interesting in the future to further investigate those 310 genes to determine and validate their potential as prognostic biomarkers for faster recurrence.

In regard to the 2 genes from the EMC DASL dataset, one is part of the *Sprouty* gene family which appears in the Glioblastoma Disease Map as regulators of the RAS/RAF/MAPK cascade, and has received increased attention in the recent years for its role in cancers[149–151], including gliomas and glioblastomas[152–154]. Thus, this result would confirm relevance of this protein family for Glioblastoma. The second gene itself, *PRKAR1B*, does not appear to have been widely characterized as a key driver of oncogenesis, but it may be worth investigating how the Protein Kinase A complex to which it belongs may be involved in Glioblastoma processes.

In addition, to further identify other potentially relevant genes associated to these results, projecting these genes onto the ACSN network and using a diffusion algorithm[29,30], this projection may then be used to identify and extract subnetworks relevant to Glioblastoma which could be used to enrich the Glioblastoma Disease Map.

However, these results were obtained using only the subset of short-term and long-term survivors from datasets, which constituted only a limited number of samples. It would have been interesting to also run an analysis including intermediate-term survivors with overall survival as a continuous response rather than an ordinal one. More samples included may have

granted more analysis power, and it would have been interesting to compare the results with the existing DEA results.

In addition, these results are already interesting in themselves, but they rely on individual analysis of each dataset separately, and the Glioblastoma Disease Map is barely used downstream of these analyses as a resource to locate the identified genes in the context of Glioblastoma as represented from the literature. To draw on the full potential of these materials, the initial plan for this PhD was to design and execute an integrative analysis involving all data and guided by the literature knowledge through the Disease Map. While this point was not reached over the course of my studies due to time constraints, the general pipeline was taking shape and it would have been interesting to see what would have come of it.
The next steps for the integrative data analysis would have been to:

- for each patient,
    - for each GLIOTRAIN dataset, project the data of the patient onto the ACSN network, and run a diffusion algorithm to attribute weights or score to molecular interactions based on their profile in the dataset
    - merge the networks corresponding to the different datasets for the patients using an algorithm similar to the Similarity Network Fusion[155] iterative merging method
    - Extract the weight associated to each interaction from the merged network as a vector for the patient
- run an unsupervised clustering algorithm on the network weights of the patients
- characterize the resulting clusters, including the main interactions defining them and the clinical profile of the patients they contain
- The interactions driving cluster definitions can be extracted from the ACSN to enrich the Glioblastoma Disease Map
- The clusters definition may be validated using the EMC DASL and TCGA data

This integrative pipeline makes use of both the Disease Map, with ACSN here since as mentioned previously the current Glioblastoma Disease Map does not include most of the genes present in large omics dataset, to guide the analysis and the data from multiple omics dataset. Furthermore, if the clusters definitions do not appear to correlate with Overall Survival, a supervised clustering algorithm may be considered as well to bring that dimension into consideration in the definition of clusters. In addition, this approach does not require all samples to have data available in all datasets. As such, we could also consider calculating weights for

EMC data samples and TCGA samples to include them in the clustering instead of as validation sets, although I would be less leaning towards this approach of mixing completely different samples, collected and sequenced in completely different conditions over which we have little knowledge and control.

While the general pipeline for this analysis has been formulated, the specific methods to use have yet to be determined and implemented. This includes in particular the diffusion algorithm, the adaptation of the SNF iterative merging algorithm, and the clustering method. Nevertheless, even if it could not be carried out, I believe the results could be very interesting to investigate.

6.4 Conclusion

Over the course of my PhD studies, research towards the elucidation of resistance mechanisms led me to elaborate a Glioblastoma-specific Disease Map and to analyze multiple omics datasets. Along the way, implementation of data management principles provided insights not only on considerations necessary for suitable and transparent storing of data, but also on the importance of documenting and communicating information relative to any transformation operated on it, both in the context of processing it to make it available and for the purpose of analysis, since without that knowledge mishandling the data, producing biased results or misinterpreting quickly becomes easy. The development of the Glioblastoma Disease Map allowed for the highlighting of key mechanisms of Glioblastoma and of their interconnections, facilitating investigations into the molecular emergence of resistances to treatment. The work achieved in this domain yielded both a core for the Glioblastoma Disease Map which should be further expanded in the future, and the definition of a model for genetic alterations within the Disease Map framework which was presented to the community and open for adoption in other projects. Finally, in order to analyze omics dataset both for the Identification of predictive biomarkers of drug response analyses and the GLIOTRAIN analysis, extensive multivariate analysis pipelines have been defined and partly implemented and executed. While these investigations could not be completed all the way during this PhD, the results they produced were promising and suggested that resuming and finishing them may lead to a breakthrough on our understanding of Glioblastoma resistance mechanisms and potential ways to overcome them.

7 References

1. Ostrom, Q. T. *et al.* CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2012-2016. *Neuro-Oncol.* **21**, v1–v100 (2019).
2. Malmström, A. *et al.* Temozolomide versus standard 6-week radiotherapy versus hypofractionated radiotherapy in patients older than 60 years with glioblastoma: the Nordic randomised, phase 3 trial. *Lancet Oncol.* **13**, 916–926 (2012).
3. Koshy, M. *et al.* Improved survival time trends for glioblastoma using the SEER 17 population-based registries. *J. Neurooncol.* **107**, 207–212 (2012).
4. Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol. (Berl.)* **131**, 803–820 (2016).
5. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncol.* **23**, 1231–1251 (2021).
6. Kleihues, P. & Ohgaki, H. Primary and secondary glioblastomas: from concept to clinical diagnosis. *Neuro-Oncol.* **1**, 44–51 (1999).
7. Ohgaki, H. *et al.* Genetic pathways to glioblastoma: a population-based study. *Cancer Res.* **64**, 6892–6899 (2004).
8. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
9. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
10. Nørøxe, D. S., Poulsen, H. S. & Lassen, U. Hallmarks of glioblastoma: a systematic review. *ESMO Open* **1**, e000144 (2016).

11. Mao, H., LeBrun, D. G., Yang, J., Zhu, V. F. & Li, M. Deregulated Signaling Pathways in Glioblastoma Multiforme: Molecular Mechanisms and Therapeutic Targets. *Cancer Invest.* **30**, 48–56 (2012).
12. Aldape, K., Zadeh, G., Mansouri, S., Reifenberger, G. & von Deimling, A. Glioblastoma: pathology, molecular mechanisms and markers. *Acta Neuropathol. (Berl.)* **129**, 829–848 (2015).
13. Nakada, M. *et al.* Aberrant signaling pathways in glioma. *Cancers* **3**, 3242–3278 (2011).
14. Wolbers, J. G. Novel strategies in glioblastoma surgery aim at safe, supra-maximum resection in conjunction with local therapies. *Chin. J. Cancer* **33**, 8–15 (2014).
15. Lukas, R. V. *et al.* Newly Diagnosed Glioblastoma: A Review on Clinical Management. *Oncol. Williston Park N* **33**, 91–100 (2019).
16. Stupp, R. *et al.* Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.* **352**, 987–996 (2005).
17. Lee, S. Y. Temozolomide resistance in glioblastoma multiforme. *Genes Dis.* **3**, 198–210 (2016).
18. Esteller, M. *et al.* Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N. Engl. J. Med.* **343**, 1350–1354 (2000).
19. Hegi, M. E. *et al.* MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* **352**, 997–1003 (2005).
20. Tanaka, S., Louis, D. N., Curry, W. T., Batchelor, T. T. & Dietrich, J. Diagnostic and therapeutic avenues for glioblastoma: no longer a dead end? *Nat. Rev. Clin. Oncol.* **10**, 14–26 (2013).
21. Paolillo, M., Boselli, C. & Schinelli, S. Glioblastoma under Siege: An Overview of Current Therapeutic Strategies. *Brain Sci.* **8**, (2018).
22. Vleeschouwer, S. D. *Upcoming Cutting-Edge Innovations*. *Glioblastoma [Internet]* (Codon Publications, 2017).

23. Zhang, B., Tian, Y. & Zhang, Z. Network biology in medicine and beyond. *Circ. Cardiovasc. Genet.* **7**, 536–547 (2014).
24. Sonawane, A. R., Weiss, S. T., Glass, K. & Sharma, A. Network Medicine in the Age of Biomedical Big Data. *Front. Genet.* **10**, 294 (2019).
25. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst.* **5**, 251-267.e3 (2017).
26. Huang, L. *et al.* Driver network as a biomarker: systematic integration and network modeling of multi-omics data to derive driver signaling pathways for drug combination prediction. *Bioinformatics* **35**, 3709–3717 (2019).
27. McGillivray, P. *et al.* Network Analysis as a Grand Unifier in Biomedical Data Science. *Annu. Rev. Biomed. Data Sci.* **1**, 153–180 (2018).
28. Koutrouli, M., Karatzas, E., Paez-Espino, D. & Pavlopoulos, G. A. A Guide to Conquer the Biological Network Era Using Graph Theory. *Front. Bioeng. Biotechnol.* **8**, 34 (2020).
29. Charmpi, K., Chokkalingam, M., Johnen, R. & Beyer, A. Optimizing network propagation for multi-omics data integration. *PLoS Comput. Biol.* **17**, e1009161 (2021).
30. Dimitrakopoulos, C. *et al.* Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* **34**, 2441–2448 (2018).
31. Stingo, F. C., Chen, Y. A., Tadesse, M. G. & Vannucci, M. Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Ann. Appl. Stat.* **5**, (2011).
32. Ietswaart, R., Gyori, B. M., Bachman, J. A., Sorger, P. K. & Churchman, L. S. GeneWalk identifies relevant gene functions for a biological context using network representation learning. *Genome Biol.* **22**, 55 (2021).
33. Jang, Y., Yu, N., Seo, J., Kim, S. & Lee, S. MONGKIE: an integrated tool for network analysis and visualization for multi-omics data. *Biol. Direct* **11**, 10 (2016).

34. von Mering, C. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**, D433–D437 (2004).

35. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).

36. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

37. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).

38. Licata, L. *et al.* SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. *Nucleic Acids Res.* **48**, D504–D510 (2020).

39. Kuperstein, I. *et al.* Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis* **4**, e160 (2015).

40. Mazein, A. *et al.* Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. *NPJ Syst. Biol. Appl.* **4**, 21 (2018).

41. Ostaszewski, M. *et al.* Community-driven roadmap for integrated disease maps. *Brief. Bioinform.* **20**, 659–670 (2019).

42. Le Novère, N. *et al.* The Systems Biology Graphical Notation. *Nat. Biotechnol.* **27**, 735–741 (2009).

43. Hucka, M. Systems Biology Markup Language (SBML). in *Encyclopedia of Systems Biology* (eds. Dubitzky, W., Wolkenhauer, O., Cho, K.-H. & Yokota, H.) 2057–2063 (Springer New York, 2013). doi:10.1007/978-1-4419-9863-7_1091.

44. Gawron, P. *et al.* MINERVA-a platform for visualization and curation of molecular interaction networks. *NPJ Syst. Biol. Appl.* **2**, 16020 (2016).

45. Kondratova, M., Sompairac, N., Barillot, E., Zinovyev, A. & Kuperstein, I. Signalling maps in cancer research: construction and data analysis. *Database J. Biol. Databases Curation* **2018**, (2018).

46. Dorel, M., Viara, E., Barillot, E., Zinovyev, A. & Kuperstein, I. NaviCom: a web application to create interactive molecular network portraits using multi-level omics data. *Database J. Biol. Databases Curation* **2017**, (2017).

47. Bonnet, E. *et al.* NaviCell Web Service for network-based data visualization. *Nucleic Acids Res.* **43**, W560-565 (2015).

48. Kuperstein, I., Robine, S. & Zinovyev, A. Network biology elucidates metastatic colon cancer mechanisms. *Cell Cycle Georget. Tex* **14**, 2189–2190 (2015).

49. Monraz Gomez, L. C. *et al.* Application of Atlas of Cancer Signalling Network in preclinical studies. *Brief. Bioinform.* **20**, 701–716 (2019).

50. Bonnet, E. *et al.* BiNoM 2.0, a Cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. *BMC Syst. Biol.* **7**, 18 (2013).

51. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

52. Athey, B. D., Braxenthaler, M., Haas, M. & Guo, Y. tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. *AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci.* **2013**, 6–8 (2013).

53. Scheufele, E. *et al.* tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform. *AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci.* **2014**, 96–101 (2014).

54. HDD Data Curation and Normalization - tranSMART Project wiki - Confluence. https://wiki.transmartfoundation.org/display/transmartwiki/HDD+Data+Curation+and+Normalization.

55. Supported Data Types - tranSMART Project wiki - Confluence. https://wiki.transmartfoundation.org/display/transmartwiki/Supported+Data+Types.

56. Stang, P. E. *et al.* Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership. *Ann. Intern. Med.* **153**, 600 (2010).

57. Reisinger, S. J. *et al.* Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *J. Am. Med. Inform. Assoc. JAMIA* **17**, 652–662 (2010).

58. Defalco, F. J., Ryan, P. B. & Soledad Cepeda, M. Applying standardized drug terminologies to observational healthcare databases: a case study on opioid exposure. *Health Serv. Outcomes Res. Methodol.* **13**, 58–67 (2013).

59. Voss, E. A., Ma, Q. & Ryan, P. B. The impact of standardizing the definition of visits on the consistency of multi-database observational health research. *BMC Med. Res. Methodol.* **15**, 13 (2015).

60. Voss, E. A. *et al.* Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J. Am. Med. Inform. Assoc.* **22**, 553–564 (2015).

61. Gini, R. *et al.* Data Extraction and Management in Networks of Observational Health Care Databases for Scientific Research: A Comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies. *EGEMS Wash. DC* **4**, 1189 (2016).

62. Hripcsak, G. *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud. Health Technol. Inform.* **216**, 574–578 (2015).

63. Who We Are – OHDSI. https://ohdsi.org/who-we-are/.

64. Ryan, P. B. *et al.* Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat. Med.* **31**, 4401–4415 (2012).

65. Schuemie, M. J. *et al.* Replication of the OMOP Experiment in Europe: Evaluating Methods for Risk Identification in Electronic Health Record Databases. *Drug Saf.* **36**, 159–169 (2013).
66. Ryan, P. Statistical challenges in systematic evidence generation through analysis of observational healthcare data networks. *Stat. Methods Med. Res.* **22**, 3–6 (2013).
67. Weng, C. *et al.* A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. *Appl. Clin. Inform.* **5**, 463–479 (2014).
68. Chapter 4 The Common Data Model | The Book of OHDSI. https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html.
69. Chapter 5 Standardized Vocabularies | The Book of OHDSI. https://ohdsi.github.io/TheBookOfOhdsi/StandardizedVocabularies.html.
70. *International statistical classification of diseases and related health problems*. (World Health Organization, 2004).
71. Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T. & Moore, R. Normalized names for clinical drugs: RxNorm at 6 years. *J. Am. Med. Inform. Assoc. JAMIA* **18**, 441–448 (2011).
72. Rabbit in a Hat. http://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html.
73. Usagi. http://ohdsi.github.io/Usagi/.
74. Informatics, O. H. D. S. and. *Chapter 6 Extract Transform Load | The Book of OHDSI*.
75. EHDEN Academy. https://academy.ehden.eu/.
76. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).
77. R Core Team. R: A Language and Environment for Statistical Computing. (2021).
78. RStudio Team. RStudio: Integrated Development Environment for R. (2020).
79. PubMed. *PubMed* https://pubmed.ncbi.nlm.nih.gov/.

