

INTEGRATIVE NETWORK MODELLING OF THE DASATINIB TREATMENT IN
GLIOBLASTOMA STEM CELLS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÖKÇE SENER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
BIOINFORMATICS

MARCH 2019

Approval of the thesis:

**INTEGRATIVE NETWORK MODELLING OF THE DASATINIB TREATMENT IN
GLIOBLASTOMA STEM CELLS**

Submitted by GÖKÇE SENGER in partial fulfillment of the requirements for the degree of **Master of Science in the Department of Bioinformatics, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics, METU**

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics, METU**

Assoc. Prof. Dr. Nurcan Tunçbağ
Supervisor, **Health Informatics, METU**

Examining Committee Members:

Prof. Dr. Rengül Çetin Atalay
Health Informatics Dept., METU

Assoc. Prof. Dr. Nurcan Tunçbağ
Health Informatics Dept., METU

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics Dept., METU

Assoc. Prof. Dr. Özlen Konu
Molecular Biology and Genetics Dept., Bilkent
University

Asst. Prof. Dr. Nihal Terzi Çizmecioglu
Biological Sciences Dept., METU

Date:

26.03.2019



I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : GÖKÇE SENGER

Signature : _____

ABSTRACT

INTEGRATIVE NETWORK MODELLING OF THE DASATINIB TREATMENT IN GLIOBLASTOMA STEM CELLS

Senger, Gökçe

MSc., Department of Bioinformatics

Supervisor: Assoc. Prof. Dr. Nurcan Tunçbağ

March 2019, 60 pages

Glioblastoma (GBM), the most aggressive type of the glial tumours, is thought to be widely promoted by stem-like cells. Although certain cancer types have been radically treated with Receptor Tyrosine Kinases (RTKs) inhibitors, prior studies demonstrate that treatment Glioblastoma Stem Cells (GSCs) with RTK inhibitors led to dynamic interconversion from proliferative to slow-cycling, persistent state. In this work, we use the publicly available RNA-Seq and ChIP-Seq data in naive patient-derived GBM cell line (GSC8), 12-day and chronic dasatinib treated GSC8 published by Liao et al (Liao et al., 2017) and apply an integrative approach to develop a further explanation for reversible transition of GSCs and to model the effect of the dasatinib treatment in a network context. We first used the Garnet module in Omics Integrator software which identifies transcription factor binding sites from epigenomic data, relates known and predicted transcription factor binding sites to gene expression and finds the significantly active transcription factors. Then, we used the Forest module of the Omics Integrator software to reconstruct an optimal network for each condition by integrating significantly active transcription factors and a confidence weighted protein interactome. As a result, we obtained three condition specific networks and clustered them based on the topology of these networks. Each module was analysed in terms of pathway enrichments. Then, we compared these networks based on the node, edge and pathway similarities. We reveal that GSCs tend to activate RTK-targeted genes and upregulate neurodevelopmental programs by reorganizing chromatin modifications.

Keywords: Integrative Network Modelling, Multi-omics data, Histone modifications, Receptor Tyrosine Kinases, Glioblastoma Stem Cells

ÖZ

GLİOBLASTOMA KÖK HÜCRELERİNDE DASATİNİB TEDAVİSİNİN BÜTÜNLEYİCİ AĞ MODELLEMESİ

Senger, Gökçe

Yüksek Lisans, Biyoenformatik Bölümü

Tez Yöneticisi: Doç. Dr. Nurcan Tunçbağ

Mart 2019, 60 sayfa

Glioblastoma, tüm beyin tümörleri içinde en sık rastlanan kötü huylu bir beyin tümörüdür. Kanser genom çalışmaları reseptör tirozin kinazların (RTK) glioblastoma vakalarında en çok değişen genler olduğunu göstermiştir. Fakat bilinen ve bu genleri hedef alan tedavi yöntemleri glioblastoma için yeterli olmamıştır. Son çalışmalar gösteriyor ki, glioblastomanın çok kökenli hücre fenotipi ve kalıtsal yapısı tedaviler karşısında direnç göstermesine neden oluyor. Liao ve çalışma arkadaşlarının yaptığı bir çalışma, glioblastoma kök hücrelerinin, reseptör tirozin kinazları hedef alan bir ilaç, dasatinib, etkisiyle epigenetik mekanizmalar yardımıyla kendilerini uyku durumuna soktuklarını ve ilaca dayanaklı hale geldiklerini göstermektedir. Bu tez çalışmasında, glioblastoma kök hücrelerinin tersinir epigenetik mekanizmalarının açıklanması ve dasatinib etkisiyle meydana gelen hücresel iletim trafiğinin modellenmesi hedeflenmiştir. İlk olarak epigenomik veriden transkripsiyon faktörlerinin bağlanma bölgelerini tanımlamak için Omics Integrator programının Garnet modülünü kullandık. Buradan gelen sonuçları gen ifadesi verisi ile birleştirerek önemli derecede aktif olan transkripsiyon faktörlerini belirledik. Daha sonra, belirlenen faktörlerin protein iletişim ağlarındaki yerlerini belirlemek için Omics Integrator programının Forest modülünü kullandık. Sonuç olarak üç farklı durum için ağ modelleri elde edip bunları topolojik özelliklerine göre karşılaştırılması yapıldı. Ayrıca her model için sinyal yolak analizi yapıldı. Daha sonra modeller, topolojik ve sinyal yolak benzerliklerine göre karşılaştırıldı. Bu çalışma ile glioblastoma kök hücrelerinin RTK ile aktivasyonu sağlanan genleri ve sinirsel gelişim programlarını, kromatin modifikasyonlarını düzenleyerek aktiveleştirme eğilimlerini göstermiş olduk.

Anahtar Sözcükler: Ağ Modelleme, Omik Veri, Histone modifikasyonu, Tirozin Kinaz Reseptörleri, Glioblastoma kök hücreleri



To My Family

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my adorable advisor Assoc. Prof. Dr. Nurcan Tunçbağ for her continuous support and endless patience. As an academic perspective, the capability of dealing with obstacles throughout the academic path and the corner stones of the professionalism in the field have been conveyed from her to me with an enlightened way. This work would not have been possible without her unique and unequalled guidance.

Besides my supervisor, I would like to thank the rest of my thesis committee: Prof. Dr. Rengül Çetin Atalay, Assoc. Prof. Dr. Yeşim Aydın Son, Assoc. Prof. Dr. Özlen Konu and Asst. Prof. Dr. Nihal Terzi Çizmecioğlu for accepting to be on my thesis committee members in despite of their strict schedule; also, it should be noted that their encouragement and challenging questions were most appreciated.

I would also like to thank all my colleagues from Network Modelling Lab with whom I have had the pleasure to work together. Particularly, I would like to give my gratitude to my beloved friends Cansu Demirel, Cansu Dinçer and Meriç Kınalı who made our journey enjoyable and remarkably intuitive in terms of academic development and collaboration in many fields of science. I also would like to thank Alperen Taciroğlu, Muazzez Çelebi Çınar, Elif Bozlak and Evrim Fer for giving me the pleasure overcoming by stressful times. Most importantly, I am grateful to my long-term friend Ezgi Gül Keskin who is like family to me for a long time for being empathetic and supportive.

I owe a special thanks to my family: to my father, Zafer Senger, for showing that everything would be possible even if you are in the darkest night, to my beautiful mother, Nurten Kurt, for teaching me to be strong against every impediment and to my beloved sisters, Sevgi Nur Senger and Ceyda Senger, for believing in me and supporting me in the finding of my own way. Exceptionally, I would like to thank my loving and supportive partner in crime, Uğur Can Erdem, for encouraging me and being by my side even though we are living thousand kilometers apart.

TABLE OF CONTENTS

ABSTRACT	iv
ÖZ.....	v
DEDICATION	vi
ACKNOWLEDGMENTS	vii
TABLE OF CONTENTS.....	viii
LIST OF TABLES	x
LIST OF FIGURES	xi
LIST OF ABBREVIATIONS	xii
CHAPTERS	
1.INTRODUCTION.....	1
2.LITERATURE REVIEW	5
2.1. Understanding of The Underlying Biology of GBM	5
2.1.1. Identification of GBM subtypes based on transcriptomic profiling.....	5
2.1.2. Glioblastoma Stem Cells are defined as new targets for therapy	6
2.2. Current Studies on Transcriptomic and Epigenomic Profiling of GBM.....	7
2.3. ARACNe: An Algorithm for the Reconstruction of Accurate Cellular Networks	9
2.4. Integration of Multi-Omics Data.....	10
3.MATERIALS AND METHODS	13
3.1. Overview of the Pipeline	13
3.2. Datasets	14
3.2.1. Data from ENA	14
3.2.2. Data from UCSC	15
3.3. RNA-Seq Analysis.....	16
3.3.1. Alignment RNA-Seq reads to reference transcriptome.....	16
3.3.2. Different Gene Expression Analysis	17

3.3.3. Gene Set Enrichment Analysis.....	18
3.4. ChIP-Seq Analysis.....	18
3.4.1. Alignment ChIP-Seq reads to reference genome	18
3.4.2. Differential Peak Calling	19
3.5. Network Modelling with Omics Integrator.....	20
3.5.1. Identification of transcription factors using Garnet.....	21
3.5.2. Network integration with Forest.....	22
3.5.3. Visualization of networks with Cytoscape.....	23
3.5.4. Overrepresentation Enrichment Analysis	24
4.RESULTS.....	25
4.1. Transcriptomic Profiling Reveals That Notch Pathway and Histone Modification Related Genes are Highly Enriched in Persister Cells.....	25
4.2. Gene Set Enrichment Analysis Shows That Cell-Cycle Related Biological Functions are Negatively Regulated in Drug-Treated Cells.....	29
4.3. Comparative Comparison of Significantly Active Transcription Factors Across Pairwise Comparison Conditions.....	30
4.4. Network Modelling Indicates That GSCs Prefer an Alternative Cell Surface Receptors to Activate RTK-dependent Pathways.....	34
4.5. Comparison of The Results of Overrepresentation Enrichment Analysis of Clusters in Each Condition Specific Network.....	37
5. DISCUSSION AND CONCLUSION.....	45
REFERENCES	47
APPENDICES	55
APPENDIX A	55

LIST OF TABLES

Table 2.1: Changes in genes and pathways associated with GBM subtypes..... 6



LIST OF FIGURES

Figure 3.1: Overall representation of the methodology	14
Figure 3.2: RNA-Seq and ChIP-Seq data for cell line GSC8	15
Figure 3.3: The overall representation of differential gene expression analysis	17
Figure 3.4: The overall representation of differential peak calling analysis	20
Figure 4.1: Gene expression profiles of naïve, 12d and persister cells.....	27
Figure 4.2: The expression profiles of some selected genes in naïve, 12d and persister cells	28
Figure 4.3: Differentially expressed genes across pairwise comparison conditions	29
Figure 4.4: KEGG pathways enriched in each pairwise comparison condition.....	30
Figure 4.5: Biological functions enriched in each pairwise comparison condition.....	31
Figure 4.6: Significantly active TFs in each pairwise comparison condition	32
Figure 4.7: Significantly active TFs found in at least two pairwise comparison conditions	33
Figure 4.8: Condition specific network for the pairwise comparison condition naïve vs 12d	36
Figure 4.9: Condition specific network of the pairwise comparison condition 12d vs persister	36
Figure 4.10: Condition specific network of the pairwise comparison condition naïve vs persister	37
Figure 4.11: Pathway analysis across clusters in the naïve vs 12d network.....	38
Figure 4.12: Pathway analysis across clusters in the 12d vs persister network	39
Figure 4.13: Pathway analysis across clusters in the naïve vs persister network.....	40
Figure 4.14: Biological process analysis across pairwise comparison condition networks	41
Figure 4.15: Molecular function analysis across pairwise comparison condition networks	42
Figure 4.16: Pathway analysis across pairwise comparison conditions	43