80. Vivanco, I. & Sawyers, C. L. The phosphatidylinositol 3-Kinase AKT pathway in human cancer. *Nat. Rev. Cancer* **2**, 489–501 (2002).

81. Toker, A. & Marmiroli, S. Signaling specificity in the Akt pathway in biology and disease. *Adv. Biol. Regul.* **55**, 28–38 (2014).

82. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).

83. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

84. Inoki, K., Li, Y., Zhu, T., Wu, J. & Guan, K.-L. TSC2 is phosphorylated and inhibited by Akt and suppresses mTOR signalling. *Nat. Cell Biol.* **4**, 648–657 (2002).

85. Engelman, J. A., Luo, J. & Cantley, L. C. The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. *Nat. Rev. Genet.* **7**, 606–619 (2006).

86. Tzivion, G., Dobson, M. & Ramakrishnan, G. FoxO transcription factors; Regulation by AKT and 14-3-3 proteins. *Biochim. Biophys. Acta* **1813**, 1938–1945 (2011).

87. Burgering, B. M. T. & Medema, R. H. Decisions on life and death: FOXO Forkhead transcription factors are in command when PKB/Akt is off duty. *J. Leukoc. Biol.* **73**, 689–701 (2003).

88. Accili, D. & Arden, K. C. FoxOs at the crossroads of cellular metabolism, differentiation, and transformation. *Cell* **117**, 421–426 (2004).

89. Mayo, L. D. & Donner, D. B. A phosphatidylinositol 3-kinase/Akt pathway promotes translocation of Mdm2 from the cytoplasm to the nucleus. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 11598–11603 (2001).

90. Diehl, J. A., Cheng, M., Roussel, M. F. & Sherr, C. J. Glycogen synthase kinase-3beta regulates cyclin D1 proteolysis and subcellular localization. *Genes Dev.* **12**, 3499–3511 (1998).

91. Degirmenci, U., Wang, M. & Hu, J. Targeting Aberrant RAS/RAF/MEK/ERK Signaling for Cancer Therapy. *Cells* **9**, (2020).

92. Kidger, A. M., Sipthorp, J. & Cook, S. J. ERK1/2 inhibitors: New weapons to inhibit the RAS-regulated RAF-MEK1/2-ERK1/2 pathway. *Pharmacol. Ther.* **187**, 45–60 (2018).

93. Burkhart, D. L. & Sage, J. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nat. Rev. Cancer* **8**, 671–682 (2008).

94. Dick, F. A. & Rubin, S. M. Molecular mechanisms underlying RB protein function. *Nat. Rev. Mol. Cell Biol.* **14**, 297–306 (2013).

95. Witkiewicz, A. K. & Knudsen, E. S. Retinoblastoma tumor suppressor pathway in breast cancer: prognosis, precision medicine, and therapeutic interventions. *Breast Cancer Res. BCR* **16**, 207 (2014).

96. Wade, M., Li, Y.-C. & Wahl, G. M. MDM2, MDMX and p53 in oncogenesis and cancer therapy. *Nat. Rev. Cancer* **13**, 83–96 (2013).

97. Jesionek-Kupnicka, D. *et al.* TP53 promoter methylation in primary glioblastoma: relationship with TP53 mRNA and protein expression and mutation status. *DNA Cell Biol.* **33**, 217–226 (2014).

98. Bykov, V. J. N., Eriksson, S. E., Bianchi, J. & Wiman, K. G. Targeting mutant p53 for efficient cancer therapy. *Nat. Rev. Cancer* **18**, 89–102 (2018).

99. Bieging, K. T., Mello, S. S. & Attardi, L. D. Unravelling mechanisms of p53-mediated tumour suppression. *Nat. Rev. Cancer* **14**, 359–370 (2014).

100. Mischel, P. S., Nelson, S. F. & Cloughesy, T. F. Molecular analysis of glioblastoma: pathway profiling and its implications for patient therapy. *Cancer Biol. Ther.* **2**, 242–247 (2003).

101. Ueki, K. *et al.* CDKN2/p16 or RB alterations occur in the majority of glioblastomas and are inversely correlated. *Cancer Res.* **56**, 150–153 (1996).

102. Funahashi, A., Morohashi, M., Kitano, H. & Tanimura, N. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO* **1**, 159–162 (2003).

103. Funahashi, A. *et al.* CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proc. IEEE* **96**, 1254–1265 (2008).
104. The SAS software. Copyright © 2018 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.
105. Schag, C. C., Heinrich, R. L. & Ganz, P. A. Karnofsky performance status revisited: reliability, validity, and guidelines. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **2**, 187–193 (1984).
106. EMA. Temodal. *European Medicines Agency* https://www.ema.europa.eu/en/medicines/human/EPAR/temodal (2018).
107. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
108. TCGA Barcode - GDC Docs. https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/#tcga-barcode.
109. Dray, S. & Dufour, A.-B. The **ade4** Package: Implementing the Duality Diagram for Ecologists. *J. Stat. Softw.* **22**, (2007).
110. Data loading tools - tranSMART Project wiki - Confluence. https://wiki.transmartfoundation.org/display/transmartwiki/Data+loading+tools.
111. Krämer, A., Green, J., Pollard, J. & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinforma. Oxf. Engl.* **30**, 523–530 (2014).
112. Adrian Alexa, J. R. topGO: Enrichment Analysis for Gene Ontology. (2020) doi:10.18129/B9.BIOC.TOPGO.
113. Balvers, R. K. *et al.* Serum-free culture success of glial tumors is related to specific molecular profiles and expression of extracellular matrix-associated gene modules. *Neuro-Oncol.* **15**, 1684–1695 (2013).

114. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinforma. Oxf. Engl.* **24**, 1547–1548 (2008).

115. Cheang, M. C. U. *et al.* Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J. Natl. Cancer Inst.* **101**, 736–750 (2009).

116. Pan Du, R. B. lumi. (2017) doi:10.18129/B9.BIOC.LUMI.

117. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat. Oxf. Engl.* **8**, 118–127 (2007).

118. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinforma. Oxf. Engl.* **28**, 882–883 (2012).

119. nCounter PanCancer Pathways Panel. *NanoString* https://www.nanostring.com/products/ncounter-assays-panels/oncology/ncounter-pancancer-pathways-panel/.

120. PanCancer Progression. *NanoString* https://www.nanostring.com/products/ncounter-assays-panels/oncology/pancancer-progression/.

121. TruSight Oncology 500 Assay | For pan-cancer biomarkers in DNA and RNA. https://www.illumina.com/products/by-type/clinical-research-products/trusight-oncology-500.html.

122. AmpliSeq for Illumina Comprehensive Cancer Panel Data Sheet | Illumina. https://science-docs.illumina.com/documents/LibraryPrep/ampliseq-comprehensive-cancer-panel-data-sheet-770-2017-023/Content/Source/Library-Prep/AmpliSeq/comprehensive-cancer-panel/ampliseq-comprehensive-cancer-panel-data-sheet.html.

123. rt2 profiler pcr arrays - GeneGlobe. https://geneglobe.qiagen.com/us/product-groups/rt2-profiler-pcr-arrays.

124. Qiagen Introduces Sequencing Workflow, Including NGS System and Library-Prep Instrument, at AGBT. *Genomeweb* https://www.genomeweb.com/sequencing/qiagen-introduces-sequencing-workflow-including-ngs-system-and-library-prep-inst (2013).
125. ClearSeq Cancer Research Panels. https://hpst.cz/sites/default/files/oldfiles/clearseqcancerflyer-5991-6124en1.pdf.
126. Tiriac, H. *et al.* Organoid Profiling Identifies Common Responders to Chemotherapy in Pancreatic Cancer. *Cancer Discov.* **8**, 1112–1129 (2018).
127. C. Ritz, F. Baty, J. C. Streibig, & D. Gerhard. Dose-Response Analysis Using R. (2016).
128. James M. Curran. Bolstad2. (2013).
129. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, (2010).
130. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
131. Langfelder, P. & Horvath, S. Fast R Functions for Robust Correlations and Hierarchical Clustering. *J. Stat. Softw.* **46**, i11 (2012).
132. cBioPortal for Cancer Genomics::Datasets. https://www.cbioportal.org/datasets.
133. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
134. Venables, W. N., Ripley, B. D. & Venables, W. N. *Modern applied statistics with S*. (Springer, 2002).
135. Booklet for the 4th Disease Maps Community Meeting, Sevilla. (2019).
136. Human Genome Overview - Genome Reference Consortium. https://www.ncbi.nlm.nih.gov/grc/human.
137. Verheul, C. *et al.* Generation, characterization, and drug sensitivities of 12 patient-derived IDH1-mutant glioma cell cultures. *Neuro-Oncol. Adv.* **3**, vdab103 (2021).

138. Baylin, S. B. & Jones, P. A. Epigenetic Determinants of Cancer. *Cold Spring Harb. Perspect. Biol.* **8**, a019505 (2016).

139. Zoghbi, H. Y. & Beaudet, A. L. Epigenetics and Human Disease. *Cold Spring Harb. Perspect. Biol.* **8**, a019497 (2016).

140. Fan, X. *et al.* NOTCH pathway blockade depletes CD133-positive glioblastoma cells and inhibits growth of tumor neurospheres and xenografts. *Stem Cells Dayt. Ohio* **28**, 5–16 (2010).

141. Guessous, F. *et al.* microRNA-34a is tumor suppressive in brain tumors and glioma stem cells. *Cell Cycle Georget. Tex* **9**, 1031–1036 (2010).

142. Kefas, B. *et al.* The neuronal microRNA miR-326 acts in a feedback loop with notch and has therapeutic potential against brain tumors. *J. Neurosci. Off. J. Soc. Neurosci.* **29**, 15161–15168 (2009).

143. Brown, D. V. *et al.* Expression of CD133 and CD44 in glioblastoma stem cells correlates with cell proliferation, phenotype stability and intra-tumor heterogeneity. *PloS One* **12**, e0172791 (2017).

144. Iwadate, Y. Epithelial-mesenchymal transition in glioblastoma progression. *Oncol. Lett.* **11**, 1615–1620 (2016).

145. Iser, I. C., Pereira, M. B., Lenz, G. & Wink, M. R. The Epithelial-to-Mesenchymal Transition-Like Process in Glioblastoma: An Updated Systematic Review and In Silico Investigation. *Med. Res. Rev.* **37**, 271–313 (2017).

146. Guarino, M., Rubino, B. & Ballabio, G. The role of epithelial-mesenchymal transition in cancer pathology. *Pathology (Phila.)* **39**, 305–318 (2007).

147. Hollier, B. G., Evans, K. & Mani, S. A. The epithelial-to-mesenchymal transition and cancer stem cells: a coalition against cancer therapies. *J. Mammary Gland Biol. Neoplasia* **14**, 29–43 (2009).

148. ACSN/About. https://acsn.curie.fr/ACSN2/about.html.

149. Masoumi-Moghaddam, S., Amini, A. & Morris, D. L. The developing story of Sprouty and cancer. *Cancer Metastasis Rev.* **33**, 695–720 (2014).

150. Kawazoe, T. & Taniguchi, K. The Sprouty/Spred family as tumor suppressors: Coming of age. *Cancer Sci.* **110**, 1525–1535 (2019).

151. Qiu, B. *et al.* Sprouty4 correlates with favorable prognosis in perihilar cholangiocarcinoma by blocking the FGFR-ERK signaling pathway and arresting the cell cycle. *EBioMedicine* **50**, 166–177 (2019).

152. Celik-Selvi, B. E. *et al.* Sprouty3 and Sprouty4, Two Members of a Family Known to Inhibit FGF-Mediated Signaling, Exert Opposing Roles on Proliferation and Migration of Glioblastoma-Derived Cells. *Cells* **8**, E808 (2019).

153. Hausott, B. *et al.* Subcellular Localization of Sprouty2 in Human Glioma Cells. *Front. Mol. Neurosci.* **12**, 73 (2019).

154. Park, J.-W. *et al.* Sprouty2 enhances the tumorigenic potential of glioblastoma cells. *Neuro-Oncol.* **20**, 1044–1054 (2018).

155. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).

8 Annexes

8.1 Berkeley LASSO Analysis Results

DNA-damaging agents	Target / Mechanism of Action	LASSO-selected Genes/Probesets
Dacarbazine	Alkylating agents (triazines)	OTX2
Bendamustine hydrochloride	Alkylating agents (nitrogen mustard)	DACH1.2, DUOXA2, ERP44.1, LAMB2L, MIR331, ZNF483.3
Melphalan hydrochloride	"	KHDRBS2
Uracil mustard	"	ACVR1C, BTRC.1, IL1RL1.1, MBOAT2.1, SEC24B.1, SLC25A44, SLC9A2, TAC3.2
Streptozocin	Alkylating agents (nitrosoureas)	INPP5J, RHBG
Carmustine	"	BMF.2
Lomustine	"	ACVR1C, ADPRHL1, AGAP3, ATE1.3, DOC2B, HTR1B, PDPK1, PPP4R4.2, SLC25A22, ST8SIA6, TMEM233
Busulfan	Alkylating agents (alkylsufonates)	AMBN.1, C9orf11, DACH1.2, DMGDH, FXYD2.1, KCNJ12, PMEPA1.1, THAP7.2
Oxaliplatin	"	CRH
Carboplatin	"	C2orf88, DEFB119, FBXW11.2, HTR1B, OPN4.1, PCDH21, TCERG1L, ZNF483.3
Cisplatin	"	MAPK8IP3.2
Cytarabine hydrochloride	Antimetabolite	PTPN20B
Pentostatin	"	CASP7.2, KIF20B, SLC38A2, ZDHHC18
Methotrexate	"	FOXR2, PRSS38
Floxuridine	"	OPN4.1
Fluorouracil	"	KCNC2.1
Pemetrexed	"	GLI2, HPGD, SMTN.2
Decitabine	Nucleoside analogue	DIS3
Azacitidine	Nucleoside analogue	CTXN3.1
Dexrazoxane	Topoisomerase II inhibitor	ASB12, CC2D1B.1, ICAM5, LPHN1, MRM1, OLA1.1, RTBDN.2, SEMA6D.4, SPPL2B
Bleomycin sulfate	Cytotoxic antibiotic	CBWD2, DEFB119, DEFB124, DENND2C, GLTPD2, PCDHA1, PHF6.2, RIPPLY2.1, TCERG1L
Hydroxyurea	Ribonucleoside diphosphate reductase inhibitor	ATG12, LOC286238, WNT3, ZPBP2.1
Arsenic trioxide	Metalloid oxides	CCL1, CYP4X1, DIO2, HOXB9, ISM1, LYPD6B.1, OTUD4, PMS2, SMPDL3B, UPK3B.3
Tyrosine kinase inhibitors	**Target / Mechanism of Action**	**LASSO-selected Genes/Probesets**

Imatinib	Bcr-Abl	MIRLET7D, OR2L13, PRPF40B.1, Sep.04, SLC8A3.2, SLITRK1.1, XRN2
Dasatinib	Bcr-Abl	C9orf135
Erlotinib hydrochloride	EGFR	CCR2, ITIH1
Pazopanib hydrochloride	Multi-targeted kinase inhibitor	ALOX12, C15orf27, CASQ2, CBLN4, CPLX3, CT45A5, CTXN3.1, FAM81A, FAM9A, GABRB2, XRN2, ZNF676, GPR128, LCN6, MAPK1, MYH11, NECAB1, OPRK1, OPRM1.5, PLN.1, TMEM144, TPSD1, TROVE2.3, UCN3
Sorafenib	Multi-targeted kinase inhibitor	PDE1A, SLC01A2.2
Trametinib	MEK 1	CPA3, SLC22A2.2
Vemurafenib	BRAF	CTSG, DSP.2, HBD, SLC22A2.2, SLC6A20, TNNT2
Others	***Target / Mechanism of Action***	***LASSO-selected Genes/Probesets***
Allopurinol	Xanthine oxidase inhibitor	CD207
Bortezomib	Proteasome inhibitor	CTAG2
Enzalutamide	Nonsteroidal antiandrogen (NSAA)	ACVR1C, CTXN3.1, KHDRBS2, SPTB.1
Estramustine phosphate sodium	Antigonadotropic / Antiandrogen	GNB5.1, MYO3A
Sirolimus	mTOR inhibitor	B4GALNT2, MGC57359.1, Sep.14
Tretinoin	Retinoid analogue	DAO, DNAI2, KIF19, PTPN3, RNF7.1, SEMA3E, SLC39A12, TRPM3.1
Fulvestrant	Estrogen receptor antagonist	C6orf204, DAB2IP.2, GPR116, IPP, MIR513A2, PANK2, SPAG4L, TRPM9.1, WNT10A, ZNF585A.1
Mitotane	Steroidogenesis inhibitor / diphenylmethanes	CALCA.2

Table 21: Berkeley Lasso analysis results: Genes associated to each drug by LASSO selection. *Gene names 'Sep.04' and 'Sep.14' are likely Excel artifacts for 'SEPT4' and 'SEPT14' gene names.*

8.2 Drugs Repurposing LASSO and WGCNA results

<table>
<tr><th colspan="2">Bendamustine hydrochloride</th></tr>
<tr><th colspan="2">Unbiased genes set | Primary Glioblastoma samples</th></tr>
<tr><td colspan="2">• regulation of small GTPase mediated signal transduction
• positive regulation of GTPase activity
• regulation of GTPase activity
• small GTPase mediated signal transduction</td></tr>
<tr><th colspan="2">Carboplatin</th></tr>
<tr><th colspan="2">Cancer genes set | Primary Glioblastoma samples</th></tr>
<tr><td colspan="2">• positive regulation of NF-kappaB transcription factor activity
• positive regulation of DNA-binding transcription factor activity
• transmembrane receptor protein serine/threonine kinase signaling pathway
• regulation of DNA-binding transcription factor activity
• regulation of signaling receptor activity
• cellular response to growth factor stimulus
• response to growth factor</td></tr>
<tr><th colspan="2">Dactinomycin</th></tr>
<tr><th>Cancer genes set | All Glioblastoma samples</th><th>Cancer genes set | Primary Glioblastoma samples</th></tr>
<tr><td>• preantral ovarian follicle growth
• negative regulation of ovarian follicle development
• gonadal mesoderm development
• Mullerian duct regression
• sex determination
• positive regulation of NF-kappaB transcription factor activity
• positive regulation of DNA-binding transcription factor activity
• transmembrane receptor protein serine/threonine kinase signaling pathway
• aging
• regulation of DNA-binding transcription factor activity</td><td>• canonical Wnt signaling pathway involved in positive regulation of cardiac outflow tract cell prolif...
• positive regulation of fibroblast growth factor receptor signaling pathway
• glial cell fate determination</td></tr>
<tr><th colspan="2">Enzalutamide</th></tr>
<tr><th colspan="2">Cancer genes set | All Glioblastoma samples</th></tr>
<tr><td colspan="2">• positive regulation of canonical Wnt signaling pathway
• positive regulation of stress-activated MAPK cascade
• positive regulation of stress-activated protein kinase signaling cascade
• positive regulation of Wnt signaling pathway
• negative regulation of canonical Wnt signaling pathway</td></tr>
</table>

• negative regulation of Wnt signaling pathway • regulation of stress-activated MAPK cascade • regulation of stress-activated protein kinase signaling cascade • regulation of canonical Wnt signaling pathway	
Everolimus	
Unbiased genes set | Primary Glioblastoma samples	**Cancer genes set | All Glioblastoma samples**
• cellular response to tumor necrosis factor • response to tumor necrosis factor • neutrophil degranulation • neutrophil activation involved in immune response • neutrophil mediated immunity • neutrophil activation • granulocyte activation • leukocyte degranulation • myeloid cell activation involved in immune response • myeloid leukocyte mediated immunity • myeloid leukocyte activation • leukocyte activation involved in immune response • cell activation involved in immune response • leukocyte mediated immunity	• positive regulation of canonical Wnt signaling pathway • positive regulation of stress-activated MAPK cascade • positive regulation of stress-activated protein kinase signaling cascade • positive regulation of Wnt signaling pathway • negative regulation of canonical Wnt signaling pathway • negative regulation of Wnt signaling pathway • regulation of stress-activated MAPK cascade • regulation of stress-activated protein kinase signaling cascade • regulation of canonical Wnt signaling pathway
Idarubicin hydrochloride	
Unbiased genes set | Primary Glioblastoma samples	
• mast cell chemotaxis • mast cell cytokine production • mast cell degranulation • antimicrobial humoral response • negative regulation of blood vessel diameter • regulation of blood pressure • negative regulation of neuron death • regulation of neuron death	

Megestrol acetate		
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**	**Cancer genes set | Primary Glioblastoma samples**
• daunorubicin metabolic process • doxorubicin metabolic process • progesterone metabolic process	• cellular response to tumor necrosis factor • response to tumor necrosis factor • neutrophil degranulation	• positive regulation of glial cell-derived neurotrophic factor secretion

• prostaglandin metabolic process • digestion • positive regulation of protein kinase B signaling • regulation of protein kinase B signaling • protein kinase B signaling • epithelial cell differentiation • G protein-coupled receptor signaling pathway • positive regulation of cell proliferation • positive regulation of intracellular signal transduction • positive regulation of signal transduction • regulation of cell proliferation • positive regulation of cell communication • positive regulation of signaling	• neutrophil activation involved in immune response • neutrophil mediated immunity • neutrophil activation • granulocyte activation • leukocyte degranulation • myeloid cell activation involved in immune response • myeloid leukocyte mediated immunity • myeloid leukocyte activation • leukocyte activation involved in immune response • cell activation involved in immune response • leukocyte mediated immunity	
Melphalan hydrochloride		
Unbiased genes set | All Glioblastoma samples		
• angiogenesis • heart development • blood vessel morphogenesis • blood vessel development • vasculature development • cardiovascular system development		
Mitomycin		
Unbiased genes set | Primary Glioblastoma samples		
• calcium ion transmembrane transport • calcium ion transport • divalent metal ion transport • divalent inorganic cation transport • regulation of ion transmembrane transport • regulation of transmembrane transport • inorganic cation transmembrane transport • regulation of ion transport • inorganic ion transmembrane transport • cation transmembrane transport • metal ion transport		
Nelarabine		

Unbiased genes set | Primary Glioblastoma samples
• positive regulation of ERK1 and ERK2 cascade • protein localization to nucleus • regulation of ERK1 and ERK2 cascade • ERK1 and ERK2 cascade • positive regulation of MAPK cascade
Nilotinib
Cancer genes set | Primary Glioblastoma samples
• regulation of toll-like receptor signaling pathway • TRIF-dependent toll-like receptor signaling pathway • regulation of tumor necrosis factor-mediated signaling pathway • positive regulation of I-kappaB kinase/NF-kappaB signaling • NIK/NF-kappaB signaling • regulation of I-kappaB kinase/NF-kappaB signaling • I-kappaB kinase/NF-kappaB signaling • regulation of inflammatory response
Pazopanib hydrochloride

Unbiased genes set | Primary Glioblastoma samples	Cancer genes set | All Glioblastoma samples	Cancer genes set | Primary Glioblastoma samples
• regulation of mitotic cell cycle phase transition • regulation of cell cycle phase transition • mitotic cell cycle phase transition • cell cycle phase transition • regulation of mitotic cell cycle • cell division • regulation of cell cycle process • mitotic cell cycle process • mitotic cell cycle • regulation of cell cycle • cell cycle process • cell cycle	• negative regulation of neuron remodeling • negative regulation of dendrite extension • negative regulation of branching morphogenesis of a nerve	• negative regulation of transcription by RNA polymerase II • negative regulation of transcription, DNA-templated • negative regulation of nucleic acid-templated transcription • negative regulation of RNA biosynthetic process • negative regulation of RNA metabolic process

Raloxifene
Unbiased genes set | Primary Glioblastoma samples
• transmembrane transport • transport • establishment of localization

- localization
- single strand break repair
- regulation of DNA recombination
- regulation of mitotic recombination
- mitotic spindle elongation

Romidepsin

Cancer genes set | Primary Glioblastoma samples

- signal transduction involved in mitotic G2 DNA damage checkpoint
- positive regulation of telomerase catalytic core complex assembly
- negative regulation of TORC1 signaling
- establishment of protein-containing complex localization to telomere
- meiotic telomere clustering
- positive regulation of DNA damage response, signal transduction by p53 class mediator
- phosphatidylinositol-3-phosphate biosynthetic process
- negative regulation of B cell proliferation
- positive regulation of DNA catabolic process
- regulation of microglial cell activation

Sirolimus

Unbiased genes set | All Glioblastoma samples

- chloride transport
- ion transmembrane transport
- transmembrane transport
- transport
- ion transport
- inorganic anion transport
- anion transport

Sorafenib

Unbiased genes set | All Glioblastoma samples

- post-translational protein modification
- protein modification process
- cellular protein modification process
- macromolecule modification
- cellular protein metabolic process
- protein metabolic process

Temsirolimus

Cancer genes set | All Glioblastoma samples

- positive regulation of single-stranded telomeric DNA binding
- telomere assembly
- protection from non-homologous end joining at telomere
- establishment of protein localization to telomere
- negative regulation of telomere maintenance via telomerase
- positive regulation of telomere maintenance via telomerase
- positive regulation of telomerase activity

Thiotepa

Unbiased genes set | Primary Glioblastoma samples

- regulation of membrane repolarization during action potential
- calcium ion import across plasma membrane
- regulation of calcium ion transmembrane transport via high voltage-gated calcium channel
- inorganic ion homeostasis
- calcium ion transmembrane import into cytosol
- cellular homeostasis
- positive regulation of calcium ion transmembrane transporter activity

Topotecan hydrochloride

Unbiased genes set | Primary Glioblastoma samples

- B cell receptor transport into membrane raft
- positive regulation of activated T cell proliferation
- glomerular visceral epithelial cell differentiation
- T cell costimulation
- positive regulation of protein tyrosine kinase activity
- positive regulation of MAP kinase activity
- intrinsic apoptotic signaling pathway
- positive regulation of protein serine/threonine kinase activity

Uracil mustard

Cancer genes set | All Glioblastoma samples

- regulation of Cdc42 protein signal transduction
- positive regulation of interleukin-2 secretion
- negative regulation of long-term synaptic potentiation
- B cell proliferation involved in immune response
- regulation of modification of synaptic structure
- positive regulation of Wnt signaling pathway, planar cell polarity pathway
- positive regulation of substrate adhesion-dependent cell spreading

Vandetanib

Cancer genes set | All Glioblastoma samples
• negative regulation of integrin biosynthetic process • positive regulation of insulin-like growth factor receptor signaling pathway
Vismodegib
Cancer genes set | All Glioblastoma samples / **Cancer genes set | Primary Glioblastoma samples**
• response to glucocorticoid • response to corticosteroid • neutrophil activation involved in immune response • neutrophil degranulation • neutrophil activation • granulocyte activation • neutrophil mediated immunity • leukocyte degranulation • myeloid cell activation involved in immune response • myeloid leukocyte mediated immunity • myeloid leukocyte activation • leukocyte activation involved in immune response • cell activation involved in immune response / • negative regulation of cell-substrate adhesion • negative regulation of angiogenesis • negative regulation of blood vessel morphogenesis • positive regulation of angiogenesis
Vorinostat
Cancer genes set | All Glioblastoma samples
• positive regulation of intrinsic apoptotic signaling pathway in response to osmotic stress • positive regulation of granulosa cell apoptotic process • positive regulation of B cell differentiation • positive regulation of mitochondrial membrane potential • positive regulation of release of cytochrome c from mitochondria

Table 22: Enrichment analysis results derived from the EMC Drugs Repurposing Project LASSO analysis results

Amiodarone hydrochloride		
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**	**Cancer genes set | Primary Glioblastoma samples**
All - Cluster 11 • cell division • G1/S transition of mitotic cell cycle • DNA replication initiation • DNA unwinding involved in DNA replication • positive regulation of mitotic nuclear division • mitotic chromosome condensation	Primary – Cluster 16 • DNA replication-dependent nucleosome assembly • nucleotide-excision repair, DNA gap filling • telomere maintenance via semi-conservative replication • positive regulation of DNA-directed DNA polymerase activity	Primary – Cluster 1 • DNA replication initiation • cell division • anaphase-promoting complex-dependent catabolic process • G1/S transition of mitotic cell cycle

• G2/M transition of mitotic cell cycle • meiotic cell cycle • positive regulation of DNA-dependent DNA replication initiation <u>All - Cluster 13</u> • cell division • nucleosome assembly • DNA replication-dependent nucleosome assembly • mitotic metaphase plate congression • telomere capping • cell cycle • positive regulation of cytokinesis • regulation of mitotic cell cycle spindle assembly checkpoint • attachment of mitotic spindle microtubules to kinetochore • G2/M transition of mitotic cell cycle • double-strand break repair • regulation of cytokinetic process	• base-excision repair, gap-filling • telomere capping • double-strand break repair via nonhomologous end joining • telomere organization • mismatch repair • DNA repair <u>Primary – Cluster 25</u> • cell division • anaphase-promoting complex-dependent catabolic process • regulation of attachment of spindle microtubules to kinetochore • positive regulation of cytokinesis • mitotic metaphase plate congression • nucleosome assembly • G2/M transition of mitotic cell cycle • mitotic spindle organization • DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest • positive regulation of mitotic metaphase/anaphase transition	• mitotic spindle assembly checkpoint • G2/M transition of mitotic cell cycle • DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest • regulation of response to DNA damage stimulus • telomere maintenance • mitotic centrosome separation • negative regulation of monocyte differentiation