LIST OF ABBREVIATIONS

ChIP-Seq	Chromatin Immunoprecipitation Sequencing
CSC	Cancer Stem Cell
DPI	Data Processing Inequality
ENA	European Nucleotide Archive
FFL	Feedforward Loop
GBM	Glioblastoma Multiforme
GEP	Gene Expression Profile
GSC	Glioblastoma Stem Cell
GSEA	Gene Set Enrichment Analysis
MES	Mesenchymal
MI	Mutual Information
NSC	Neural Stem Cell
ORA	Overrepresentation Enrichment Analysis
PN	Preneural
PPDE	Posterior Probability of Differentially Expressed
PPEE	Posterior Probability of Equally Expressed
PROLIF	Proliferative
REMBRANDT	Repository for Molecular Brain Neoplasia Data
RNA-Seq	RNA Sequencing
RTKs	Receptor Tyrosine Kinases
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
TFA	Transcription Factor Affinity

CHAPTER 1

1.INTRODUCTION

Glioblastoma Multiforme (GBM), the most frequent type of human brain cancers, is thought to be the most aggressive and malignant primary brain tumour propagated by stem-like cancer cells. Although new therapeutic strategies have been tried to either slow down or stop the progression of cancer and they have worked against certain cancer types, these treatments have failed to lower mortality rate of GBM. The intra- and inter-tumoral genetic and phenotypic heterogeneity of GBM is the main reason why these tumours resistance to current therapies which are mostly designed to target bulk tumours (Ikushima et al., 2009).

Transcriptomic profiling is a widely used technique to deeply understand the molecular background under the extensively diverse GBM phenotypes and to biologically characterize of Glioblastoma stem cells (GSCs). Transcriptomic profiling is basically defined as annotation and quantification of all RNA molecules in a single cell or population of cells which is so called transcriptome. Recent advances in high-throughput RNA sequencing and the decrease in sequencing cost make large scale data collection possible, opening the door to categorize anomalous events in the progression and maintenance of GBM and to develop a consciousness about vulnerability and aggressiveness of GBM (Zhao et al., 2017). These genome wide studies have led to detailed characterization of the molecular pathogenesis of GBM, continuity of clinical treatment plans and identification of new drug targets. While receptor tyrosine kinase (RTK)/Ras/PI3K, p53 and Rb signalling pathways are the most commonly affected pathways in GBM according to The Cancer Genome Atlas (TCGA) research, EGFR, HER2, PDGFRA and MET are the most altered genes among receptor tyrosine kinases (RTKs) (Tanaka, Louis, Curry, Batchelor, & Dietric, 2013). Although RTK family inhibitors has been successful in the treatment of certain cancer types (Tanaka et al., 2013), they have failed to develop a recovery in GBM because of the genetic and phenotypic heterogeneity in GBM and distinctive characteristics of GSCs (Eder & Kalman, 2014; justin d. lathia, 2015).

Previous studies on GBM showed that GSCs have a critical role in tumour progression, maintenance of the malignant behaviour of GBM and resistance to current therapies (Bao et al., 2006; Singh et al., 2004). Properties of GSCs have been well studied and it was indicated that they are capable of forming renewable neurospheres *in vitro* (Kalkan, 2015) and express neurodevelopmental transcription factors (TFs) (Alberta et al., 2011; Rheinbay et al., 2013; Wortman et al., 2014) which is thought to be significantly associated with GSC maintenance and tumorigenic capacity of GBM (Laks et al., 2011). Furthermore, GSCs can express multipotent neural stem cells (NSCs)-like cells which can further differentiate into neurons, astrocytes, and oligodendrocytes within the tumour mass (Huang, Cheng, Guryanova, Wu, & Bao, 2010). According to single-cell RNA sequencing studies, cell-cycle related genes indicate relatively low expression in GSCs (Patel et al., 2014) suggesting these cells may adopt slow-cycling states (Sosa, Bragado, & Aguirre-Ghiso, 2014).

The transcriptome is regulated by multiple mechanisms including the dynamic chromatin reorganization such as DNA methylation and histone modifications, defined as epigenome (O'Geen et al., 2011). Changes in the nucleosomes alter the DNA packaging and affect the gene expression by turning on/off a certain fraction of upstream regions of genes, referred as transcriptional start site. For example, histone methylation is typically associated with open or compacted chromatin regions while histone acetylation is associated with open and accessible chromatin regions. The chromatin immunoprecipitation (ChIP-Seq) method is used to determine direct physical interactions between DNA and protein across different modification markers and to improve a perspective on how proteins bind to DNA to regulate gene expression. The interaction between TFs and DNA can also be studied using ChIP-Seq. Therefore, epigenomic profiling is also crucial to reveal tumour maintenance mechanisms of GBM (Dirks, Stunnenberg, & Marks, 2016).

It is still not completely known how GSCs' stemness is maintained despite the well-studied transcriptomic and epigenomic profiling of GBM. However, we can gain new insights from the perspective of system biology and we can extensively analyse how the transcriptomic entities and regulators communicate each other in GBM.

In this thesis we used advanced system biology approaches, namely multi-omic data integration followed by network modelling to elucidate maintenance of stemness and drug resistance of GSCs. We used the transcriptomic and epigenomic data of GBM cells in multiple conditions. In this work, the publicly available RNA-Seq and ChIP-Seq data from patient-derived GBM cell line treated with dasatinib has been used to model the effect of the dasatinib treatment in a network context. The important point in the method is to reconstruct transcription factor network for different time points of the treatment. For this aim, we first

used the Garnet module in Omics Integrator software which identifies transcription factor binding sites from epigenomic data, relates known and predicted transcription factor binding sites to gene expression and finds the significantly active transcription factors. Then, we used the Forest module of the Omics Integrator software to reconstruct an optimal network for each condition by integrating significantly active transcription factors and a confidence weighted protein interactome.

In Chapter 2, we first explained the current studies in transcriptomic profiling in detail. Then we reviewed recent approaches in the analysis of epigenomic data. We also recovered recent tools for identification of transcription factor binding and reconstructing networks through the integration of multi-omics data.

In Chapter 3, we described the methodology starting from analysing gene expression profiles of GSCs with different drug exposure time, comparing these expression profiles, identification of condition specific transcription factors using histone modification data to reconstructing final optimal networks.

In Chapter 4, we emphasized that how our results would contribute to understanding dynamic chromatin reorganization mechanisms of GSCs which control rapid conversion from proliferative to slow-cycling, drug-persistent state. we also indicated that the network-based integrative approach that we proposed in the study highlighted new signalling pathways which could be clinically important and potential targets in new treatments.

CHAPTER 2

2.LITERATURE REVIEW

2.1. Understanding of The Underlying Biology of GBM

2.1.1. *Identification of GBM subtypes based on transcriptomic profiling*

Prior studies have demonstrated that defining subtypes for GBM using genomic and transcriptomic analyses is hard because of the phenotypic and genotypic heterogeneity of GBM. In 2006, Phillips and his colleagues identified three subtypes which are mesenchymal (MES), proliferative (PROLIF) and preneural (PN) by using DNA microarray data (Phillips et al., 2006). Then Verhaak and his co-workers added a critical fourth class and named it as Classical (Verhaak et al., 2010). The fourth class share similar characteristics with both PN and MESS groups, so they defined the fourth class as an intermediate group. Further studies have shown that GBM cells classified into the classical group have increased level of cell cycle and proliferation genes (Huse, Holland, & DeAngelis, 2013).

Although transcriptomic and genomic approaches have been the most commonly used approaches, various methods have also been developed to identify different subtypes of GBM. Toedt et al. used an integrative approach by using array-based comparative genomic hybridization and expression profiling analyses. They found three GBM subtypes similar to those of Phillips et al. (Toedt et al., 2011). The extensive studies in the identification of GBM subtypes has introduced new requirements to deeply characterize and understand each subtype. Morokoff et al. conducted a detailed study to define signalling pathways and gene expression patterns associated with each subtype. **Table 2.1** indicates which pathways are

significantly associated with which GBM subtype and also highlights differentially expressed genes across different subtypes (Morokoff, Ng, Gogos, & Kaye, 2015). It appears that Notch signalling pathway is critical for the proneural subtype (Saito et al., 2014). However, the MES phenotype is more dependent on TGF β signalling and tend to be more invasive (Mahabir et al., 2014).

Table 2.1: Changes in genes and pathways associated with GBM subtypes.

Glioblastoma subtypes by Phillips et al.	Glioblastoma subtypes by Verhaak et al.	Genes or chromosomes amplified, overexpressed	Genes or chromosomes deleted, mutated, downregulated
Proneural	Proneural	PDGFRA amplification MYC amplification OLIG2 PI3K/Mtor Hedgehog, Wnt, Notch CDK4 amplification SOX2 amplification, DCX, DLL3, ASCL1, TCF4 CXCR4 ALT-positive	IDH1 mutations TP53 mutations ATRX 1p/19q loss, CIC, TERT, FUBP1 (oligo) COX2 IGFBP2 Annexin1 TAZ PIK3CA/PIK3R1 4EBP1
Proliferative	Classical	Chr7, 19, 20 amp EGFR amp EGFRvIII NES Sonic HH	EGFRvIII (activating) Chr10 loss CDKN2A-deletion FAT1 inactivation MGMT methylation
	Neural	NEFL, GABRA1, SYT1, SLC12A5	ND
Mesenchymal	Mesenchymal	EGFR, PI3K/Akt CHI3L1, YKL40, vimentin MET, CD44, MERTK TGF β /BMP, TNF family NF-kappaB CXCR4 CD31, VEGFR-2 Snail	NF1 (17q11.2 deletion) TSC2 tuberin

2.1.2. Glioblastoma Stem Cells are defined as new targets for therapy

GBM is the most prevalent and aggressive type of primary intrinsic human brain tumor and contains tumorigenic cancer stem cells (CSCs). Studies on genetic profiling of these cells have revealed that they can mimic neural stem cells (Galli et al., 2004) and thus CSCs in GBM are named as stem-like cells or glioblastoma stem cells (GSCs). GSCs contribute tissue development processes and help cells to regenerate and grow (Lathia, Mack, Mulkearns-Hubert, Valentim, & Rich, 2015). Furthermore, GSCs show self-renewal and they are ability to differentiate into various cell types like neural stem cells. In addition, tumor propagation in GBM is widely promoted by GSCs. Because of the properties of these stem cells, recent therapies have been developed to target GSCs rather than bulk tumor tissues.

In recent decades, cancer therapies have initially particularly focused on the characterization of signaling pathways underlying GBM biology to develop new treatments which target GSCs. It was found that RTK signaling plays a critical role and EGFR, VEGFR and PDGFR are the most studied RTK receptors which have central role in GBM. Furthermore, it was also found that EGFRvIII-positive GSCs, an active mutation of EGFR, positively regulate cell proliferation related pathways while they negatively regulate apoptotic signaling pathways. Nevertheless, paradoxically it has been revealed that EGFRvIII is associated with better prognosis. In addition, normal GBM cells have lower expression of EGFRvIII or they lost it and EGFRvIII-negative cells are resistance to RTK-inhibitors (Montano et al., 2011; Schulte et al., 2012). It is possible to conclude that while EGFRvIII-positive GSCs are differentiating into proliferating normal tumor cells, they lose EGFRvIII expression and develop a cellular adaptive resistance which may explain why EGFR-inhibitors have failed to improve overall survival in GBM (Inda et al., 2010). MET and PI3K/Akt signaling pathways are also thought to have important role in maintenance of stem-like phenotypes in GSCs (Jun et al., 2012; Molina, Hayashi, Stephens, & Georgescu, 2010). In addition, there are other studies which emphasized the role of the Hedgehog, Notch and the canonical Wnt signaling pathways in GBM development and progression (Fan et al., 2009; Sandberg et al., 2013).

2.2. Current Studies on Transcriptomic and Epigenomic Profiling of GBM

The process of gene expression within a cell is regulated by transcription regulation mechanisms in which transcription factors (TFs) bind specific DNA regions, which are called motifs, encourage other proteins assemble and help start transcription of specific gene. Thus, TFs perform key functions in the regulation of gene expression. Furthermore, TFs can also affect regulation of

another TF by making direct or indirect interactions (Neph et al., 2012). These cross-regulations are involved in the various regulatory subnetworks and dysregulation of these interactions are strongly related with various types of cancer (Stergachis et al., 2014). Much progress has been done in the understanding of the role of TF-TF interaction in cellular identity and function. However, more efforts should be made for the comprehensive understanding of the topology of human TF-TF networks.

Li and colleagues (Y. Li et al., 2015) analysed genome-wide expression profiles of TFs from two different gene expression datasets, TCGA and REMBRANDT and constructed grade-specific TF regulatory networks for glioma grade I, III and IV. Then they compared the resulting networks based on their topology and dynamics to improve an understanding for how GBM transcription regulatory networks change during different progression states of GBM.

They demonstrated that although key regulatory interactions are shared by all types of grades, human transcription regulatory interactions of glioma are generally specific to glioma grade types, with ratios between 45% and 60%. In addition, they compared the sub interactions within each network to understand whether some topological structures might be responsible for the conserved architecture across different glioma grades. They identified the feedforward loop (FFL) which are the most common structures in the grade-specific networks. They drew attention to RARG-NR1|2-CDX2 FFL which is observed in each grade-specific network and associated with prognosis.

Integration of multi-omics data may provide new insights to understand intratumoral heterogeneity in GBM. Lemée et al put forward a new approach in which they used transcriptome and proteome profiling together to identify pathogenic mechanisms underlying the biology of GBM (Lemée et al., 2018). They used both RNA microarray chips and proteome data from five GBM biopsies vs their related peritumoral brain zone. Then they compared the transcriptome data with their corresponding proteome data in terms of shared characteristics, altered biological processes, functional pathways and network topology (Haider & Pal, n.d.). They found that there is a poor relation between the transcriptome and its corresponding proteome data in GBM. However, they revealed that neurofilament light polypeptide and synapsin 1 protein abundances are strongly correlated with the mRNA abundances of the related genes. Furthermore, both transcriptomic and proteomic data support that biological processes related to cell-cell communication, synaptic transmission and nervous systems are the most commonly altered ones across the five GBM samples.

2.3. ARACNe: An Algorithm for the Reconstruction of Accurate Cellular Networks

Cellular phenotypes and cell physiology are largely dependent on the activity of cellular functions which are controlled by dynamic activity of complex networks of coregulated genes. Thus, clustering cells based on their phenotypic characteristics requires elaborative work in which genes are classified in the context of the networks in which they do functions. Proteins are synthesized from the gene products, mRNA, and regulate expression of genes by directly or indirectly binding to regulatory regions of DNA. However, there are post transcriptional mechanisms that control conversion of the gene products to proteins, functional units of cells. Because of the post regulation mechanisms, the abundance level of proteins cannot be directly proportional to level of mRNA within the cell. Consequently, the indirect relation has spawn requirements for additional algorithms for the reconstruction of gene co-expression networks by using the data coming from high throughput analysis, microarray and RNA-Seq studies.

Genome-wide clustering approach (Eisen, Spellman, Brown, & Botstein, 1998) groups together genes which are responsible for similar transcriptional responses to different cellular conditions, and provides a crucial first step to reconstruct interaction networks. Nevertheless, it cannot eliminate interactions arising from indirect relations of cellular cascades and provides biased networks that include many non-interacting genes (A.A. et al., 2006). Thus, gene expression profiles and correlations between genes cannot be used to reconstruct interaction networks without additional statistical assumptions.

New graph-based approaches have been developed to model cellular networks from large-scale gene expression profiles (Friedman, 2004). The aim of these approaches is to represent gene regulatory circuits by using topology of graphs in which genes are represented as vertices and the direct interactions between them are represented as edges.

Differently from genome-wide clustering methods, the new strategy uses statistical inferences methods to control whether a physical interaction is strongly related with the data or not and therefore provides more realistic network models (Ideker et al., 2001). Among the computational approaches, ARACNe, Algorithm for the Reconstruction of Accurate Cellular Networks, is the most widely applied algorithm by the scientific community to model accurate and systematic gene regulatory networks (A.A. et al., 2006; Floratos, Smith, Ji, Watkinson, & Califano, 2010). More generally, ARACNe assigns an irreducible statistical dependency to an edge between genes that interact directly, and this

interaction is mediated by a transcription factor binding promoter region of a target gene.

Basically, ARACNe takes Gene Expression Profile (GEP) data and a list of transcription factors and then reconstruct context-specific transcriptional networks (Lefebvre et al., 2010). There are three key steps to run ARACNe which are Mutual Information (MI) threshold estimation, Bootstrapping/MI network reconstruction and Building consensus network (Lachmann, Giorgi, Lopez, & Califano, 2016). First, ARACNe estimates MI threshold by using GEP data. Then ARACNe defines genes included in GEP data within a fixed window which is regulated by a certain TF for each TF in the predefined list. In the network reconstruction step, it computes MI, which is the measurement of statistical dependence between two genes, for every TF/Target gene pair and removes non-statistically significant pairs by using the MI threshold. In addition, ARACNe removes indirect interactions by using Data Processing Inequality tolerance filter (DPI) (A.A. et al., 2006) and reconstructs final networks. At the final step, ARACNe does optimization for the resulting network based on a Poisson distribution in which it calculates the number of times a certain edge is found across all bootstrap runs and keeps the ones with p value lower than 0.05.

2.4. Integration of Multi-Omics Data

Recent developments in the high-throughput methods have resulted in accumulation of large amount of omics data in the biology era such as genomics, transcriptomics, epigenomics, proteomics and metabolomics (Suravajhala, Kogelman, & Kadarmideen, 2016). Previously, these omics measurements have been used in single-level analysis in which each data type is analysed separately. However, while the level of knowledge regarding molecular complexity of biological systems has been increasing, more comprehensive analysis of the omics data has become a necessity. The increase in the volume of omics data and the necessity for multi-layer analysis to enhance comprehension of this complexity have led new approaches where data from different omics studies are combined (Kadarmideen, von Rohr, & Janss, 2006).

As a concept, multi-omics data integration covers the system biology approaches which include obtaining biological data from different layer of living systems, using these data together and applying computational model to reconstruct whole system organization (Cisek, Krochmal, Klein, & Mischak, 2016). Currently, integration of various types of omics data is widely used for different aims such as defining cell-specific phenotypes, characterization of cellular pathways,

developing patient-specific treatments and understanding gene regulatory circuits.

Although multi-omics data integration approaches have more potential to shed light on complex mechanisms underlying living systems' biology than single-layer methods, making meaningful correlations and identifying true interactions among thousands of measurements obtained from various omics methods remain a challenge (Misra, Langefeld, Olivier, & Cox, 2018). Thus, data integration usually needs statistical implementations and machine-learning tools (Min, Lee, & Yoon, 2016). Most of the omics integrative frameworks use multivariate analysis tools to reduce data dimensionality and to implement genomic, proteomic and metabolomic datasets together (de Tayrac, Le, Aubry, Mosser, & Husson, 2009; Parkhomenko, Tritchler, & Beyene, 2009).

There are many integrated omics methods which address different challenges. Pavel and her co-workers used a fuzzy logic modelling presented by Xu et al. (Xu, 2008) to identify patient-specific gene expression and cancer drivers. In this study, they used direct integrative clustering of samples, clustering of pre-formed clusters and as a third category, regulatory integrative clustering (Pavel, Sonkin, & Reddy, 2016). Another study focused on integrating transcriptomic and proteomic data by using proteomics-first approach to identify cancer related sub-networks (Nibbe, Koyutürk, & Chance, 2010). In this work, Nibbe et al. identified significant proteomic targets by analysing fold changes between tumour and control tissues. Then, they used these targets to construct protein-protein interaction sub-networks which is associated with disease phenotypes. Although the network generation method used in this study has revealed interactions among molecules with known functions, the ideal network construction approaches involve investigation of key molecules with novel functions. There are tools which specifically aim to reduce false positives and negatives in the data and identify novel interactions within the final network. For instance, SteinerNet (Tuncbag, McCallum, Huang, & Fraenkel, 2012) which integrates transcriptomic, proteomic and interactome data and Omics Integrator (Tuncbag, Gosline, et al., 2016) which integrates transcriptomic, epigenomic and interactome data. The two tools were generated to search for the solution to the prize collecting Steiner tree problem.



CHAPTER 3

3.MATERIALS AND METHODS

In this chapter, we explain the methodology of this study which covers transcriptomic and epigenomic profiling of RTK-dependent GSC lines and network modelling by integrating multi-omics data

3.1. Overview of the Pipeline

GSCs show rapid dynamic interconversion from their naïve, proliferative states to slow-cycling, persister states by reorganizing chromatin modifications under the treatment of dasatinib. In this study, we developed an integrative understanding for the characteristics of GSCs by using transcriptomic, epigenomic and interactome data, which is summarized in **Figure 3.1**. The transcriptomic and epigenomic data are previously published by Bernstein (Liau et al., 2017) and his co-workers and accessible for academic usage. First, we calculated the gene expression changes between different drug-treated time series profiles: GSC8 naïve, GSC8 12d and GSC8 persister using RSEM-EBSeq pipeline. Then we performed differential peak calling analysis between each pairwise comparison condition from histone modifications ChIP-Seq data for the markers, H3K4me3, H3K27me3 and H3K27ac with MACS2 pipeline. After that, we integrated gene expression changes and differential peak calling results to find significantly active transcription factors by using the Garnet module of Omics Integrator which have potential to explain condition-specific regulatory changes. Finally, we used the Forest module of the Omics Integrator software to reconstruct an optimal network for each pairwise comparison condition by

integrating significantly active transcription factors and a confidence weighted protein interactome.