Arsenic trioxide	
Unbiased genes set | All Glioblastoma samples	**Cancer genes set | All Glioblastoma samples**
<u>All - Cluster 8</u> • Golgi inheritance • Golgi localization <u>All - Cluster 9</u> • central nervous system myelination • positive regulation of oligodendrocyte progenitor proliferation <u>All - Cluster 14</u> • cellular response to cell-matrix adhesion • mitotic cell cycle	<u>All - Cluster 1</u> • cell proliferation • MAPK cascade • cellular response to calcium ion • negative regulation of signal transduction • cell surface receptor signaling pathway • positive regulation of GTPase activity • positive regulation of intracellular signal transduction • neurogenesis • positive regulation of transporter activity • regulation of Ras protein signal transduction • negative regulation of neuron apoptotic process

All - Cluster 15 • cellular macromolecule biosynthetic process • canonical Wnt signaling pathway involved in stem cell proliferation • regulation of gene expression • cellular response to retinoic acid • regulation of macromolecule biosynthetic process • canonical Wnt signaling pathway involved in midbrain dopaminergic neuron differentiation • positive regulation of DNA methylation • canonical Wnt signaling pathway involved in mesenchymal stem cell differentiation	• regulation of cell death • renal sodium ion absorption
Azacitidine	
Unbiased genes set | Primary Glioblastoma samples	**Cancer genes set | Primary Glioblastoma samples**
Primary – Cluster 1 • positive regulation of cellular senescence • negative regulation of ERK1 and ERK2 cascade • negative regulation of phosphatase activity • negative regulation of dephosphorylation Primary – Cluster 25 • cell division • anaphase-promoting complex-dependent catabolic process • mitotic sister chromatid segregation • mitotic spindle assembly checkpoint • mitotic spindle organization • mitotic metaphase plate congression • mitotic cytokinesis • negative regulation of stress-activated MAPK cascade • DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest • regulation of chromosome segregation • chromosome segregation Primary – Cluster 2 • positive regulation of apoptotic signaling pathway • negative regulation of insulin-like growth factor receptor signaling pathway	Primary – Cluster 1 • cell division • mitotic spindle organization • G2/M transition of mitotic cell cycle • mitotic cytokinesis • positive regulation of chromosome segregation • mitotic centrosome separation • DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest • mitotic spindle assembly checkpoint • regulation of chromosome organization • cell proliferation • positive regulation of mitotic nuclear division Primary – Cluster 2 • negative regulation of double-strand break repair via nonhomologous end joining • positive regulation of brain-derived neurotrophic factor receptor signaling pathway

• negative regulation of myelination • mitotic cell cycle arrest • positive regulation of microglial cell activation	

Bleomycin sulfate			
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**	**Cancer genes set | All Glioblastoma samples**	**Cancer genes set | Primary Glioblastoma samples**
<u>All - Cluster 12</u> • cell division • mitotic chromosome movement towards spindle pole • G1/S transition of mitotic cell cycle • retrograde vesicle-mediated transport, Golgi to ER • regulation of mitotic metaphase/anaphase transition • mitotic centrosome separation • mitotic cell cycle checkpoint • G2/M transition of mitotic cell cycle • positive regulation of somatic stem cell population maintenance • positive regulation of somatic stem cell division • positive regulation of DNA-dependent DNA replication initiation	<u>Primary – Cluster 27</u> • susceptibility to natural killer cell mediated cytotoxicity • susceptibility to T cell mediated cytotoxicity • negative regulation of macrophage chemotaxis • positive regulation of natural killer cell mediated cytotoxicity directed against tumor cell target <u>Primary – Cluster 23</u> • positive regulation of fibroblast apoptotic process • negative regulation of ERK1 and ERK2 cascade • regulation of non-canonical Wnt signaling pathway <u>Primary – Cluster 15</u> • positive regulation of forebrain neuron differentiation • cell-cell adhesion in response to extracellular stimulus • acinar cell differentiation • hepatocyte cell migration • cardiac neuron differentiation	<u>All - Cluster 2</u> • negative regulation of ERK1 and ERK2 cascade • positive regulation of B cell proliferation • cell surface receptor signaling pathway • positive regulation of MAP kinase activity • negative regulation of neurotrophin TRK receptor signaling pathway • motor neuron migration	<u>Primary – Cluster 3</u> • negative regulation of ERK1 and ERK2 cascade • positive regulation of MAP kinase activity • receptor localization to synapse • angiogenesis

Bortezomib	
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**
<u>All - Cluster 1</u> • chemical synaptic transmission • memory • brain development	<u>Primary – Cluster 26</u> • calcium ion transmembrane transport • neuron development • brain development

• neurotransmitter transport • inhibitory synapse assembly • regulation of postsynaptic membrane potential • regulation of neuronal synaptic plasticity	• regulation of synaptic vesicle exocytosis • chemical synaptic transmission • neurotransmitter secretion • synaptic vesicle clustering • postsynaptic intermediate filament cytoskeleton organization • positive regulation of phospholipase C-activating G protein-coupled receptor signaling pathway • neurofilament bundle assembly • exocytic insertion of neurotransmitter receptor to postsynaptic membrane • regulation of postsynaptic membrane potential • neurotransmitter receptor internalization
Busulfan	
Unbiased genes set | Primary Glioblastoma samples	
Primary – Cluster 3 • Notch receptor processing, ligand-dependent	
Carfilzomib	
Unbiased genes set | Primary Glioblastoma samples	**Cancer genes set | Primary Glioblastoma samples**
Primary – Cluster 2 • regulation of macrophage migration inhibitory factor signaling pathway • positive regulation of phosphatidylinositol 3-kinase signaling • negative regulation of insulin-like growth factor receptor signaling pathway	Primary – Cluster 4 • cellular response to interleukin-4 • positive regulation of signal transduction by p53 class mediator • cytokine-mediated signaling pathway • positive regulation of programmed cell death • protein phosphorylation • positive regulation of protein kinase B signaling • cell migration • presynaptic modulation of chemical synaptic transmission • activation of transmembrane receptor protein tyrosine kinase activity • positive regulation of T cell receptor signaling pathway • negative regulation of extrinsic apoptotic signaling pathway in absence of ligand
Carmustine	
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**
All - Cluster 2 • cellular response to topologically incorrect protein • protein folding • negative regulation of protein localization to endosome • regulation of mRNA splicing, via spliceosome • mRNA export from nucleus	Primary – Cluster 3 • DNA ligation involved in DNA recombination • double-strand break repair via classical nonhomologous end joining • neutrophil clearance • negative regulation of dendritic cell apoptotic process

• cellular protein-containing complex assembly • protein folding in endoplasmic reticulum All - Cluster 3 • ER to Golgi vesicle-mediated transport • Golgi organization • positive regulation of protein targeting to mitochondrion • DNA modification • vesicle fusion with Golgi apparatus • autophagy • mitochondrial mRNA processing • retrograde transport, endosome to Golgi • vesicle targeting, rough ER to cis-Golgi	• single strand break repair • DNA ligation involved in DNA repair Primary – Cluster 5 • positive regulation of TORC1 signaling Primary – Cluster 1 • regulation of ERK1 and ERK2 cascade Primary – Cluster 6 • negative regulation of myeloid progenitor cell differentiation

Celecoxib
Cancer genes set | Primary Glioblastoma samples
Primary – Cluster 3 • negative regulation of ERK1 and ERK2 cascade • cell-matrix adhesion • negative regulation of neurotrophin TRK receptor signaling pathway • positive regulation of vascular endothelial growth factor receptor signaling pathway • positive regulation of vasculogenesis • epithelial cell development • positive regulation of blood vessel endothelial cell migration • positive regulation of endothelial cell proliferation

Dactinomycin			
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**	**Cancer genes set | All Glioblastoma samples**	**Cancer genes set | Primary Glioblastoma samples**
All - Cluster 16 • RNA processing • negative regulation of non-canonical Wnt signaling pathway All - Cluster 13 • nucleosome assembly	Primary – Cluster 27 • positive regulation of central B cell tolerance induction • basophil homeostasis • eosinophil homeostasis • positive regulation of blood vessel diameter • monocyte homeostasis	All - Cluster 3 • cellular response to retinoic acid • mesenchymal cell differentiation	Primary – Cluster 5 • positive regulation of I-kappaB kinase/NF-kappaB signaling

• DNA replication-dependent nucleosome assembly • double-strand break repair via nonhomologous end joining • telomere capping • DNA replication initiation • negative regulation of cell cycle checkpoint • telomere organization • G1/S transition of mitotic cell cycle	<u>Primary – Cluster 32</u> • negative regulation of epidermal growth factor receptor signaling pathway • signal transduction involved in G2 DNA damage checkpoint • regulation of phosphatidylinositol 3-kinase activity • autophagy • negative regulation of cysteine-type endopeptidase activity involved in apoptotic process • double-strand break repair via homologous recombination • regulation of G1/S transition of mitotic cell cycle • regulation of TORC1 signaling <u>Primary – Cluster 24</u> • positive regulation of interleukin-2 biosynthetic process • negative regulation of retinoic acid receptor signaling pathway • negative regulation of cAMP-dependent protein kinase activity • regulation of cell-cell adhesion mediated by integrin • interleukin-2 secretion • retinoic acid metabolic process <u>Primary – Cluster 28</u> • positive regulation of hippocampal neuron apoptotic process • positive regulation of microglial cell mediated cytotoxicity • synapse pruning • negative regulation of long-term synaptic potentiation • neutrophil degranulation • positive regulation of macrophage fusion • positive regulation of receptor localization to synapse • regulation of tumor necrosis factor biosynthetic process • negative regulation of interleukin-1 beta production • macrophage activation • immune response-inhibiting signal transduction		• G protein-coupled receptor signaling pathway • lymphocyte differentiation • synaptic transmission, glutamatergic • leukocyte degranulation

	• regulation of microglial cell migration • defense response to virus • cytokine-mediated signaling pathway • cellular response to chemokine • regulation of T cell proliferation <u>Primary – Cluster 17</u> • cellular response to tumor necrosis factor • positive regulation of double-strand break repair via nonhomologous end joining <u>Primary – Cluster 4</u> • negative regulation of Rho-dependent protein serine/threonine kinase activity • negative regulation of dendritic cell apoptotic process • positive regulation of natural killer cell differentiation <u>Primary – Cluster 29</u> • positive regulation of double-strand break repair via nonhomologous end joining • protein localization to site of double-strand break <u>Primary – Cluster 15</u> • neurotransmitter secretion • hepatocyte cell migration • cell-cell adhesion in response to extracellular stimulus • membrane to membrane docking • regulation of transcription involved in lymphatic endothelial cell fate commitment <u>Primary – Cluster 21</u> • regulation of T cell proliferation • response to tumor necrosis factor • dendrite regeneration <u>Primary – Cluster 16</u> • chromatin silencing at rDNA • DNA replication-dependent nucleosome assembly		

	• nucleosome assembly • CENP-A containing nucleosome assembly • telomere capping • double-strand break repair via nonhomologous end joining • interleukin-7-mediated signaling pathway • nucleotide-excision repair, DNA gap filling • DNA-templated transcription, initiation • antibacterial humoral response • telomere maintenance via semi-conservative replication • antimicrobial humoral immune response mediated by antimicrobial peptide • negative regulation of stem cell differentiation • negative regulation of DNA recombination		

Dasatinib
Cancer genes set | Primary Glioblastoma samples
Primary – Cluster 6 • transmission of nerve impulse • positive regulation of neurological system process • positive regulation of neuron projection regeneration • positive regulation of phagocytosis • regulation of excitatory synapse assembly • negative regulation of astrocyte differentiation • postsynapse assembly • myelination in peripheral nervous system

Dexrazoxane		
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**	**Cancer genes set | All Glioblastoma samples**
All - Cluster 4 • negative regulation of vascular endothelial growth factor receptor signaling pathway • regulation of microglial cell migration • negative regulation of neuron apoptotic process	Primary – Cluster 25 • cell division • mitotic cytokinesis • DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest • G2/M transition of mitotic cell cycle • regulation of signal transduction by p53 class mediator • positive regulation of DNA endoreduplication	All - Cluster 2 • negative regulation of transforming growth factor beta production • negative regulation of ERK1 and ERK2 cascade • regulation of cellular response to growth factor stimulus

• positive regulation of protein kinase B signaling • cytosolic calcium signaling involved in initiation of cell movement in glial-mediated radial cell mi... • negative regulation of phosphatidylinositol 3-kinase activity • positive regulation of interleukin-4 biosynthetic process	Primary – Cluster 7 • axonogenesis • epithelial cell morphogenesis Primary – Cluster 6 • positive regulation of Wnt signaling pathway • regulation of presynaptic cytosolic calcium ion concentration • regulation of Wnt signaling pathway, planar cell polarity pathway • positive regulation of non-canonical Wnt signaling pathway • negative regulation of cell cycle arrest Primary – Cluster 8 • regulation of postsynaptic cytosolic calcium ion concentration • negative regulation of necrotic cell death • negative regulation of cell cycle arrest • regulation of necroptotic process • regulation of necrotic cell death	• positive regulation of protein tyrosine kinase activity

Doxorubicin hydrochloride		
Unbiased genes set | Primary Glioblastoma samples	**Cancer genes set | All Glioblastoma samples**	**Cancer genes set | Primary Glioblastoma samples**
Primary – Cluster 12 • negative regulation of intrinsic apoptotic signaling pathway in response to hydrogen peroxide • cytoplasm protein quality control by the ubiquitin-proteasome system • canonical Wnt signaling pathway involved in regulation of cell proliferation • base-excision repair, gap-filling	All - Cluster 3 • negative regulation of B cell differentiation • regulation of transcription, DNA-templated • cell cycle arrest • regulation of cell proliferation • positive regulation of erythrocyte differentiation • positive regulation of extrinsic apoptotic signaling pathway in absence of ligand • cellular response to stress All - Cluster 4 • epidermal cell differentiation	Primary – Cluster 7 • cell cycle arrest • regulation of neuron differentiation • regulation of Ras protein signal transduction • protein dephosphorylation • cellular senescence • Wnt signaling pathway • positive regulation of transcription of Notch receptor target • positive regulation of Notch signaling pathway Primary – Cluster 6 • positive regulation of extrinsic apoptotic signaling pathway in absence of ligand • regulation of macrophage differentiation • extrinsic apoptotic signaling pathway