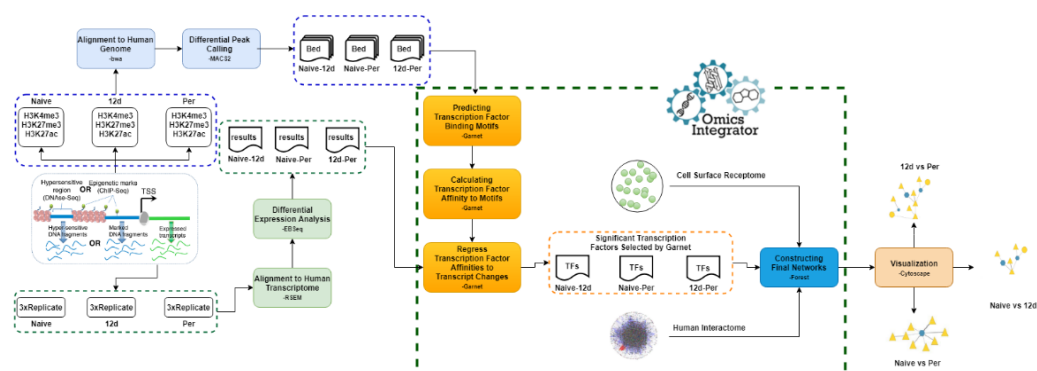


Figure 3.1: Overall representation of the methodology.

3.2. Datasets

3.2.1. Data from ENA

The European Nucleotide Archive is a nucleotide sequencing based repository which covers corresponding information about sequencing experiments; input data, experimental design, raw sequences and quality score and functional annotation. We downloaded raw sequences from ENA for the experiment performed RNA-Seq in different GSC lines with various drug treatments (GEO accession: GSE74557). RNA-Seq data for two GSC lines; GSC4 and GSC8 and four different drug treatments; Dasatinib, PD0325901, GSKJ4 and KDM5-C70 are available for this study. We chose to start downstream analysis with GSC8 as cell line and dasatinib as drug treatment because different time points for this treatment and corresponding replicates are available only for this condition. Paired-end reads for three biological replicates per drug treatment time point were downloaded from ENA (<http://www.ebi.ac.uk/ena/data/view/SRR4417704-SRR4417712>) in the form of FASTQ file, shown in **Figure 3.2**. FASTQ is a text-based format designed for storing biological sequences and their quality information. For a usual FASTQ file, there are four lines for each sequence. The first line begins with “@” character and contains sequence identifier. The second line is for the sequence itself and contains raw sequence letters. The line 3 begins with a “+” character.

For some FASTQ files, the third line contains the sequence identifier same as in the first line. The last line is for the quality scores for raw letters in the line 2 and should be the same length with the raw sequence.

After downloading the data, we had FASTQ files which store reads of RNA-Seq experiment for GSCs in their native status (naïve) and different drug treatment time points which are twelve-day dasatinib treatment (12d) and more than eight weeks dasatinib treatment (persister). Each time point has three biological replicates. To perform epigenetic profiling and reveal transcription factors that are most probably regulator of transcriptomic profiles of corresponding condition, we also downloaded ChIP-Seq sequence single-end read files from ENA in FASTQ format for the following antibodies: H3K4me3, H3K27me3 and H3K27ac per condition. The depth of the ChIP-Seq reads are usually different from each other. To make the samples comparable, input files that contains sequenced fragments do not originate from histone markers of interest are also prepared (Flensburg, Kinkel, Keniry, Blewitt, & Oshlack, 2014) (Liang & Keleş, 2012). Thus, the input read files for each antibody were also downloaded from ENA. The detailed information about the ChIP-Seq read files is presented in **Figure 3.2**.




	 GSC naïve	 GSC ^{12d}	 GSC ^{Per}
RNA-Seq Data	Replicate 1 Replicate 2 Replicate 3	Replicate 1 Replicate 2 Replicate 3	Replicate 1 Replicate 2 Replicate 3
ChIP-Seq Data	H3K4me3 H3K27me3 H3K27ac	H3K4me3 H3K27me3 H3K27ac	H3K4me3 H3K27me3 H3K27ac

Figure 3.2: RNA-Seq and ChIP-Seq data for cell line GSC8. Conditions are represented as GSC naïve for naïve, GSC^{12d} for 12d and GSC^{Per} for persister. There are three replicates per condition for the transcriptomic analysis. ChIP-Seq datasets come from the experiments that were conducted with the following antibodies: *H3K4me3*, *H3K27me3* and *H3K27ac*.

3.2.2. Data from UCSC

The human genome was downloaded from UCSC with the version of GRCh37/hg19 (assembly Feb. 2009). UCSC stores assemblies and their corresponding annotations for a wide range of organism from vertebrate to model organisms. UCSC also provides various tools to view, analyse and

download data. Sequence information for each chromosome was downloaded separately in FASTA format and stored under one main directory. We first aligned RNA-Seq reads to the human transcriptome. To obtain human transcriptomic data from human genome, the two annotation files, GTF and knownIsoforms, are also needed.

The UCSC Genes transcript annotations file in GTF format for UCSC hg19 version of human genome has downloaded using the UCSC's Table Browser Table in GTF format. Isoform-gene relationship information for UCSC hg19 version of human genome was obtained from UCSC Genome Browser as knownIsoforms.txt.

3.3. RNA-Seq Analysis

3.3.1. Alignment RNA-Seq reads to reference transcriptome

To measure transcript abundances in each condition, paired-end RNA-Seq reads were aligned to UCSC human transcriptome (hg19) by using Bowtie (version 0.12.7) (Langmead, Trapnell, Pop, & Salzberg, 2009) first and then quantification was done by using RSEM (version 1.3.1). RSEM (B. Li & Dewey, 2014) is a software tool for both alignment and quantification of single-end or paired-end RNA-Seq reads. This program, in its default mode, uses bowtie to align reads against a reference transcriptome and provides an alternative way to users to choose a different alignment program. Before alignment, users should prepare a reference transcriptome and genome indices by using “*rsem-prepare-reference*”. There are “- -bowtie” and “- - star” options to generate both; however, genome indices must be generated again using “*rsem-prepare-reference*” for those who uses alternative aligner and provides an alignment file.

In this study, the UCSC human genome, GTF file downloaded from UCSC Table Browser and knownIsoforms file obtained from UCSC Genome Browser were used to generate transcriptome reference using “*rsem-prepare-reference*” code. Bowtie indices were also created by supplying “- -bowtie” and “- - bowtie-path” parameters to the code. Then alignment was done by using Bowtie with the following parameters: *--chunkmbs 512 -q --phred33-quals -n 0 -l 25 -I 1 -X 2000 -p 10 -a -m 15 -S*. To be sure that the resulted alignment file satisfies the requirements mentioned in *rsem-calculate-expression* protocol, the resulted alignment files were converted to bam files by using *convert-sam-for-rsem* script. In the step of estimating gene expression, transcript-level abundances were quantified as transcript per million (TPM) and Fragments per Kilobase million (FPKM) by using *rsem-calculate-expression* script with the following

parameters: `--paired-end --alignments --fragment-length-max 1000 --bam --estimate-rspd`.

3.3.2. Different Gene Expression Analysis

After the alignment and quantification processes, differential expression analysis was conducted with EBSeq (Leng et al., 2015). EBSeq is a R package based on Bayesian inference methods and designed for differential expression analysis from RNA-Seq data. RSEM installation comes with EBSeq package in its folder named “EBSeq” needed to be compiled. The resulting gene-counts from RSEM run were used for differential gene expression analysis. Firstly, *rsem-generate-data-matrix* command was used to extract input matrix from expression results. As a result, we had three input matrices for the following pairwise comparisons: naïve vs 12d, 12d vs persister and naïve vs persister, in which columns represent biological replicates, rows stores genes and then matrix was filled with raw gene counts. Then variable genes across pairwise conditions were detected with *rsem-run-ebseq* command. The analysis gave an output file which provides four statistics; PPEE (posterior probability of equally expressed), PPDE (posterior probability of differentially expressed), PostFC, RealFC, but does not calculate log2FC. Thus, we calculated log2FC for each gene by taking log2 of RealFC column. For visualization and further analysis, we continued with all genes with posterior probability greater than 0.95 and log2FC higher than 2. The schematic representation of the differential gene expression analysis was shown in **Figure 3.3**.

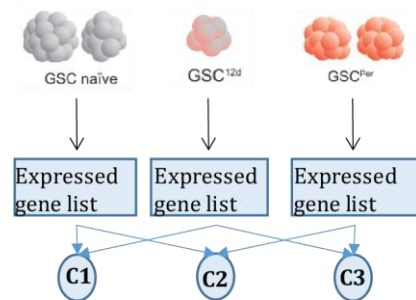


Figure 3.3: The overall representation of differential gene expression analysis. The figure indicates the result of differential gene expression analysis. At the end, there is one list for each condition (C1, C2 and C3 represent pairwise comparison conditions; naïve vs 12d, 12d vs persister, naïve vs persister, respectively). Each list stores significantly differentially expressed genes between conditions.

3.3.3. Gene Set Enrichment Analysis

To understand functional profiling of the resulting transcriptomic data and identify biological processes and/or pathways in which significantly enriched or depleted gene sets are involved, we conducted gene set enrichment analysis by using an R package WebGestaltR (version 0.3.0) (Wang, Duncan, Shi, & Zhang, 2013; Wang, Vasaikar, Shi, Greer, & Zhang, 2017; Zhang, Kirov, & Snoddy, 2005). We inputted rnk file containing two columns separated by tab: the gene list and the corresponding scores. We removed all genes which have tpm value lower than 0.1 across all samples from differentially expressed genes obtained from EBSeq and the resulting gene list generated the first column of the rnk file. Then these genes were listed based on their PPDE value in descending order and log2FC values were assigned to corresponding genes and defined as their scores in the second column of the rnk file. After prepared rnk files for all pairwise comparisons: naïve vs 12d, 12d vs persister and naïve vs persister., we run GSEA by specifying the following parameters: *enrichMethod = "GSEA", organism = "hsapiens", enrichDatabase = "KEGG pathway & gene ontology Biological processes", interestGeneType = "genesymbol", collapseMethod = "mean", minNum = 20, maxNum = 300, sigMethod = "fdr", fdrThr = 1*. Afterthought, the resulting enrichment files were subjected to FDR threshold: 0.1 and 0.25, to mark out significant pathways and biological process. Lastly, we removed generic pathways/biological processes which are most frequently enriched in many conditions.

3.4. ChIP-Seq Analysis

3.4.1. Alignment ChIP-Seq reads to reference genome

We used FastQC tool to do quality control checks on raw single-end ChIP-Seq reads. The quality check reports were good for all read files and there was no requirement to trim any sequences. Thus, we kept on with all read files for alignment with BWA (version 0.7.17) (H. Li & Durbin, 2010) and downstream analysis.

To be able to conduct proper alignment, BWA requires the FM-index for reference genome so we constructed bwa indices for the UCSC human reference genome (hg19) by using the *bwa index* command with its default parameters at first place. Then single-end ChIP-Seq reads were aligned to hg19 using *bwa aln* and *bwa samse* commands consecutively again with their default parameters. BWA outputs alignment files in the SAM format and BAM formatted files are needed for further analysis, hence the reason why resulting SAM files were

converted BAM files using Samtools (version 1.8). To reduce the effect of PCR amplification bias during the sample preparation, we used Picard MarkDuplicate tools (version 1.118) to remove PCR duplicates. After sorted the resulting alignment files using Samtools (version 1.8), we used these files as input to MarkDuplicates. The final alignment files without duplicates were subjected to differential peak calling.

3.4.2. Differential Peak Calling

After alignment single-end ChIP-Seq reads to reference genome, the next step was to measure how many reads differentially mapped to enriched regions, compared with any two conditions in our case which is called differential peak calling. Although this kind of analysis is generally used to determine differential expression genes in the gene expression analysis, to be able to define transcription factors and construct networks specific to each pairwise conditions: naïve vs 12d, 12d vs persister and naïve vs persister, we identified differential peak regions of two conditions using MACS2 (version 2.1.2).

A wide range of tools have been evolved for detection of differential enriched regions between conditions. Therefore, it is crucial to determine which tool should be used to analyse the data on hand. The kind of peaks (sharp peak or broad enrichment), presence of biological replicates and presence of predefined regions are main characteristics of ChIP-Seq reads which should be considered before selecting a tool for further analysis (Steinhauser, Kurzawa, Eils, & Herrmann, 2016). In this study, we used ChIP-Seq data specific for histone modifications and these regions are most probably broad enriched regions. In addition to that, there is no biological replicates thus we decided to use MACS2 tool for differential peak calling. The resulting alignment files were used to call peak using MACS2 *predictd* and *callpeak* modules, respectively. Firstly, we run *predictd* with its default parameters to get a uniform extension size which is the average of two fragment size, condition1 and condition2, given in the output of *predictd* module. Since differential peak calling performed for three histone modifications per pairwise comparison condition, the extension size was calculated as the average of fragment size of condition1 and condition2 for the same histone marker.

Secondly, we carried out peak calling with *callpeak* module by giving aligned file and its control file as input with additional parameters: --nomodel, --extsize for extension size. The important point in here was to keep using the same extension size for both compared conditions. After successfully running *callpeak* module, we had two output files for each condition; cond_treat_pileup.bdg and cond_control_lambda.bdg containing enriched

regions for each chromosome for treatment and control file, respectively. Besides, differential peak calling was conducted using MACS2 *bdgdiff* module with the following parameters: --t1 for condition1 treat_pileup.bdg file, --c1 for condition1 control_lambda.bdg file, --t2 for condition2 treat_pileup.bdg file, --c2 for condition2 control_lambda.bdg file, --d1 and --d2 for actual effective depths for condition 1 and 2 learned by extracting the “tags after filtering in control” line from output file of *callpeak* run.

At the end, *bdgdiff* module resulted with three differential peak files. One of those files stores regions that are highly enriched in condition 1 compared to conditions 2. Another one stores the regions having more enrichment in condition 2 over condition 1. The final file stores regions showing similar enrichment in both conditions. As a result, there are 9 bed files for H3K4me3, H3K27me3 and H3K27ac per pairwise comparison condition. **Figure 3.4** summarizes differential peak calling analysis.

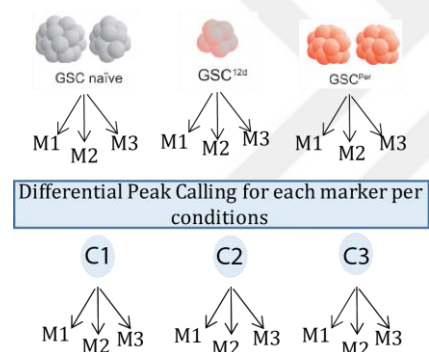


Figure 3.4: The overall representation of differential peak calling analysis. The figure indicates the result of peak calling analysis. At the end, there are three lists for each condition per markers (C1, C2 and C3 represent pairwise comparison conditions; naïve vs 12d, 12d vs persister, naïve vs persister, respectively. M1, M2 and M3 shows histone markers; H3K27ac, H3K27me3 and H3K4me3 respectively.). Each list stores significantly differentially changed peak regions specific to each marker between conditions.

3.5. Network Modelling with Omics Integrator

In this study, we mainly focused on mapping transcriptomic and epigenomic data into interaction networks and reconstructing conditions specific networks to develop an understanding for reversible transition mechanisms of GSCs under the dasatinib treatment. After properly analysed gene expression and chromatin accessibility data, we firstly identified transcription factor (TF) set using Garnet

module of Omics Integrator and then reconstructed networks by integrating them with human proteomic data using Forest module of Omics Integrator (Tuncbag, Gosline, et al., 2016).

3.5.1. Identification of transcription factors using Garnet

Garnet outputs a set of transcription factors that potentially responsible for gene expression changes by associating chromatin accessibility data and nearby expressed genes. To do this, Garnet first uses chromatin accessibility data which gives open chromatin regions, histone modification ChIP-Seq data in this case, and scans regions proximal to transcribed genes obtained from transcriptomic data within a defined window to detect transcription factor binding motifs. Although transcription factors have preferences to bind specific motifs, these proteins can tolerate a few possible base changes at certain positions in motifs. Therefore, Garnet defines transcription factor affinity (TFA) scores for each motif across all regions (**Equation 1**). Then, Garnet uses linear regression model to relate TFA scores and gene expression changes and gives a set of significant transcription factors.

$$TFA_j = \frac{\sum_i e^{w_j \cdot m_i}}{\beta_j + \sum_i e^{w_j \cdot m_i}} \quad (1)$$

In the equation, TFA_j is the estimated probability of binding for motif j which is calculated by taking scores of all possible binding windows i in the region. m_i represents the likelihood of a transcription factor to bind a region at the i^{th} window. w_j and β_j are the tuning parameters which control the probability that the motif is not false positive, and the motif is false positive, respectively.

Garnet needs four type of data as input; bed-formatted file containing open chromatin regions, fasta-formatted file stores sequence information of regions in BED file, gene expression data file as tab-delimited file and configuration file. We already prepared epigenomic data in a bed-formatted file by merging differential peak calling results for each histone marker per pairwise comparison conditions. We downloaded FASTA files by using Galaxy webserver “Extracting Genomic DNA” tool for each bed file. Then gene expression files were prepared by choosing statistically significant genes having FDR lower than 0.01 for three pairwise comparison conditions. The expression file contains two columns, first one is the name of the gene and the second one is the log-fold-change of that gene.

For each Garnet run, we created configuration files which specifies full paths to the bedfile, fastafile and expressionfile and run-related parameters: 2000 for

window size, 0.05 for pvalTresh. It should be noted that we also provided the full paths to the annotation files; genefile and xreffile, for the human genome version hg19. As the last file, the motif data in the TAMO format was specified in the configuration file. Garnet was performed for each histone marker separately and the resulting TFs lists in the output file with the extension of “regression_results.tsv” specific to each pairwise comparison condition were concatenated. The file contains four columns. These are motif column for TFs binding same motifs, slope, p-value and q-value for the regarding TF, respectively. These transcription factors are symbolized with *transfac* id and needed to be converted into official gene symbols. HGNC (HUGO Gene Nomenclature Committee) *multi-symbol checker* tool was used to retrieve official gene symbols for *transfac* ids of each TF. After that prize for each TF was calculated as negative of log₂ p-value. As a result, there were three different set of significant TFs and their prizes associated with each pairwise comparison condition: naïve vs 12d, 12d vs persister and naïve vs persister.

3.5.2. Network integration with Forest

The TFs from each pairwise comparison condition list are the molecules considered as crucially related with the mechanisms which control rapid dynamic interconversion of GSCs from sensitive, proliferative to slow-cycling, persister state. We used these TFs as terminal node set and their negative log₂ p-values as prizes and mapped them to human interactome by using Forest module of Omics Integrator.

Forest use node prizes to determine how strongly that node should be included in the final network by assigning negative weights to nodes based on a generalized prize function indicated below (**Equation 2**).

$$p'(v) = \beta \cdot p(v) - \mu \cdot degree(v) \quad (2)$$

While $p(v)$ denotes for terminal node prize where v is a vertex (node), $degree(v)$ is the number of connections that a node v has in the interactome. The parameters, β and μ , are the scaling factors to control the effect of terminal and hub nodes in the final network, respectively. To avoid negative evidences which caused from having high degree of a node just because of involving in many interactions or studying more, Forest uses these two parameters. While increasing μ makes harder to be included a hub node in the final network, increasing β means that more terminal nodes to be included in the final network.

Forest uses another scoring function to calculate the probability $p(e)$ reflecting the confidence of an edge between two proteins and allows users to avoid false positive edges. The scoring function (**Equation 3**) is:

$$c(e) = 1 - p(e) \quad (3)$$

where $c(e)$ is the cost assigned to edge based on this function by using a set of edge weights, denoted $p(e)$.

In this study, we inputted three terminal sets with prize values for each pairwise comparison conditions as text file and human interactome (iRefIndex v13) which contains interacting molecules and associated edge weight in a text file format. Forest takes the network data $G(V, E, c(e), p'(v))$ where V is the node set, E is the edge set and $c(e)$ and $p'(v)$ functions assign a cost to each edge and a prize to each node, respectively. The aim of the Forest is to find final optimal network $F(V_F, E_F)$ which minimizes the objective function (**Equation 4**):

$$f'(F) = \sum_{v \in V_F} p'(v) + \sum_{e \in E_F} c(e) + \omega \cdot k \quad (4)$$

where the ω is the parameter which controls edge cost between a dummy node and a node in the node set, N . k is the number of trees in the forest. The optimal way to construct a network is to give different values to parameters and then choose the optimal combination of the parameters that give a final network where there are maximum number of nodes from prize file.

In this study, we used different values for the parameters; 0.5 and 1 for ω , 0 and 0.01 for μ , 1, 2, 3, 4, and 5 for β and 10 for D . The combination of each value of parameters were supplied to Forest in a configuration file. In addition to prize files and human interactome, we also inputted dummy node lists containing cell surface receptors by specifying `--dummyNode` option. After run Forest, we had 20 networks for each pairwise comparison condition which were outputs of different configuration files containing different values for the parameters. Then we merged those 20 networks using python NetworkX package and obtained one final optimal network for each condition in sif-formatted file.

3.5.3. Visualization of networks with Cytoscape

The optimal final networks were visualized in Cytoscape (version 3.6.0). Each node type was specified with a different shape; triangle represents transcription factors coming from the prize file, steiner nodes are symbolized with hexagon shape and cell surface receptors are shown with V shape. Furthermore, the networks were clustered by using Cytoscape clusterMaker Community cluster (GLay). The gene expression changes, as \log_2FC values, were also added to

nodes as color scale where bluish color indicates downregulated genes and reddish color represents upregulated genes. Finally, we add histone marker information for each TFs in the final networks in the form of pie-chart. Red, purple and light blue colours represent for H3K27ac, H3K27me3 and H3K4me markers, respectively.