		• positive regulation of cyclin-dependent protein serine/threonine kinase activity • neuron differentiation • negative regulation of cell growth • negative regulation of extrinsic apoptotic signaling pathway
Enzalutamide		
Unbiased genes set | Primary Glioblastoma samples		
Primary – Cluster 16 • DNA replication-dependent nucleosome assembly • chromatin silencing at rDNA • nucleosome assembly • CENP-A containing nucleosome assembly • telomere capping • double-strand break repair via nonhomologous end joining • DNA-templated transcription, initiation • DNA repair		
Epirubicin hydrochloride		
Unbiased genes set | Primary Glioblastoma samples	**Cancer genes set | All Glioblastoma samples**	**Cancer genes set | Primary Glioblastoma samples**
Primary – Cluster 12 • neurotransmitter receptor biosynthetic process • negative regulation of intrinsic apoptotic signaling pathway in response to hydrogen peroxide • negative regulation of synaptic transmission, cholinergic • canonical Wnt signaling pathway involved in regulation of cell proliferation • positive regulation of mitotic cell cycle spindle assembly checkpoint • positive regulation of cell cycle checkpoint • regulation of spindle checkpoint • positive regulation of B cell differentiation	All - Cluster 5 • positive regulation of cell proliferation involved in heart morphogenesis • positive regulation of DNA catabolic process • regulation of apoptotic process • glial cell fate commitment • regulation of mitotic spindle assembly • mitotic G2 DNA damage checkpoint All - Cluster 3 • negative regulation of B cell differentiation • positive regulation of neuron differentiation • mesenchymal cell differentiation • cellular response to retinoic acid • negative regulation of intracellular signal transduction	Primary – Cluster 7 • cell cycle arrest • regulation of Ras protein signal transduction • regulation of neuron differentiation • cellular senescence • positive regulation of transcription of Notch receptor target • centrosome localization
Floxuridine		
Cancer genes set | Primary Glioblastoma samples		

Primary – Cluster 4
- negative regulation of GTPase activity
- retinoic acid metabolic process

Fludarabine phosphate

Unbiased genes set | Primary Glioblastoma samples

Primary – Cluster 34
- positive regulation of endothelial cell proliferation
- phosphatidylinositol 3-kinase signaling
- positive regulation of blood vessel diameter
- positive regulation of blood vessel endothelial cell migration
- positive regulation of phosphorylation
- negative regulation of phosphatidylinositol biosynthetic process
- cellular response to growth factor stimulus
- leukocyte migration

Primary – Cluster 27
- susceptibility to natural killer cell mediated cytotoxicity
- susceptibility to T cell mediated cytotoxicity
- cell-cell adhesion via plasma-membrane adhesion molecules
- positive regulation of DNA damage response, signal transduction by p53 class mediator resulting in t...
- positive regulation of type B pancreatic cell apoptotic process
- positive regulation of natural killer cell mediated cytotoxicity directed against tumor cell target

Primary – Cluster 6
- regulation of eIF2 alpha phosphorylation by dsRNA
- negative regulation of myeloid progenitor cell differentiation

Fluorouracil

Unbiased genes set | All Glioblastoma samples

All - Cluster 5
- intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress
- regulation of autophagy
- intrinsic apoptotic signaling pathway in response to oxidative stress
- negative regulation of protein kinase B signaling
- positive regulation of oligodendrocyte apoptotic process

Gefitinib

Unbiased genes set | Primary Glioblastoma samples	Cancer genes set | Primary Glioblastoma samples
Primary – Cluster 6	Primary – Cluster 8

• positive regulation of Wnt signaling pathway, planar cell polarity pathway • positive regulation of vascular endothelial cell proliferation • positive regulation of endothelial cell apoptotic process • regulation of vascular endothelial cell proliferation • vascular endothelial cell proliferation • positive regulation of epithelial cell apoptotic process • regulation of endothelial cell apoptotic process • cellular response to epidermal growth factor stimulus	• macrophage activation • regulation of response to DNA damage stimulus • DNA unwinding involved in DNA replication • telomere maintenance via recombination • regulation of T cell differentiation in thymus • DNA replication initiation
Gemcitabine hydrochloride	
Unbiased genes set | All Glioblastoma samples	
All - Cluster 9 • central nervous system myelination • negative regulation of neuron differentiation • regulation of synaptic vesicle fusion to presynaptic active zone membrane • positive regulation of glial cell differentiation • positive regulation of myelination • myelination • tissue regeneration • oligodendrocyte differentiation • neuron fate specification • spinal cord motor neuron differentiation • regulation of myelination • positive regulation of gliogenesis • positive regulation of G protein-coupled receptor signaling pathway	
Hydroxyurea	
Unbiased genes set | Primary Glioblastoma samples	
Primary – Cluster 1 • positive regulation of protein localization to synapse • negative regulation of autophagosome assembly • regulation of presynapse assembly • regulation of presynapse organization • synaptic transmission, GABAergic	
Idarubicin hydrochloride	
Unbiased genes set | Primary Glioblastoma samples	**Cancer genes set | Primary Glioblastoma samples**
Primary – Cluster 20 • cell-matrix adhesion	Primary – Cluster 3 • negative regulation of ERK1 and ERK2 cascade

• extracellular matrix organization • branching involved in blood vessel morphogenesis • blood vessel development • negative regulation of angiogenesis • endothelial cell differentiation Primary – Cluster 18 • base-excision repair, base-free sugar-phosphate removal • telomere maintenance via base-excision repair Primary – Cluster 22 • positive regulation of retinoic acid biosynthetic process • negative regulation of immature T cell proliferation in thymus	• negative regulation of neurotrophin TRK receptor signaling pathway • negative regulation of GTPase activity • establishment of mitotic spindle orientation • positive regulation of MAP kinase activity • protein localization to postsynaptic membrane • positive regulation of excitatory postsynaptic potential

Imatinib			
Unbiased genes set | Primary Glioblastoma samples			
Primary – Cluster 16 • chromatin silencing at rDNA • DNA replication-dependent nucleosome assembly • telomere organization • transcription, RNA-templated • rRNA processing			
Ixabepilone			
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**	**Cancer genes set | All Glioblastoma samples**	**Cancer genes set | Primary Glioblastoma samples**
All - Cluster 9 • regulation of ARF protein signal transduction • negative regulation of epithelial to mesenchymal transition • positive regulation of ARF protein signal transduction • negative regulation of Wnt signaling pathway involved in dorsal/ventral axis specification	Primary – Cluster 20 • positive regulation of angiogenesis • extracellular matrix organization • negative regulation of endodermal cell differentiation • anatomical structure formation involved in morphogenesis • cell adhesion • cell surface receptor signaling pathway	All - Cluster 4 • nucleotide-excision repair, DNA duplex unwinding • global genome nucleotide-excision repair • nucleotide-excision repair, preincision complex assembly • regulation of G2/M transition of mitotic cell cycle • Wnt signaling pathway • nucleotide-excision repair, DNA incision, 5'-to lesion	Primary – Cluster 9 • interleukin-21-mediated signaling pathway • interleukin-4-mediated signaling pathway • interleukin-9-mediated signaling pathway • regulation of epithelial cell proliferation • interleukin-2-mediated signaling pathway • negative regulation of T cell differentiation

• negative regulation of canonical Wnt signaling pathway involved in controlling type B pancreatic cel... All - Cluster 10 • epidermal growth factor receptor signaling pathway	• positive regulation of focal adhesion assembly • blood vessel development	• transcription-coupled nucleotide-excision repair • nucleotide-excision repair, DNA damage recognition	• interleukin-15-mediated signaling pathway • interleukin-7-mediated signaling pathway

Lapatinib	
Unbiased genes set | Primary Glioblastoma samples	
Primary – Cluster 6 • common myeloid progenitor cell proliferation • myeloid progenitor cell differentiation • positive regulation of Wnt signaling pathway, planar cell polarity pathway	
Lomustine	
Unbiased genes set | Primary Glioblastoma samples	**Cancer genes set | Primary Glioblastoma samples**
Primary – Cluster 9 • positive regulation of NF-kappaB transcription factor activity • regulation of postsynaptic neurotransmitter receptor internalization • postsynaptic neurotransmitter receptor internalization • postsynaptic endocytosis • neurotransmitter receptor internalization • lymph vessel development Primary – Cluster 10 • myelination in peripheral nervous system • hematopoietic progenitor cell differentiation • negative regulation of cell adhesion • regulation of cell adhesion • hemopoiesis • hematopoietic or lymphoid organ development Primary – Cluster 26 • calcium ion transmembrane transport • spontaneous neurotransmitter secretion • regulation of synaptic vesicle fusion to presynaptic active zone membrane	Primary – Cluster 10 • cell adhesion • negative regulation of STAT cascade • positive regulation of gene expression • immune response • regulation of blood vessel endothelial cell migration • regulation of epidermal cell differentiation Primary – Cluster 3 • positive regulation of CREB transcription factor activity • negative regulation of ERK1 and ERK2 cascade • activation of cysteine-type endopeptidase activity involved in apoptotic process • negative regulation of neurotrophin TRK receptor signaling pathway • regulation of axonogenesis • negative regulation of GTPase activity • positive regulation of p38MAPK cascade • negative regulation of cell growth

• synaptic vesicle clustering • potassium ion transport • exocytic insertion of neurotransmitter receptor to postsynaptic membrane • synaptic vesicle exocytosis • calcium ion export across plasma membrane • calcium ion-regulated exocytosis of neurotransmitter • synapse assembly • positive regulation of excitatory postsynaptic potential • regulation of ion transmembrane transport <u>Primary – Cluster 7</u> • positive regulation of TOR signaling	
Mechlorethamine hydrochloride	
Unbiased genes set | All Glioblastoma samples	
<u>All - Cluster 11</u> • extracellular matrix organization • positive regulation of cytokine biosynthetic process • angiogenesis • negative regulation of intracellular signal transduction • positive regulation of angiogenesis • negative regulation of apoptotic process • negative regulation of cell adhesion • cell adhesion • cell activation • regulation of endothelial cell apoptotic process	
Melphalan hydrochloride	
Unbiased genes set | Primary Glioblastoma samples	**Cancer genes set | Primary Glioblastoma samples**
<u>Primary – Cluster 26</u> • neurotransmitter secretion • calcium ion transmembrane transport • regulation of synaptic vesicle fusion to presynaptic active zone membrane • synaptic vesicle clustering • glutamate secretion • exocytic insertion of neurotransmitter receptor to postsynaptic membrane • sodium ion transmembrane transport	<u>Primary – Cluster 3</u> • negative regulation of neurotrophin TRK receptor signaling pathway • regulation of signaling receptor activity • negative regulation of GTPase activity • regulation of axonogenesis • establishment of mitotic spindle orientation • positive regulation of CREB transcription factor activity • positive regulation of ion transport • positive regulation of excitatory postsynaptic potential • activation of MAPKKK activity

• calcium ion-regulated exocytosis of neurotransmitter • spontaneous neurotransmitter secretion • synapse assembly • positive regulation of excitatory postsynaptic potential <u>Primary – Cluster 22</u> • Notch signaling pathway • positive regulation of retinoic acid biosynthetic process • regulation of Fas signaling pathway • positive regulation of Wnt signaling pathway by BMP signaling pathway • negative regulation of immature T cell proliferation in thymus <u>Primary – Cluster 9</u> • positive regulation of NF-kappaB transcription factor activity • intrinsic apoptotic signaling pathway by p53 class mediator • positive regulation of stress-activated MAPK cascade • synaptic vesicle transport • positive regulation of stress-activated protein kinase signaling cascade • establishment of synaptic vesicle localization	• mitotic cell cycle arrest

Methotrexate			
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**	**Cancer genes set | All Glioblastoma samples**	**Cancer genes set | Primary Glioblastoma samples**
<u>All - Cluster 6</u> • positive regulation of angiogenesis • vasculogenesis • defense response to tumor cell • negative regulation of angiogenesis • angiotensin maturation • negative regulation of blood vessel endothelial cell migration • vascular endothelial growth factor signaling pathway	<u>Primary – Cluster 10</u> • negative regulation of T cell extravasation • negative regulation of NK T cell proliferation • negative regulation of CD8-positive, alpha-beta T cell differentiation • negative regulation of T-helper 17 type immune response • interleukin-17 secretion • negative regulation of interleukin-17 secretion	<u>All - Cluster 6</u> • positive regulation of ERK1 and ERK2 cascade • extracellular matrix organization • blood vessel remodeling • glomerular mesangial cell development • vascular endothelial growth factor signaling pathway • positive regulation of endothelial cell migration	<u>Primary – Cluster 10</u> • negative regulation of cell migration • transmembrane receptor protein tyrosine kinase signaling pathway • neutrophil degranulation • neutrophil mediated immunity • response to ammonium ion

• phosphatidylinositol 3-kinase signaling • cell-matrix adhesion • negative regulation of endothelial cell differentiation • blood vessel maturation • blood vessel remodeling • positive regulation of endothelial cell differentiation • lymphangiogenesis • positive regulation of phosphoprotein phosphatase activity • negative regulation of androgen receptor signaling pathway	• negative regulation of T-helper 17 cell differentiation • negative regulation of tumor necrosis factor secretion • NK T cell differentiation • negative regulation of double-strand break repair via homologous recombination • negative regulation of type 2 immune response <u>Primary – Cluster 34</u> • venous blood vessel development • vasculogenesis • transforming growth factor beta receptor signaling pathway • negative regulation of angiogenesis • endothelium development • negative regulation of Rho-dependent protein serine/threonine kinase activity • cell adhesion • extracellular matrix organization	• positive regulation of endothelial cell proliferation • positive regulation of angiogenesis	
Methoxsalen			
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**	**Cancer genes set | All Glioblastoma samples**	**Cancer genes set | Primary Glioblastoma samples**
<u>All - Cluster 4</u> • regulation of microglial cell migration • mature B cell differentiation involved in immune response • leukocyte activation • immune response • positive regulation of neuroblast proliferation	<u>Primary – Cluster 11</u> • regulation of release of cytochrome c from mitochondria • mitochondrial nucleoid organization • positive regulation of mitochondrial transcription • negative regulation of mitochondrial membrane permeability involved in apoptotic process	<u>All - Cluster 2</u> • negative regulation of neurotrophin TRK receptor signaling pathway • negative regulation of ERK1 and ERK2 cascade • positive regulation of CREB transcription factor activity • establishment of mitotic spindle orientation	<u>Primary – Cluster 3</u> • positive regulation of CREB transcription factor activity • negative regulation of angiogenesis • negative regulation of neurotrophin TRK receptor signaling pathway • regulation of cysteine-type endopeptidase activity