3.5.4. Overrepresentation Enrichment Analysis

Overrepresentation Enrichment Analysis (ORA) was performed for each cluster having more than 5 nodes per condition specific networks using WebGestaltR (version 0.3.0). Differently from GSEA, there is no need for scores in the input file so that we supplied node lists as well as reference gene file to carry out ORA. We prepared one column txt files for all clusters in each pairwise comparison specific network; there were 10 clusters out of 13 for naïve vs 12d network, 5 clusters out of 12 for naïve vs persister network and 10 out of 13 for 12d vs persister network. We run ORA by specifying the following parameters: *enrichMethod = "ORA", organism = "hsapiens", enrichDatabase = "KEGG pathway & pathway Reactome & gene ontology Biological processes & gene ontology Molecular Function", referenceSet = "genome_protein_coding", minNum = 10, maxNum = 300, sigMethod = "fdr", fdrThr = 1*. Afterthought, the resulting enrichment files were subjected to FDR threshold: 0.05 to mark out significant pathways and biological process. Lastly, we removed generic pathways/biological processes which are most frequently enriched in many conditions.

CHAPTER 4

4.RESULTS

4.1. Transcriptomic Profiling Reveals That Notch Pathway and Histone Modification Related Genes are Highly Enriched in Persister Cells

Glioblastoma stem cells show the characteristics of neural stem cells and express stemness marker genes (Justin d. lathia, 2015). These markers are mostly enriched in cell cycle-related, neuron cell differentiation-related biological functions and some of them are transcription factors that directly bind to DNA. *SOX2* (Hemmati et al., 2003), *OLIG2* (Ligon et al., 2007) are two of these markers. In this study, we analyzed three conditions of dasatinib treated GSCs which can be expressed as naïve, GSCs in their proliferative state, 12d, 12 days treated GSCs and persister, more than 8 weeks treated GSCs. The analysis of gene expression profiles revealed that *SOX*-related genes (Figure 4.1) and *OLIG2* are highly enriched in dasatinib treated cell lines, especially in persister and 12d, respectively. The results also show that dasatinib treatment trigger the expression of *SMAD3* and *SMAD7* genes. *SMAD* genes play a role in the transmission of signal from cell surface to nucleus and in the activation of transcription through TGF-beta signalling pathway (Macias, Martin-Malpartida, & Massagué, 2015). TGF-beta activated SMADs perform various functions which are negative regulation of cell growth, formation of fibrosis and modulating immune-related pathways (Weinstein, Yang, & Deng, 2000). While *SMAD3* shows higher-level expression pattern in 12d and persister status than naïve status, *SMAD7* is highly expressed in 12d status and these two genes are related to inhibiting growth factors and growth-related signals within the cell.

FOX family genes are another group regulated differentially across conditions. *FOXO3*, *FOXG1*, *FOXK1* and *FOXN3* are the genes having opposite trend when

expression profiles of them in dasatinib treated cells; 12d and persister, and in naïve cells were compared. Furthermore, expression of *FOXA3* gene is relatively lower in persister cells than naïve and 12d cells (Figure 4.2). *FOX* genes are involved in transcriptional regulation, cell growth and differentiation by coding DNA-binding FOX proteins and known as their upregulation in tumor cells (Katoh, Igarashi, Fukuda, Nakagama, & Katoh, 2013). It should be highlighted that although *FOXO3* and *FOXG1* function in negative regulation of neuron migration and differentiation, respectively and show similar expression pattern in 12d and persister cells, *FOXA3* plays a role in the regulation of neuron differentiation, chromatin remodelling and Notch signalling pathways and shows different expression pattern in persister cells. Besides indirectly regulation of Notch signalling pathways by *FOXA1*, genes directly implicated in these pathways, *NES*, *HES5*, *HEY1*, are activated in persister cells (Figure 4.2). Cell replication and differentiation, prolongation of stemness markers are the functions controlled by Notch signalling pathways (Borggreffe & Oswald, 2009). Further, some studies emphasized that Notch signalling is activated particularly in cancer stem cells (Venkatesh et al., 2018).

Genes associated with chromatin remodelling, especially the ones involving histone demethylases, showed an overexpression in 12d and persister cells. Persister GSCs, which are insensitive to dasatinib, have upregulated *KDM5B* and *KDM6B* (Figure 4.2) genes from *KDM* family. Previous studies have demonstrated significant regulation of these genes in cells that are insensitive to drug and slow-cycling (Roesch et al., 2013; Sharma et al., 2010).

The differential expression profiles were also studied to check whether the differences in the expression patterns of genes are valuable to make a biological inference. We performed the analysis on the pairwise comparison conditions; naïve sensitive cells versus 12d cells, naïve sensitive cells versus slow cycling persister cells and 12d cells versus slow cycling persister cells (Figure 4.3).

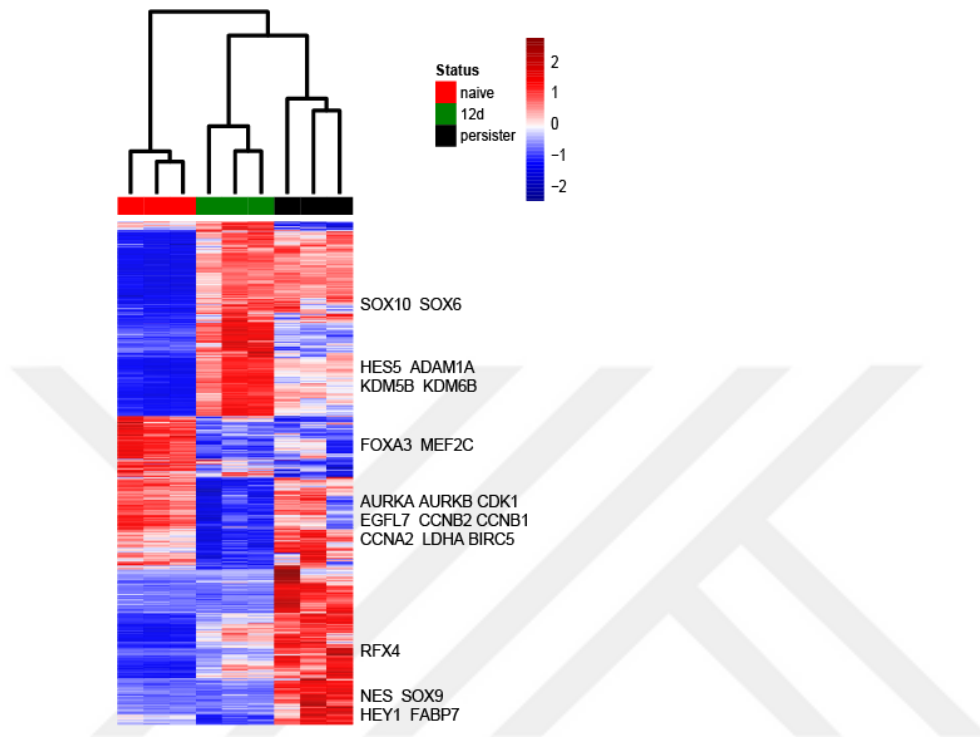


Figure 4.1: Gene expression profiles of naïve, 12d and persister cells. Heatmap indicates gene expression profiles of the most significant 1483 genes ($FDR < 0.01$, $\log_2FC > 2$) across GSC8 naïve, GSC8 12d and GSC8 persister. Genes having TPM values lower than 0.1 across all conditions were removed. Data were generated from $\log_2(\text{tpm} + 1)$ transformed TPM scores of three biological replicates. Red-blue colour scale represent z-scores from positive to negative values. Cell status are indicated by red, black and green colour.

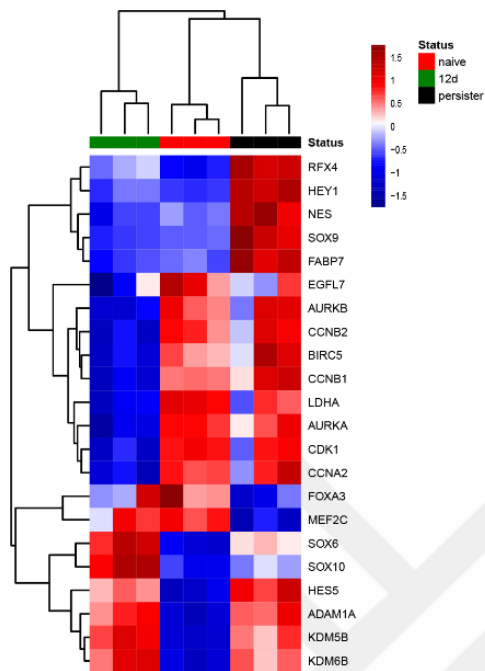


Figure 4.2: The expression profiles of some selected genes in naïve, 12d and persister cells. Heatmap indicates gene expression profiles of highlighted genes in the Figure 4.1 ($FDR < 0.01$, $\log_2FC > 2$) across GSC8 naïve, GSC8 12d and GSC8 persister. Data were generated from $\log_2(\text{tpm}+1)$ transformed TPM scores of three biological replicates. Red-blue colour scale represent z-scores from positive to negative values. Cell status are indicated by red, black and green colour.

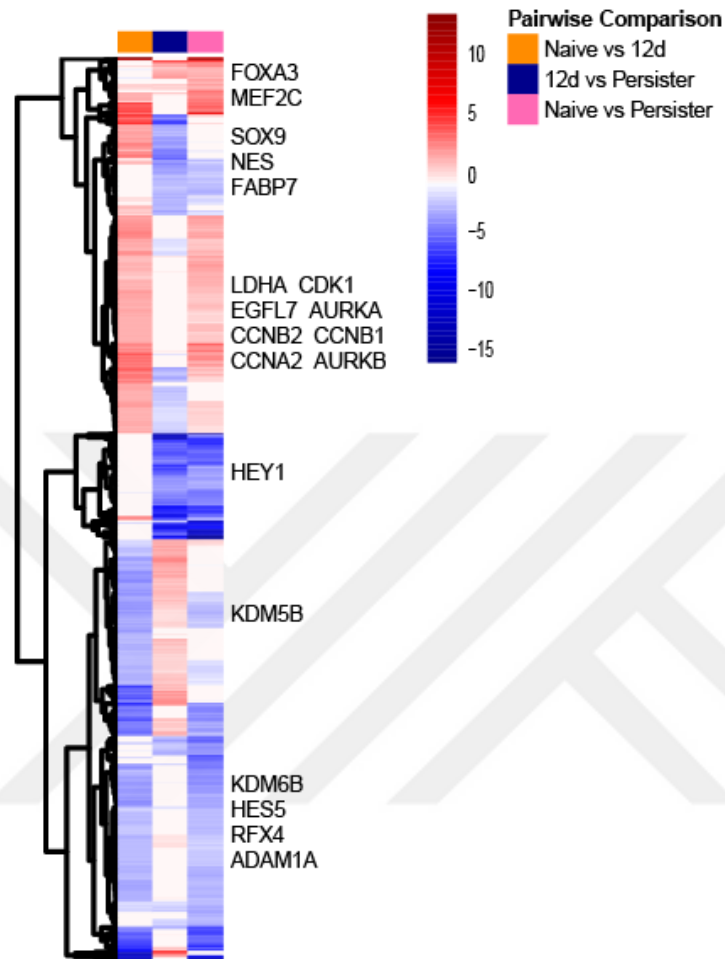


Figure 4.3: Differentially expressed genes across pairwise comparison conditions. Heatmap shows 1256 significantly differentially expressed genes between each comparison conditions; naïve vs 12d, 12d vs persister, naïve vs persister (represented by orange, dark blue and pink colours, respectively). Genes were selected based on having FDR lower than 0.01 and log₂FC higher than 2. Genes having TPM values lower than 0.1 across all conditions were removed. Red-blue colour scale represents log₂FC values.