• regulation of vascular endothelial growth factor receptor signaling pathway • negative regulation of autophagic cell death • positive regulation of CD40 signaling pathway • negative regulation of phosphatidylinositol 3-kinase activity <u>All - Cluster 5</u> • regulation of signaling receptor activity • positive regulation of MAP kinase activity • nucleotide-excision repair, DNA damage recognition • regulation of transcription from RNA polymerase II promoter in response to hypoxia • regulation of postsynaptic specialization assembly • nucleotide-excision repair, DNA duplex unwinding • global genome nucleotide-excision repair • nucleotide-excision repair, preincision complex assembly • positive regulation of syncytium formation by plasma membrane fusion <u>All - Cluster 7</u> • neutrophil degranulation • positive regulation of tumor necrosis factor production • macrophage differentiation	• positive regulation of cytochrome-c oxidase activity • inhibition of cysteine-type endopeptidase activity involved in apoptotic process • negative regulation of hypoxia-induced intrinsic apoptotic signaling pathway • DNA damage induced protein phosphorylation <u>Primary – Cluster 36</u> • Golgi vesicle prefusion complex stabilization • positive regulation of single-stranded telomeric DNA binding • vesicle-mediated intercellular transport • vascular endothelial growth factor receptor-1 signaling pathway • telomere assembly • negative regulation of double-strand break repair via nonhomologous end joining • negative regulation of vascular endothelial cell proliferation • cell-cell adhesion via plasma-membrane adhesion molecules <u>Primary – Cluster 12</u> • regulation of cell cycle arrest • positive regulation of cell cycle process • cellular respiration • neurotransmitter receptor biosynthetic process • B cell receptor apoptotic signaling pathway • B cell negative selection • regulation of mitochondrial membrane permeability involved in programmed necrotic cell death	<u>All - Cluster 1</u> • cellular response to calcium ion • positive regulation of GTPase activity • Ras protein signal transduction • regulation of Rho protein signal transduction • regulation of apoptotic process	involved in apoptotic process • epidermal growth factor receptor signaling pathway • negative regulation of neuron differentiation • activation of protein kinase activity • establishment of mitotic spindle orientation

• negative regulation of B cell proliferation • positive regulation of T cell proliferation • positive regulation of cytokine secretion • phagocytosis, engulfment • positive regulation of B cell differentiation • cellular response to macrophage colony-stimulating factor stimulus • positive regulation of defense response to bacterium • leukocyte migration • inflammatory response • regulation of neutrophil differentiation • B cell receptor signaling pathway • receptor-mediated endocytosis	• release of matrix enzymes from mitochondria • negative regulation of intrinsic apoptotic signaling pathway in response to hydrogen peroxide <u>Primary – Cluster 28</u> • neutrophil degranulation • innate immune response • positive regulation of macrophage fusion • negative regulation of leukocyte apoptotic process • toll-like receptor 2 signaling pathway • T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenti... • negative regulation of tumor necrosis factor production • inflammatory response • defense response to virus <u>Primary – Cluster 13</u> • inflammatory response • peptide antigen assembly with MHC class II protein complex • neutrophil degranulation • MyD88-dependent toll-like receptor signaling pathway • positive regulation of T cell proliferation • negative regulation of interleukin-6 production • antigen processing and presentation of exogenous peptide antigen via MHC class II • positive regulation of B cell differentiation • negative regulation of immune effector process • regulation of neutrophil differentiation		

	• polysaccharide assembly with MHC class II protein complex • innate immune response • regulation of lymphocyte mediated immunity • lymphocyte differentiation • regulation of adaptive immune response based on somatic recombination of immune receptors built from... • phagocytosis		

Mitomycin	
Unbiased genes set | Primary Glioblastoma samples	**Cancer genes set | Primary Glioblastoma samples**
Primary – Cluster 9 • positive regulation of aorta morphogenesis • positive regulation of somatic stem cell population maintenance • positive regulation of somatic stem cell division • positive regulation of mammary stem cell proliferation • positive regulation of NF-kappaB transcription factor activity • ectodermal cell differentiation • venous blood vessel morphogenesis • mammary gland epithelial cell differentiation • positive regulation of cardiac muscle cell differentiation • negative regulation of stem cell differentiation • positive regulation of cardiocyte differentiation	Primary – Cluster 3 • regulation of signaling receptor activity • negative regulation of cell proliferation • negative regulation of angiogenesis • positive regulation of CREB transcription factor activity • epithelial cell differentiation • positive regulation of cell proliferation • cell morphogenesis involved in differentiation • negative regulation of ERK1 and ERK2 cascade • positive regulation of protein phosphorylation • cell-cell signaling • positive regulation of integrin biosynthetic process • negative regulation of receptor biosynthetic process

Mitoxantrone	
Unbiased genes set | Primary Glioblastoma samples	**Cancer genes set | Primary Glioblastoma samples**
Primary – Cluster 25 • cell division • DNA replication • attachment of mitotic spindle microtubules to kinetochore • anaphase-promoting complex-dependent catabolic process • regulation of attachment of spindle microtubules to kinetochore • regulation of transcription involved in G1/S transition of mitotic cell cycle • mitotic spindle assembly checkpoint • cell cycle	Primary – Cluster 1 • cell division • mitotic spindle assembly checkpoint • DNA replication • cell proliferation • DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest • anaphase-promoting complex-dependent catabolic process • regulation of mitotic nuclear division • DNA damage induced protein phosphorylation

• DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest • regulation of mitotic cell cycle spindle assembly checkpoint • regulation of chromosome segregation • positive regulation of exit from mitosis • regulation of G2/M transition of mitotic cell cycle • mitotic spindle organization • mitotic sister chromatid cohesion • DNA biosynthetic process • mitotic sister chromatid segregation	• positive regulation of chromosome segregation • G2/M transition of mitotic cell cycle • mitotic centrosome separation • chromatin remodeling • replicative senescence • regulation of signal transduction by p53 class mediator • regulation of transcription, DNA-templated • mitotic spindle assembly
Nilotinib	
Unbiased genes set | Primary Glioblastoma samples	
Primary – Cluster 9 • positive regulation of protein autophosphorylation • ARF protein signal transduction • regulation of ARF protein signal transduction • negative regulation of insulin receptor signaling pathway • negative regulation of cellular response to insulin stimulus • regulation of protein autophosphorylation • negative regulation of mitotic nuclear division • negative regulation of nuclear division • regulation of insulin receptor signaling pathway • regulation of cellular response to insulin stimulus • insulin receptor signaling pathway • regulation of mitotic nuclear division Primary – Cluster 22 • regulation of calcium import into the mitochondrion • establishment of glial blood-brain barrier • response to aluminum ion • substrate-dependent cell migration, cell attachment to substrate • response to selenium ion • Notch receptor processing, ligand-dependent	
Omacetaxine mepesuccinate	
Cancer genes set | Primary Glioblastoma samples	

Primary – Cluster 4
- regulation of signaling
- phosphatidylserine acyl-chain remodeling
- phosphatidylglycerol acyl-chain remodeling
- phosphatidylinositol acyl-chain remodeling
- positive regulation of DNA metabolic process
- mammary gland morphogenesis
- phosphatidylcholine acyl-chain remodeling
- phosphatidylethanolamine acyl-chain remodeling

Oxaliplatin		
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**	**Cancer genes set | Primary Glioblastoma samples**
All - Cluster 1 • neural retina development • neuron development • long-term synaptic potentiation • regulation of short-term neuronal synaptic plasticity • brain development • modulation of excitatory postsynaptic potential	Primary – Cluster 21 • positive regulation of glial cell-derived neurotrophic factor secretion Primary – Cluster 26 • glutamate secretion • calcium ion transmembrane transport • regulation of synaptic vesicle exocytosis • neurotransmitter secretion • neuron development • synaptic vesicle clustering • exocytic insertion of neurotransmitter receptor to postsynaptic membrane • positive regulation of phospholipase C-activating G protein-coupled receptor signaling pathway • positive regulation of dendrite extension • sodium ion transmembrane transport • regulation of synapse assembly • calcium ion export across plasma membrane • calcium ion-regulated exocytosis of neurotransmitter • spontaneous neurotransmitter secretion Primary – Cluster 9 • positive regulation of high voltage-gated calcium channel activity	Primary – Cluster 4 • regulation of neurotransmitter receptor localization to postsynaptic specialization membrane • negative regulation of GTPase activity • cell surface receptor signaling pathway • intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress • positive regulation of apoptotic signaling pathway • blood vessel remodeling • cell cycle checkpoint

	• synaptic vesicle cycle • ectodermal cell differentiation • synaptic vesicle priming • regulation of synaptic vesicle recycling • positive regulation of protein autophosphorylation • regulation of protein autophosphorylation • endocytic recycling • negative regulation of mitotic nuclear division • synaptic vesicle recycling • negative regulation of nuclear division • exocytic process • intrinsic apoptotic signaling pathway by p53 class mediator <u>Primary – Cluster 29</u> • negative regulation of stem cell differentiation • somatic stem cell division • negative regulation of ectodermal cell fate specification • regulation of hematopoietic stem cell differentiation • stem cell population maintenance • negative regulation of fat cell differentiation • positive regulation of epithelial to mesenchymal transition • regulation of T cell homeostatic proliferation	
Paclitaxel		
Unbiased genes set | Primary Glioblastoma samples		
<u>Primary – Cluster 26</u> • positive regulation of phospholipase C-activating G protein-coupled receptor signaling pathway • positive regulation of excitatory postsynaptic potential		
Pazopanib hydrochloride		
Unbiased genes set | Primary Glioblastoma samples		
<u>Primary – Cluster 17</u> • regulation of type III interferon production • positive regulation of interferon-beta secretion • positive regulation of interferon-gamma-mediated signaling pathway • cellular response to interferon-gamma • positive regulation of interferon-alpha secretion • positive regulation of double-strand break repair via nonhomologous end joining		

- positive regulation of neuron migration
- positive regulation of tumor necrosis factor-mediated signaling pathway

Primary – Cluster 4
- DNA ligation involved in DNA recombination
- negative regulation of Rho-dependent protein serine/threonine kinase activity
- double-strand break repair via classical nonhomologous end joining
- single strand break repair
- DNA ligation involved in DNA repair

Pemetrexed

Unbiased genes set | All Glioblastoma samples

All - Cluster 6
- vasculogenesis
- vascular endothelial growth factor signaling pathway
- blood vessel remodeling
- positive regulation of angiogenesis
- blood vessel maturation
- lymphatic endothelial cell differentiation
- regulation of cell proliferation
- lymphangiogenesis
- positive regulation of endothelial cell differentiation
- positive regulation of phosphoprotein phosphatase activity
- positive regulation of phosphatidylinositol 3-kinase signaling

Pipobroman

Unbiased genes set | Primary Glioblastoma samples	Cancer genes set | Primary Glioblastoma samples
Primary – Cluster 9 • positive regulation of somatic stem cell population maintenance • positive regulation of somatic stem cell division • positive regulation of mammary stem cell proliferation • ectodermal cell differentiation • positive regulation of NF-kappaB transcription factor activity • venous blood vessel morphogenesis • vesicle-mediated transport in synapse • mammary gland epithelial cell differentiation • positive regulation of cardiac muscle cell differentiation	Primary – Cluster 3 • negative regulation of ERK1 and ERK2 cascade • regulation of signaling receptor activity • activation of cysteine-type endopeptidase activity involved in apoptotic process • negative regulation of angiogenesis • negative regulation of neurotrophin TRK receptor signaling pathway • positive regulation of endothelial cell proliferation • negative regulation of GTPase activity • establishment of mitotic spindle orientation • positive regulation of CREB transcription factor activity

Plicamycin

<table>
<tr><th colspan="3">Cancer genes set | Primary Glioblastoma samples</th></tr>
<tr><td colspan="3"><u>Primary – Cluster 4</u>
• positive regulation of signal transduction by p53 class mediator
• positive regulation of MAP kinase activity
• inositol phosphate-mediated signaling
• negative regulation of cell death
• nucleotide-excision repair, DNA gap filling</td></tr>
<tr><th colspan="3">Ponatinib</th></tr>
<tr><th colspan="3">Unbiased genes set | Primary Glioblastoma samples</th></tr>
<tr><td colspan="3"><u>Primary – Cluster 30</u>
• vascular endothelial growth factor receptor signaling pathway
• negative regulation of mature B cell apoptotic process
• positive regulation of mast cell cytokine production</td></tr>
<tr><th colspan="3">Pralatrexate</th></tr>
<tr><th>Unbiased genes set | All Glioblastoma samples</th><th>Unbiased genes set | Primary Glioblastoma samples</th><th>Cancer genes set | Primary Glioblastoma samples</th></tr>
<tr><td><u>All - Cluster 11</u>
• extracellular matrix organization
• positive regulation of cell migration
• regulation of cell-substrate adhesion
• positive regulation of epithelial to mesenchymal transition
• endodermal cell differentiation
• regulation of inflammatory response
• extracellular matrix disassembly
• epithelial to mesenchymal transition involved in cardiac fibroblast development
• cell adhesion</td><td><u>Primary – Cluster 19</u>
• meiotic DNA double-strand break formation

<u>Primary – Cluster 17</u>
• cellular response to tumor necrosis factor

<u>Primary – Cluster 24</u>
• regulation of cell-cell adhesion mediated by integrin
• retinoic acid metabolic process

<u>Primary – Cluster 10</u>
• positive regulation of activated CD8-positive, alpha-beta T cell apoptotic process
• positive regulation of tolerance induction to tumor cell
• negative regulation of NK T cell proliferation
• negative regulation of CD8-positive, alpha-beta T cell activation
• negative regulation of CD8-positive, alpha-beta T cell differentiation
• positive regulation of synapse structural plasticity
• positive regulation of interleukin-10 secretion
• negative regulation of CD4-positive, alpha-beta T cell proliferation</td><td><u>Primary – Cluster 11</u>
• T cell differentiation involved in immune response
• positive regulation of epithelial to mesenchymal transition involved in endocardial cushion formatio...
• regulation of NK T cell differentiation

<u>Primary – Cluster 10</u>
• artery morphogenesis
• positive regulation of cell proliferation
• positive regulation of angiogenesis
• response to cytokine
• positive regulation of blood vessel endothelial cell migration</td></tr>
</table>

• substrate adhesion-dependent cell spreading	• NK T cell differentiation • negative regulation of interleukin-10 production • negative regulation of activated T cell proliferation <u>Primary – Cluster 9</u> • positive regulation of mammary stem cell proliferation • positive regulation of somatic stem cell population maintenance • positive regulation of somatic stem cell division • positive regulation of NF-kappaB transcription factor activity • regulation of postsynaptic neurotransmitter receptor internalization • mammary gland epithelial cell differentiation	
Procarbazine hydrochloride		
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**	**Cancer genes set | Primary Glioblastoma samples**
<u>All - Cluster 2</u> • protein folding in endoplasmic reticulum • regulation of G2/M transition of mitotic cell cycle • negative regulation of RNA splicing • negative regulation of mRNA processing	<u>Primary – Cluster 37</u> • double-strand break repair via nonhomologous end joining • regulation of transcription involved in G1/S transition of mitotic cell cycle <u>Primary – Cluster 14</u> • translational initiation • rRNA (guanine-N7)-methylation • ribosomal small subunit export from nucleus • rRNA processing • positive regulation of ribosome biogenesis • positive regulation of rRNA processing	<u>Primary – Cluster 11</u> • positive regulation of autophagy • proteasomal ubiquitin-independent protein catabolic process • DNA recombinase assembly • negative regulation of telomerase activity
Raloxifene		
Unbiased genes set | Primary Glioblastoma samples		
<u>Primary – Cluster 16</u> • DNA replication-dependent nucleosome assembly • chromatin silencing at rDNA • CENP-A containing nucleosome assembly • telomere capping • double-strand break repair via nonhomologous end joining • nucleosome assembly • re-entry into mitotic cell cycle		

- telomere organization

Romidepsin

Cancer genes set | Primary Glioblastoma samples

Primary – Cluster 9

- positive regulation of interleukin-2 biosynthetic process
- negative regulation of T cell differentiation
- regulation of megakaryocyte differentiation
- calcium ion transmembrane transport
- regulation of antigen receptor-mediated signaling pathway
- cell surface receptor signaling pathway
- cellular response to epidermal growth factor stimulus
- B cell receptor signaling pathway
- negative regulation of adaptive immune memory response
- telomeric heterochromatin assembly
- negative regulation of chromosome condensation
- positive regulation of CD8-positive, alpha-beta T cell differentiation

Sunitinib

Unbiased genes set | Primary Glioblastoma samples

Primary – Cluster 26

- commitment of neuronal cell to specific neuron type in forebrain
- neurotransmitter secretion
- positive regulation of excitatory postsynaptic potential
- neuron development
- synaptic vesicle clustering
- regulation of ion transmembrane transport
- positive regulation of dendrite extension
- negative regulation of G1/S transition of mitotic cell cycle by negative regulation of transcription...
- negative regulation of short-term neuronal synaptic plasticity
- positive regulation of ARF protein signal transduction

Primary – Cluster 15

- neuron migration
- regulation of postsynaptic neurotransmitter receptor activity
- positive regulation of cyclin-dependent protein serine/threonine kinase activity

Tamoxifen citrate

Unbiased genes set | Primary Glioblastoma samples

Primary – Cluster 4

- DNA ligation involved in DNA recombination
- mitochondrial proton-transporting ATP synthase complex assembly
- negative regulation of Rho-dependent protein serine/threonine kinase activity
- double-strand break repair via classical nonhomologous end joining
- single strand break repair
- DNA ligation involved in DNA repair
- pro-B cell differentiation
- lymphoid progenitor cell differentiation
- regulation of centrosome duplication
- nucleotide-excision repair, DNA gap filling

Temsirolimus

Unbiased genes set | Primary Glioblastoma samples

Primary – Cluster 7
- ubiquitin-dependent SMAD protein catabolic process
- oncogene-induced cell senescence
- negative regulation of toll-like receptor 5 signaling pathway

Primary – Cluster 20
- extracellular matrix organization
- cell-cell adhesion mediated by integrin
- regulation of cell adhesion mediated by integrin
- extracellular matrix disassembly

Thiotepa

Unbiased genes set | All Glioblastoma samples	Unbiased genes set | Primary Glioblastoma samples	Cancer genes set | Primary Glioblastoma samples
All - Cluster 8 • Golgi inheritance • Golgi localization • cytokine secretion involved in immune response • positive regulation of MAPK cascade	Primary – Cluster 9 • positive regulation of mitotic cell cycle, embryonic • negative regulation of stomach neuroendocrine cell differentiation • negative regulation of pancreatic A cell differentiation • negative regulation of inner ear auditory receptor cell differentiation • negative regulation of forebrain neuron differentiation • lateral inhibition	Primary – Cluster 3 • negative regulation of ERK1 and ERK2 cascade • regulation of signaling receptor activity • intestinal epithelial cell maturation • negative regulation of neurotrophin TRK receptor signaling pathway • negative regulation of GTPase activity • establishment of mitotic spindle orientation • positive regulation of excitatory postsynaptic potential

Topotecan hydrochloride		
Unbiased genes set | Primary Glioblastoma samples	**Cancer genes set | All Glioblastoma samples**	**Cancer genes set | Primary Glioblastoma samples**
Primary – Cluster 9 • negative regulation of stem cell differentiation • positive regulation of somatic stem cell population maintenance • positive regulation of somatic stem cell division • positive regulation of mammary stem cell proliferation • negative regulation of stomach neuroendocrine cell differentiation • positive regulation of mitotic cell cycle, embryonic • negative regulation of pro-B cell differentiation • forebrain radial glial cell differentiation Primary – Cluster 23 • negative regulation of sprouting angiogenesis • positive regulation of mesenchymal cell proliferation • regulation of cell migration involved in sprouting angiogenesis • negative regulation of endothelial cell migration • endothelial cell-matrix adhesion • negative regulation of phosphatidylinositol biosynthetic process Primary – Cluster 21 • negative regulation of intrinsic apoptotic signaling pathway Primary – Cluster 22 • positive regulation of Wnt signaling pathway by BMP signaling pathway Primary – Cluster 18 • negative regulation of oligodendrocyte apoptotic process Primary – Cluster 20 • transforming growth factor beta receptor signaling pathway • positive regulation of angiogenesis • angiogenesis • cell adhesion	All - Cluster 2 • negative regulation of ERK1 and ERK2 cascade • negative regulation of neurotrophin TRK receptor signaling pathway • negative regulation of protein kinase B signaling • positive regulation of oxidative stress-induced neuron death	Primary – Cluster 3 • negative regulation of cell proliferation • angiogenesis • negative regulation of ERK1 and ERK2 cascade Primary – Cluster 10 • negative regulation of apoptotic process • negative regulation of cell migration • positive regulation of angiogenesis • positive regulation of cell proliferation • positive regulation of blood vessel endothelial cell migration

• integrin-mediated signaling pathway		
Uracil mustard		
Unbiased genes set | All Glioblastoma samples	**Unbiased genes set | Primary Glioblastoma samples**	
All - Cluster 9 • central nervous system myelination • positive regulation of oligodendrocyte progenitor proliferation • positive regulation of glial cell differentiation • neuron fate specification	Primary – Cluster 33 • regulation of telomere maintenance • positive regulation of telomeric DNA binding • DNA repair • negative regulation of telomerase activity • negative regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway	
Valrubicin		
Unbiased genes set | Primary Glioblastoma samples	**Cancer genes set | Primary Glioblastoma samples**	
Primary – Cluster 13 • neutrophil degranulation • innate immune response • positive regulation of T cell proliferation • inflammatory response • immune response • interleukin-3 production • mast cell degranulation • negative regulation of immune response • regulation of leukocyte tethering or rolling • positive regulation of blood vessel endothelial cell proliferation involved in sprouting angiogenesi... • B cell receptor signaling pathway • activation of immune response • myeloid dendritic cell activation	Primary – Cluster 7 • positive regulation of transcription of Notch receptor target • insulin receptor signaling pathway • cellular response to epidermal growth factor stimulus • cell cycle arrest Primary – Cluster 6 • transmission of nerve impulse • regulation of presynapse assembly • detection of cell density by contact stimulus involved in contact inhibition • canonical Wnt signaling pathway involved in metanephric kidney development	
Vemurafenib		
Unbiased genes set | Primary Glioblastoma samples		
Primary – Cluster 31 • regulation of nucleotide-excision repair • DNA double-strand break processing involved in repair via single-strand annealing • G1/S transition of mitotic cell cycle		

Table 23: Enrichment analysis results derived from the EMC Drugs Repurposing Project WGCNA analysis results

8.3 Results from the RNA-Seq DEA

	pvalue	padj	geneName
ENSG00000249359.2	6.83E-42	2.33E-37	RP11-374A4.1
ENSG00000226521.7	1.04E-29	1.78E-25	AC126365.1
ENSG00000227713.1	1.24E-17	1.41E-13	AC116609.1
ENSG00000214883.4	7.33E-17	6.26E-13	RP11-574M7.2
ENSG00000261316.1	8.56E-15	5.85E-11	LINC01834
ENSG00000225366.4	4.82E-14	2.75E-10	TDGF1P3
ENSG00000217331.1	1.97E-13	9.62E-10	RP11-304C16.3
ENSG00000248131.5	8.28E-11	3.54E-07	LINC01194
ENSG00000267382.1	9.57E-11	3.63E-07	RP11-325K19.2
ENSG00000255047.1	4.71E-10	1.61E-06	HNRNPRP2
ENSG00000254979.5	1.42E-09	4.41E-06	RP11-872D17.8
ENSG00000277210.3	1.62E-09	4.62E-06	RP11-14C10.6
ENSG00000277762.1	3.09E-09	8.13E-06	RN7SL261P
ENSG00000214891.9	3.61E-09	8.82E-06	TRIM64C
ENSG00000160401.14	1.18E-08	2.69E-05	CFAP157
ENSG00000206738.1	1.54E-08	3.09E-05	Y_RNA
ENSG00000237770.2	1.52E-08	3.09E-05	SPATA31D2P
ENSG00000182329.12	2.06E-08	3.70E-05	KIAA2012
ENSG00000258752.1	2.00E-08	3.70E-05	RP11-356K23.1
ENSG00000154099.17	2.20E-08	3.76E-05	DNAAF1
ENSG00000234363.1	2.32E-08	3.78E-05	PPIAP27
ENSG00000199595.1	2.91E-08	4.53E-05	Y_RNA
ENSG00000181085.14	3.94E-08	5.86E-05	MAPK15
ENSG00000152611.11	5.64E-08	7.95E-05	CAPSL
ENSG00000167858.12	6.05E-08	7.95E-05	TEKT1
ENSG00000238926.1	5.93E-08	7.95E-05	Y_RNA
ENSG00000137473.17	7.74E-08	9.45E-05	TTC29
ENSG00000145491.11	7.66E-08	9.45E-05	ROPN1L
ENSG00000174844.14	8.28E-08	9.76E-05	DNAH12
ENSG00000117222.13	8.67E-08	9.88E-05	RBBP5
ENSG00000159625.14	1.25E-07	0.000136	DRC7
ENSG00000271070.1	1.27E-07	0.000136	GMCL1P2
ENSG00000156206.13	1.32E-07	0.000137	CFAP161
ENSG00000271580.1	2.24E-07	0.000225	RP11-536L3.4
ENSG00000206113.10	2.51E-07	0.000245	CFAP99
ENSG00000205959.3	3.27E-07	0.000311	RP11-689P11.2
ENSG00000243710.7	3.75E-07	0.000346	CFAP57
ENSG00000163885.11	4.01E-07	0.000358	CFAP100
ENSG00000237077.1	4.09E-07	0.000358	AC105399.2
ENSG00000131044.16	5.14E-07	0.000439	TTLL9

ENSG00000213612.3	5.28E-07	0.00044	FAM220CP
ENSG00000171595.13	6.54E-07	0.000532	DNAI2
ENSG00000166596.14	6.92E-07	0.00055	CFAP52
ENSG00000175267.14	7.20E-07	0.000559	VWA3A
ENSG00000146221.9	8.23E-07	0.000625	TCTE1
ENSG00000136918.7	9.01E-07	0.000628	WDR38
ENSG00000213085.9	8.50E-07	0.000628	CFAP45
ENSG00000265316.1	8.89E-07	0.000628	RP11-286N3.1
ENSG00000270765.5	8.95E-07	0.000628	GAS2L2
ENSG00000170893.3	1.11E-06	0.000761	TRH
ENSG00000226837.2	1.15E-06	0.000768	HMGB1P32
ENSG00000133665.12	1.28E-06	0.000843	DYDC2
ENSG00000152760.9	1.46E-06	0.000943	TCTEX1D1
ENSG00000169314.14	1.62E-06	0.001024	C22orf15
ENSG00000144031.11	1.70E-06	0.001039	ANKRD53
ENSG00000158816.15	1.76E-06	0.001039	VWA5B1
ENSG00000197385.5	1.76E-06	0.001039	ZNF860
ENSG00000279328.1	1.75E-06	0.001039	RP11-203H19.2
ENSG00000141744.3	1.99E-06	0.001136	PNMT
ENSG00000228611.2	1.99E-06	0.001136	HNF4GP1
ENSG00000139537.10	2.11E-06	0.001182	CCDC65
ENSG00000122735.15	2.17E-06	0.001194	DNAI1
ENSG00000182759.3	2.30E-06	0.001245	MAFA
ENSG00000183644.13	2.53E-06	0.001353	C11orf88
ENSG00000197816.13	2.71E-06	0.001425	CCDC180
ENSG00000155761.13	3.17E-06	0.001617	SPAG17
ENSG00000188523.8	3.15E-06	0.001617	CFAP77
ENSG00000170231.15	3.46E-06	0.001741	FABP6
ENSG00000248712.7	3.54E-06	0.001755	CCDC153
ENSG00000257057.2	3.85E-06	0.001879	C11orf97
ENSG00000162004.16	4.03E-06	0.00194	CCDC78
ENSG00000140057.8	4.19E-06	0.001991	AK7
ENSG00000186471.12	4.29E-06	0.002009	AKAP14
ENSG00000165164.13	4.45E-06	0.002043	CFAP47
ENSG00000179902.12	4.48E-06	0.002043	C1orf194
ENSG00000203799.12	4.58E-06	0.00206	CCDC162P
ENSG00000120051.14	4.97E-06	0.002206	CFAP58
ENSG00000164746.13	5.15E-06	0.002255	C7orf57
ENSG00000131771.13	6.70E-06	0.0029	PPP1R1B
ENSG00000135205.14	6.89E-06	0.002907	CCDC146
ENSG00000169064.12	6.87E-06	0.002907	ZBBX
ENSG00000173013.5	7.25E-06	0.003022	CCDC96
ENSG00000168658.18	7.35E-06	0.003028	VWA3B