4.2. Gene Set Enrichment Analysis Shows That Cell-Cycle Related Biological Functions are Negatively Regulated in Drug-Treated Cells

We then performed functional enrichment analysis on significantly differentially expressed genes for each pairwise comparison condition; naïve vs 12d, 12d vs persister and naïve vs persister, to identify classes of genes and pathways where these gene sets are involved. In this consideration, we are able to better

understand the biology underlying resistance of GSCs. Following our analysis, we found that mitotic cell-cycle related functions such as mitotic G1 DNA damage checkpoint1, mitotic G1/S transition checkpoint1, regulation of G2/M transition of mitotic cell cycle are negatively regulated in the pairwise comparison condition, 12d vs persister (Figure 4.5). Furthermore, cellular functions associated with the positive regulation of mitotic cell cycle transitions are negatively enriched for the same condition. However, histone modifications, chromatin remodelling, and DNA replication-independent nucleosome organization related functions are positively enriched in the pairwise comparison conditions; naive vs 12d and naive vs persister.

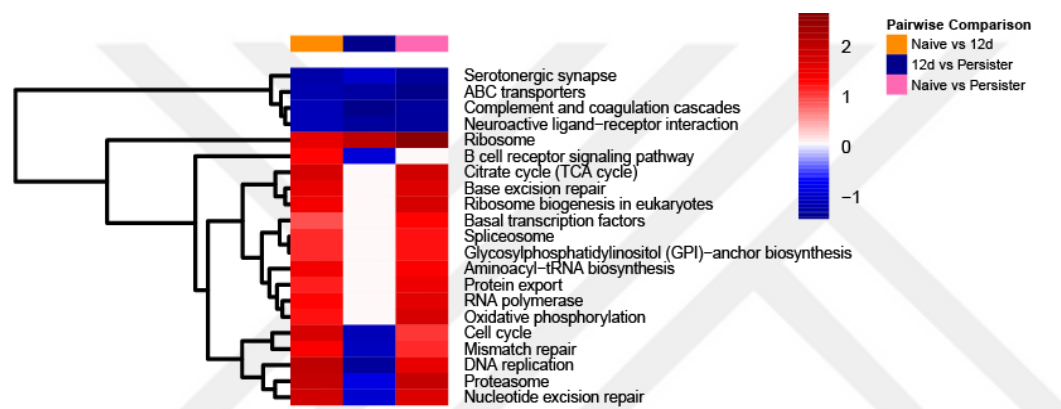


Figure 4.4: KEGG pathways enriched in each pairwise comparison condition. In the heatmap, columns show these comparison conditions; naive vs 12d, 12d vs persister and naive vs persister (represented by orange, dark blue and pink colours, respectively), and rows represent union list of enriched KEGG pathways finding by GSEA. While blue colour is to define downregulated pathways (negative enrichment score), red is to define upregulated ones (positive enrichment score). White colour means that there is no enrichment for related pathways.

4.3. Comparative Comparison of Significantly Active Transcription Factors Across Pairwise Comparison Conditions

After comprehensively analyze gene expression data, we identified significantly active TFs for each pairwise comparison condition; naive vs 12d, 12d vs persister and naive vs persister by using the Garnet module of Omics Integrator in which we integrated expression level data and chromatin accessibility data obtained from differential peak calling analysis. Then, we compared pairwise comparison conditions in terms of similarity and difference for having TFs.

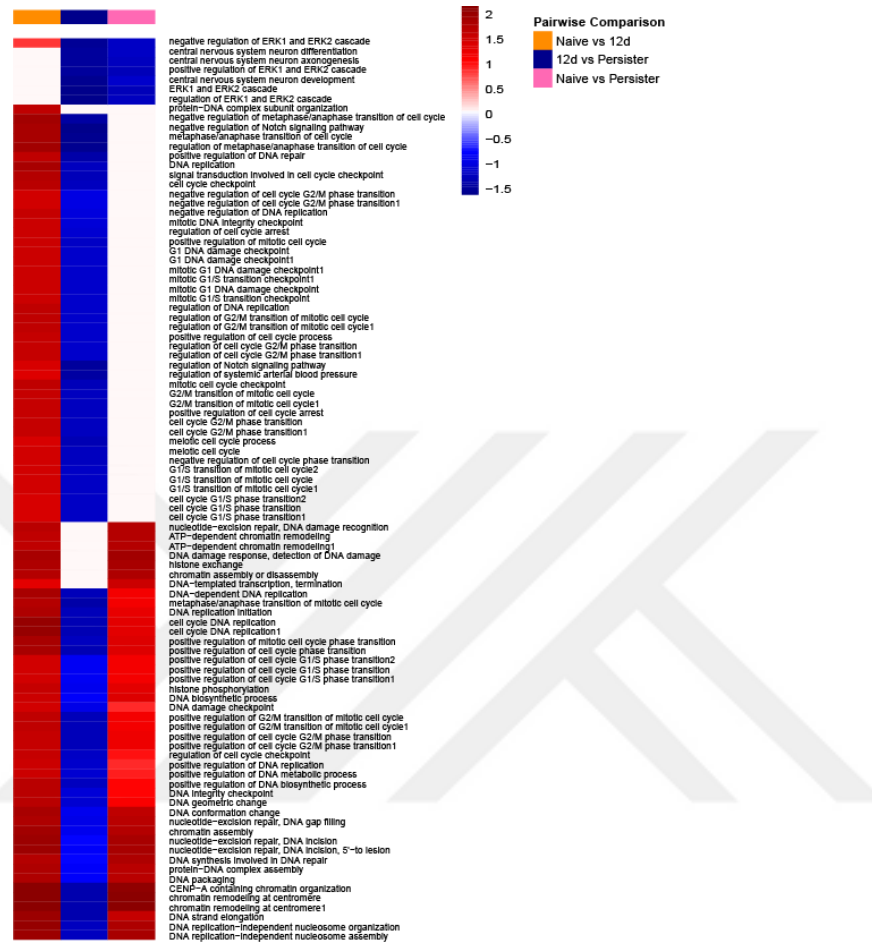


Figure 4.5: Biological functions enriched in each pairwise comparison condition. In the heatmap, columns show these comparison conditions; naive vs 12d, 12d vs persister and naive vs persister (represented by orange, dark blue and pink colours, respectively), and rows represent union list of enriched biological functions finding by GSEA. While blue colour is to define downregulated functions (negative enrichment score), red is to define upregulated ones (positive enrichment score). White colour means that there is no enrichment for related functions.

We found that different TFs are significantly active in different conditions (Figure 4.6) while fewer of them are shared by all three pairwise comparison conditions (Figure 4.7). Interestingly, cell growth and cell differentiation related TFs, EGFR family, WT1, ELF family, are found in the naive vs 12d and 12d vs persister comparison conditions. Montano and colleagues demonstrated that there is a negative regulation on apoptotic cell functions in the EGFRvIII-negative GBM cells promoted by EGFRvIII-positive GSCs (Montano et al.,

2011). EGFRvIII positive GBM cells are variants and generally associated with high level expression of cell proliferation genes. They also revealed that EGFRvIII-negative cells are resistant to certain treatments. Our findings with the result of Montono's work emphasized that cell growth related genes may be in a strong relation with genes which are responsible for GBM resistance. In addition, SOX TFs, SOX21, SOX10, SOX9, SOX7, SOX6, SOX30 and SOX14 are only found in the naive vs 12d comparison condition. These TFs play roles in the regulation of cell differentiation, cell migration and negative regulation of apoptotic process and canonical Wnt signalling pathway which is another well studied crucial pathway in GSCs (Sandberg et al., 2013).

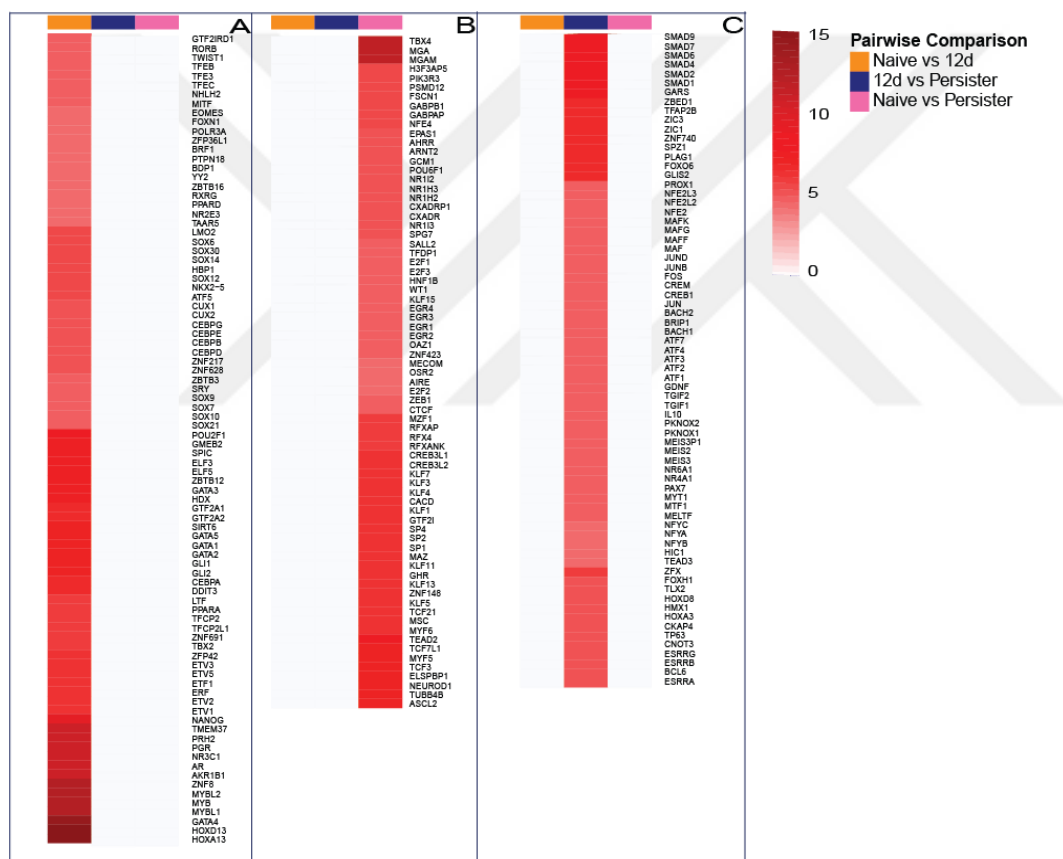


Figure 4.6: Significantly active TFs in each pairwise comparison condition. In this heatmap, each panel A, B and C, shows significantly active TFs detected by the Garnet module for each pairwise comparison condition; naive vs 12d, 12d vs persister and naive vs persister (represented by orange, dark blue and pink colours, respectively). Colour bar represent weight of each TF which is related with the importance of associated TF.