ENSG00000110723.11	7.61E-06	0.003095	EXPH5
ENSG00000072858.10	8.23E-06	0.003247	SIDT1
ENSG00000114473.13	8.25E-06	0.003247	IQCG
ENSG00000197826.11	8.27E-06	0.003247	C4orf22
ENSG00000260259.1	8.49E-06	0.00326	LINC02166
ENSG00000268736.1	8.49E-06	0.00326	MTCO3P39
ENSG00000248399.1	9.17E-06	0.003484	RP11-503N18.4
ENSG00000162814.10	9.60E-06	0.003604	SPATA17
ENSG00000269984.1	1.00E-05	0.003717	RP11-362K14.5
ENSG00000158428.3	1.03E-05	0.003758	CATIP
ENSG00000230599.2	1.03E-05	0.003758	AC018495.3
ENSG00000092850.11	1.05E-05	0.003791	TEKT2
ENSG00000186529.15	1.09E-05	0.003883	CYP4F3
ENSG00000226644.5	1.18E-05	0.004174	RP11-128M1.1
ENSG00000160188.9	1.26E-05	0.004366	RSPH1
ENSG00000162148.10	1.26E-05	0.004366	PPP1R32
ENSG00000153347.9	1.28E-05	0.004385	FAM81B
ENSG00000157703.15	1.30E-05	0.004385	SVOPL
ENSG00000196666.4	1.33E-05	0.004397	FAM180B
ENSG00000237542.1	1.32E-05	0.004397	MTCO3P17
ENSG00000140795.12	1.35E-05	0.004442	MYLK3
ENSG00000215612.7	1.42E-05	0.004597	HMX1
ENSG00000230173.1	1.43E-05	0.004597	LINC01790
ENSG00000154479.12	1.49E-05	0.004745	CCDC173
ENSG00000163263.6	1.52E-05	0.004796	C1orf189
ENSG00000187942.11	1.55E-05	0.004849	LDLRAD2
ENSG00000008226.19	1.80E-05	0.005489	DLEC1
ENSG00000129654.7	1.80E-05	0.005489	FOXJ1
ENSG00000181780.4	1.77E-05	0.005489	OR5J1P
ENSG00000166535.19	1.93E-05	0.005664	A2ML1
ENSG00000175318.11	1.88E-05	0.005664	GRAMD2
ENSG00000187905.10	1.91E-05	0.005664	LRRC74B
ENSG00000197057.9	1.92E-05	0.005664	DTHD1
ENSG00000230062.5	1.94E-05	0.005664	ANKRD66
ENSG00000185681.12	2.00E-05	0.005803	MORN5
ENSG00000223197.1	2.06E-05	0.005929	RNU6-1001P
ENSG00000248844.6	2.09E-05	0.005951	RP11-626H12.3
ENSG00000269956.1	2.29E-05	0.006455	MKNK1-AS1
ENSG00000004838.13	2.44E-05	0.006725	ZMYND10
ENSG00000141294.9	2.46E-05	0.006725	LRRC46
ENSG00000146038.11	2.41E-05	0.006725	DCDC2
ENSG00000181656.6	2.44E-05	0.006725	GPR88
ENSG00000112183.14	2.59E-05	0.007009	RBM24

ENSG00000159712.10	2.62E-05	0.007009	ANKRD18CP
ENSG00000260266.1	2.63E-05	0.007009	CTD-2311M21.2
ENSG00000176601.12	2.78E-05	0.007307	MAP3K19
ENSG00000257296.1	2.77E-05	0.007307	RP11-701B6.1
ENSG00000158423.16	2.84E-05	0.007418	RIBC1
ENSG00000167646.13	3.05E-05	0.007902	DNAAF3
ENSG00000189350.12	3.09E-05	0.007947	TOGARAM2
ENSG00000163736.3	3.26E-05	0.008242	PPBP
ENSG00000168589.14	3.24E-05	0.008242	DYNLRB2
ENSG00000111834.12	3.28E-05	0.008246	RSPH4A
ENSG00000188596.10	3.33E-05	0.008257	CFAP54
ENSG00000212766.9	3.33E-05	0.008257	EWSAT1
ENSG00000129991.12	3.44E-05	0.00841	TNNI3
ENSG00000197748.12	3.45E-05	0.00841	CFAP43
ENSG00000276578.1	3.58E-05	0.008688	LLNLR-285B5.1
ENSG00000157856.10	3.68E-05	0.008857	DRC1
ENSG00000266718.1	3.76E-05	0.008975	RP11-466A19.1
ENSG00000152763.16	4.05E-05	0.009617	WDR78
ENSG00000164946.19	4.14E-05	0.009734	FREM1
ENSG00000220267.1	4.16E-05	0.009734	ACTBP8
ENSG00000260198.1	4.26E-05	0.009909	RP11-441F2.2
ENSG00000115339.13	4.35E-05	0.010025	GALNT3
ENSG00000215187.10	4.37E-05	0.010025	FAM166B
ENSG00000187726.8	4.57E-05	0.010365	DNAJB13
ENSG00000267193.5	4.58E-05	0.010365	RP11-116O18.3
ENSG00000168970.22	4.64E-05	0.010372	JMJD7-PLA2G4B
ENSG00000203666.12	4.63E-05	0.010372	EFCAB2
ENSG00000161249.20	4.99E-05	0.011002	DMKN
ENSG00000173557.14	4.97E-05	0.011002	C2orf70
ENSG00000115423.18	5.19E-05	0.011292	DNAH6
ENSG00000182791.4	5.19E-05	0.011292	CCDC87
ENSG00000230565.1	5.33E-05	0.011523	ZNF32-AS2
ENSG00000273449.1	5.43E-05	0.011664	RP11-218F10.3
ENSG00000173947.13	5.50E-05	0.011741	PIFO
ENSG00000272514.5	5.65E-05	0.011998	CFAP206
ENSG00000197153.4	5.95E-05	0.012558	HIST1H3J
ENSG00000215912.12	6.04E-05	0.012663	TTC34
ENSG00000231453.1	6.20E-05	0.012921	LINC01305
ENSG00000133640.18	6.32E-05	0.013038	LRRIQ1
ENSG00000180626.9	6.33E-05	0.013038	ZNF594
ENSG00000135951.14	6.38E-05	0.013058	TSGA10
ENSG00000283538.1	6.60E-05	0.013421	RP11-180P8.1
ENSG00000183914.14	6.80E-05	0.013669	DNAH2

ENSG00000205835.8	6.78E-05	0.013669	GMNC
ENSG00000206172.8	6.95E-05	0.013883	HBA1
ENSG00000134533.6	7.05E-05	0.014002	RERG
ENSG00000283383.1	7.10E-05	0.014026	RP11-499F19.3
ENSG00000272442.2	7.23E-05	0.014206	RP11-444E17.6
ENSG00000141519.14	7.52E-05	0.014693	CCDC40
ENSG00000248464.1	7.71E-05	0.014963	FGF10-AS1
ENSG00000113924.11	8.04E-05	0.015518	HGD
ENSG00000175920.16	8.37E-05	0.016074	DOK7
ENSG00000166473.17	8.59E-05	0.016394	PKD1L2
ENSG00000165309.13	8.73E-05	0.01658	ARMC3
ENSG00000172771.11	8.93E-05	0.016857	EFCAB12
ENSG00000279400.1	9.06E-05	0.017006	CTD-2353F22.2
ENSG00000158578.18	9.13E-05	0.017058	ALAS2
ENSG00000118307.18	9.21E-05	0.017114	CASC1
ENSG00000171811.13	9.37E-05	0.017307	CFAP46
ENSG00000183690.12	9.51E-05	0.017387	EFHC2
ENSG00000232233.1	9.48E-05	0.017387	LINC02043
ENSG00000070731.10	9.86E-05	0.017919	ST6GALNAC2
ENSG00000156042.17	0.0001	0.017938	CFAP70
ENSG00000196565.13	9.99E-05	0.017938	HBG2
ENSG00000229657.2	9.96E-05	0.017938	RP11-494K3.2
ENSG00000115850.9	0.000105	0.018343	LCT
ENSG00000121101.15	0.000104	0.018343	TEX14
ENSG00000141469.16	0.000105	0.018343	SLC14A1
ENSG00000162456.9	0.000103	0.018343	KNCN
ENSG00000178125.14	0.000106	0.018343	PPP1R42
ENSG00000241458.1	0.000105	0.018343	RPL7P19
ENSG00000174529.7	0.000109	0.018843	TMEM81
ENSG00000104450.12	0.000113	0.019334	SPAG1
ENSG00000204666.3	0.000113	0.019334	CTD-2126E3.1
ENSG00000228038.1	0.000115	0.019536	VN1R51P
ENSG00000250990.1	0.000116	0.019659	AC073635.5
ENSG00000117245.12	0.000119	0.019962	KIF17
ENSG00000138615.5	0.000119	0.019962	CILP
ENSG00000139304.12	0.000121	0.019994	PTPRQ
ENSG00000204323.5	0.00012	0.019994	SMIM5
ENSG00000114204.14	0.000124	0.020472	SERPINI2
ENSG00000256552.2	0.000127	0.020809	RP11-113C12.4
ENSG00000133115.11	0.000129	0.020927	STOML3
ENSG00000165383.11	0.000128	0.020927	LRRC18
ENSG00000183562.3	0.000132	0.02138	CTC-343N3.1
ENSG00000258699.1	0.000133	0.02146	RP11-356K23.2

ENSG00000138435.15	0.000134	0.021524	CHRNA1
ENSG00000140527.14	0.000136	0.021669	WDR93
ENSG00000142621.19	0.000136	0.021669	FHAD1
ENSG00000147724.11	0.000137	0.021724	FAM135B
ENSG00000199565.1	0.000139	0.021844	Y_RNA
ENSG00000188659.9	0.000147	0.023056	SAXO2
ENSG00000243802.2	0.00015	0.02342	RP11-390K5.1
ENSG00000254211.5	0.000153	0.023726	LINC01485
ENSG00000214688.5	0.000155	0.023996	C10orf105
ENSG00000171533.11	0.000158	0.024353	MAP6
ENSG00000186354.10	0.000161	0.024638	C9orf47
ENSG00000173838.11	0.000166	0.025288	Mar-10
ENSG00000280927.1	0.000167	0.025432	CTBP1-AS
ENSG00000128408.8	0.000171	0.02589	RIBC2
ENSG00000169126.15	0.000173	0.026	ARMC4
ENSG00000129990.14	0.000175	0.026195	SYT5
ENSG00000007174.17	0.000176	0.026299	DNAH9
ENSG00000234841.4	0.00018	0.026704	RP11-119H12.4
ENSG00000256618.2	0.00018	0.026704	MTRNR2L1
ENSG00000091181.19	0.000183	0.02675	IL5RA
ENSG00000149927.17	0.000183	0.02675	DOC2A
ENSG00000261488.1	0.000183	0.02675	RP11-757F18.5
ENSG00000197921.5	0.000185	0.026844	HES5
ENSG00000215217.6	0.000185	0.026855	C5orf49
ENSG00000089101.17	0.000188	0.026979	CFAP61
ENSG00000205129.8	0.000187	0.026979	C4orf47
ENSG00000139714.12	0.000193	0.027483	MORN3
ENSG00000211663.2	0.000193	0.027483	IGLV3-19
ENSG00000269054.1	0.000197	0.027907	CTD-2619J13.3
ENSG00000133101.9	0.000199	0.028001	CCNA1
ENSG00000155966.13	0.000199	0.028001	AFF2
ENSG00000188729.6	0.000203	0.028468	OSTN
ENSG00000198648.10	0.000204	0.028509	STK39
ENSG00000171962.17	0.000207	0.028725	DRC3
ENSG00000267493.3	0.000211	0.029182	CIRBP-AS1
ENSG00000167434.9	0.000212	0.029218	CA4
ENSG00000109846.7	0.000213	0.029244	CRYAB
ENSG00000175455.14	0.00022	0.030096	CCDC14
ENSG00000010626.14	0.000227	0.030948	LRRC23
ENSG00000218793.1	0.000229	0.030948	RP3-382I10.3
ENSG00000243730.2	0.000228	0.030948	RPL29P3
ENSG00000244300.2	0.000231	0.031107	GATA2-AS1
ENSG00000114670.13	0.000235	0.031493	NEK11

ENSG00000226690.7	0.000241	0.03221	AC005281.1
ENSG00000096093.15	0.000243	0.032334	EFHC1
ENSG00000188536.12	0.000244	0.032341	HBA2
ENSG00000197557.6	0.000246	0.032464	TTC30A
ENSG00000146776.14	0.000253	0.03321	ATXN7L1
ENSG00000173467.8	0.000256	0.033585	AGR3
ENSG00000277103.1	0.000262	0.034239	RP11-520B13.8
ENSG00000203499.11	0.000264	0.034273	FAM83H-AS1
ENSG00000162643.12	0.000279	0.035958	WDR63
ENSG00000227579.5	0.000278	0.035958	RP1-35C21.2
ENSG00000185055.10	0.000283	0.036381	EFCAB10
ENSG00000100583.4	0.000286	0.0366	SAMD15
ENSG00000184471.7	0.00029	0.036927	C1QTNF8
ENSG00000165084.15	0.000296	0.037205	C8orf34
ENSG00000184845.3	0.000294	0.037205	DRD1
ENSG00000187695.8	0.000295	0.037205	RP11-723O4.6
ENSG00000252185.1	0.000296	0.037205	RNU6-752P
ENSG00000261787.1	0.0003	0.037508	TCF24
ENSG00000125845.6	0.000306	0.038148	BMP2
ENSG00000244398.1	0.000321	0.03981	RP11-466H18.1
ENSG00000284526.1	0.000322	0.03981	RP11-666A8.13
ENSG00000199497.1	0.00033	0.040691	RNU1-94P
ENSG00000228848.3	0.000343	0.042202	AC105402.2
ENSG00000103599.19	0.000353	0.043224	IQCH
ENSG00000125122.15	0.00036	0.043853	LRRC29
ENSG00000163737.3	0.000362	0.043853	PF4
ENSG00000163879.10	0.000362	0.043853	DNALI1
ENSG00000102904.14	0.000371	0.044396	TSNAXIP1
ENSG00000165923.15	0.000369	0.044396	AGBL2
ENSG00000224543.4	0.000368	0.044396	SNRPGP15
ENSG00000256973.1	0.000373	0.044396	RP11-359J14.2
ENSG00000265554.1	0.000373	0.044396	RP11-419J16.1
ENSG00000279467.1	0.000377	0.044682	KB-1125A3.12
ENSG00000164530.13	0.000379	0.044808	PI16
ENSG00000103145.10	0.000381	0.044904	HCFC1R1
ENSG00000130957.4	0.000383	0.044954	FBP2
ENSG00000279419.1	0.000401	0.046931	RP5-907C10.3
ENSG00000231933.7	0.000407	0.047242	CTA-125H2.2
ENSG00000235207.1	0.000405	0.047242	TUBBP6
ENSG00000274105.1	0.000408	0.047242	RP11-278C7.3
ENSG00000181619.11	0.00041	0.047348	GPR135
ENSG00000144134.18	0.000418	0.048072	RABL2A
ENSG00000111837.11	0.000428	0.048948	MAK

ENSG00000150873.11	0.00043	0.048948	C2orf50
ENSG00000260776.5	0.000428	0.048948	RP11-114H24.2
ENSG00000160345.12	0.000435	0.049292	C9orf116
ENSG00000270011.6	0.000436	0.049292	ZNF559-ZNF177
ENSG00000102575.10	0.00044	0.049675	ACP5
ENSG00000133454.15	0.000447	0.049921	MYO18B
ENSG00000142530.10	0.000446	0.049921	FAM71E1
ENSG00000150773.10	0.000449	0.049921	PIH1D2
ENSG00000165606.8	0.00045	0.049921	DRGX
ENSG00000265148.5	0.000449	0.049921	TSPOAP1-AS1
ENSG00000107014.8	0.000453	0.049942	RLN2
ENSG00000204389.9	0.000452	0.049942	HSPA1A

Table 24: Significant genes identified in the DEA of the GLIOTRAIN RNA-Seq data.