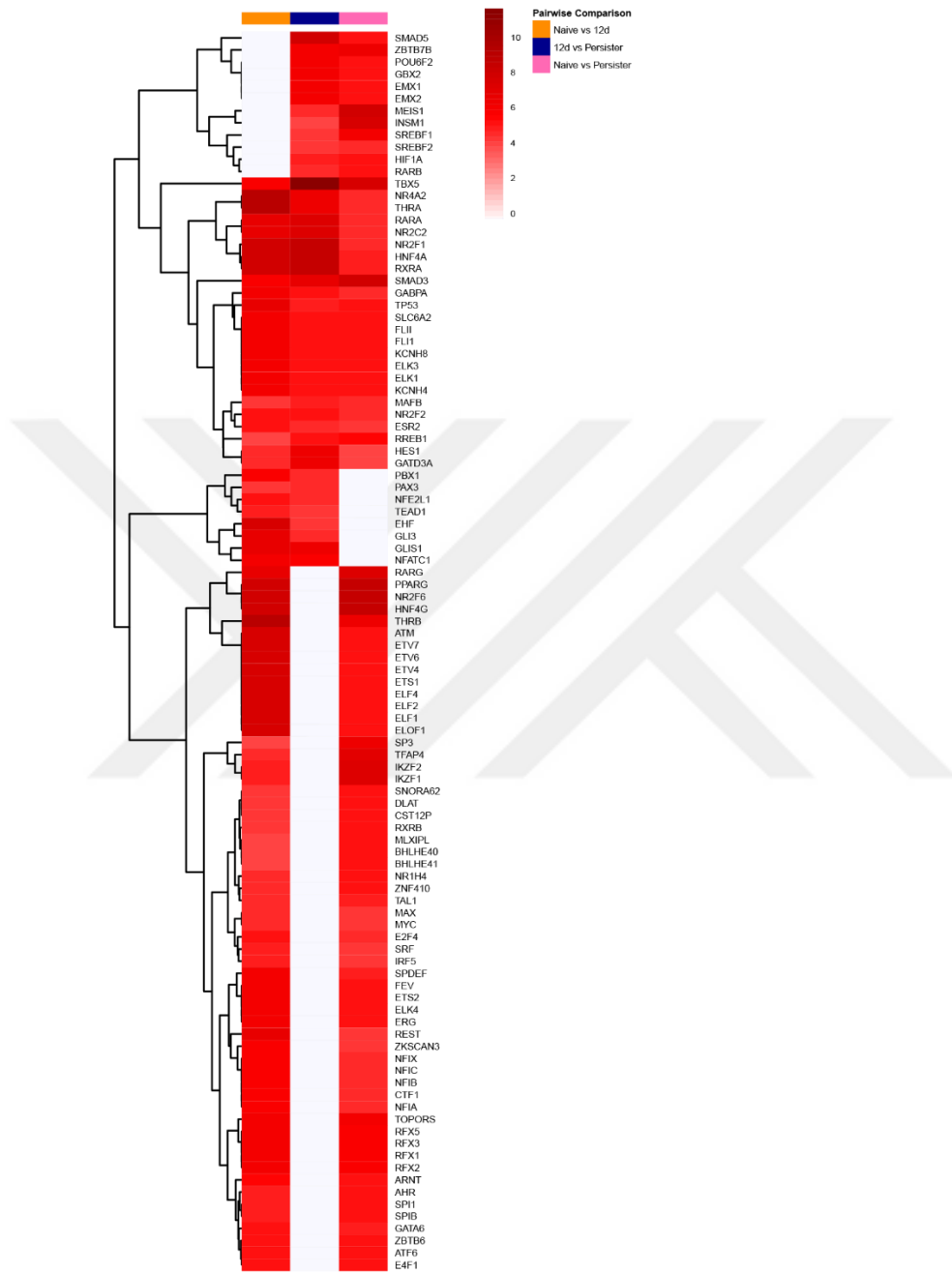


Figure 4.7: Significantly active TFs found in at least two pairwise comparison conditions. The heatmap significantly active TFs detected by the Garnet module for each pairwise comparison condition; naïve vs 12d, 12d vs persister and naïve vs persister (represented by orange, dark blue and pink colours, respectively). Colour bar represent weight of each TF which is related with the importance of associated TF.

4.4. Network Modelling Indicates That GSCs Prefer an Alternative Cell Surface Receptors to Activate RTK-dependent Pathways

We next integrated omic datasets (epigenomic and transcriptomic data) to reconstruct the signaling and regulatory networks for each condition and further compared these networks to elucidate the commonalities and differences in each condition. We need to note that the network reconstruction is important in two terms: i. we are able to analyze multiple omic data together ii. we are able to show the interactions between omic entities beyond the list of genes or proteins. In this way, it is possible to analyze the conditions at pathway level. For this purpose, we used Omics Integrator software. The Garnet module is used to integrate ChIP-Seq data with RNA-Seq data to obtain significantly active transcription factors and the Forest module is used to reconstruct signaling networks starting from cell surface receptors and terminating at the significant transcription factors found by Garnet. As a result, “naive vs 12d” network has 188 node and 394 edges, “12d vs persister” network has 188 node and 407 edges, and “naive vs persister” network has 111 node and 281 edges.

Rheinbay et al. highlighted that Wnt signalling is crucial in GSCs and highly associated with tumour propagation (Rheinbay et al., 2013). Dasatinib inhibits receptor tyrosine kinases and its specific targets are BCR/ABL and Src family receptor tyrosine kinases (Das et al., 2006; Talpaz et al., 2006). Among the Src family, SRC, LCK, YES FYN, PDGFR β , C-KIT and tPHA2 are the main targets while SRC involves in the positive regulation of canonical Wnt pathway. In this study, network modelling of pairwise comparison conditions reveals that GSCs maintain their tumorigenicity and show resistance to dasatinib treatment by activating Wnt signalling pathway via a different cell surface receptor.

Among the cell surface receptors, SELL, LIFR, IL6ST and BAMBI are the receptors enriched in the resulting network for the pairwise comparison condition naïve vs 12d (Figure 4.4A). The activation of BAMBI receptor activates proteins performing function in the positive regulation of canonical Wnt signalling. In addition to role of BAMBI in the activation Wnt signalling pathway, IL6ST receptor enriched in the naïve vs 12d network plays a role in the Notch signalling which is known with its relatively high activity in stem cells. IL6ST is also related with cell proliferation and target for growth factor.

TGF- β signalling pathway has been studied well and these studies shed light on the importance of TGF- β signalling pathway in the activation of stemness markers in GSCs (Ikushima et al., 2009). Although BAMBI receptor positively regulates tumour related pathways, Wnt signalling, this receptor negatively affects the regulation of TGF- β signalling pathway. Besides, it cannot be ignored that there is a general trend for binding glycoproteins and being in a relation with

energy-related pathways for the receptors enriched in the naïve vs 12d condition specific network.

Ras/Raf/MAPK signalling is another glioma related pathway which includes regulation of cell proliferation, differentiation and survival (Halfon et al., 2000; Tuncbag, Milani, et al., 2016). The network of naïve vs persister condition represents the interactions of significantly changing transcription factors between sensitive and insensitive cells (Figure 4.10) and it contains of a transmembrane receptor, GFRA2. The receptor GFRA2 is involved in the tyrosine kinase signalling pathway which is inhibited by dasatinib binding to other receptor tyrosine kinase receptors. Additionally, it activates Raf/MAPK cascade as well as nervous system development. Furthermore, dasatinib targeted RTKs particularly act in the positive regulation of MAP kinases and MAPK cascade. Besides, GFRA2 activates glial cell line-derived neurotrophic factor (GDNF) and positively regulate the survival and differentiation of neurons. Other two receptors, IL10RA and IFNAR2, in the naïve vs persister network are known as receptors that regulates JAK/STAT signalling pathway. Another receptor, ITGAX, is shared the characteristics of binding integrins with the RTKs.

There is a general trend among receptors in the 12d vs persister condition specific network (Figure 4.9) for binding growth factor and regulating the growth-related pathways. The receptor, GHR, plays a role in the indirect activation of MAPK signalling by activating JAK/STAT pathway upon protein kinase binding. Another receptor, NGFR, can be important for the stemness maintenance and drug resistance of GSCs in their insensitive status in terms of being involved in binding to nerve growth factor, cell survival and differentiation.

The other proteins in the resulting optimal networks, transcription factors and proteins called as Steiner node, are also crucial to understand how GSCs become persistent under the dasatinib treatment. For the naïve vs 12d network (Figure 4.8), neural stem cell related markers; histone demethylases, SMAD proteins and SOX transcription factors, are highly connected to other proteins in the network.

For naïve vs persister network (Figure 4.10) SMAD family transcription factors are grouped in a cluster and make interactions with the member of another cluster which is regulated by the cell surface receptor, ITGAX in the network. The second cluster in the interaction of two clusters, involves MAPK10 protein and MAF transcription factor.

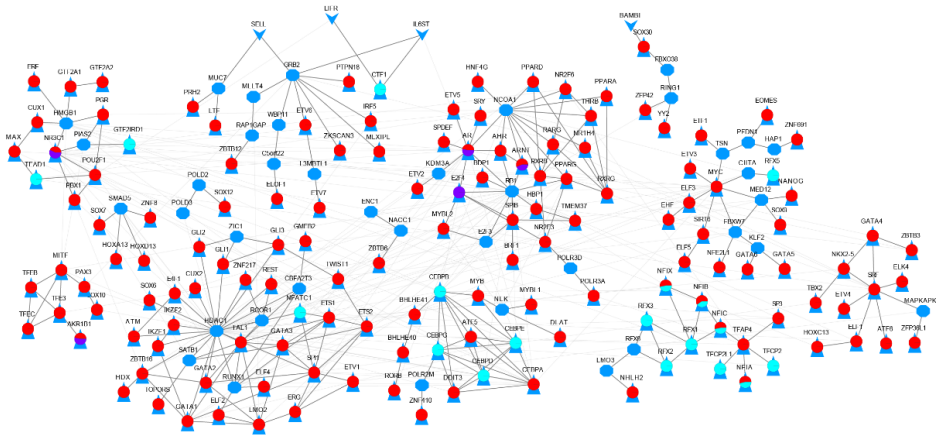


Figure 4.8: Condition specific network for the pairwise comparison condition naïve vs 12d. The network was constructed by merging all augmented network results from Forest with μ 0.01. There are 188 nodes and 394 edges. Triangles with a pie chart represent TF, hexagons are Steiner nodes and V shape shows cell surface receptors. Each pie chart on TFs represents the histone mark that the TF is obtained from. The colours of the pie chart represent different histone markers; red for H3K27ac, purple for H3K27me3 and light blue for H3K4me3.

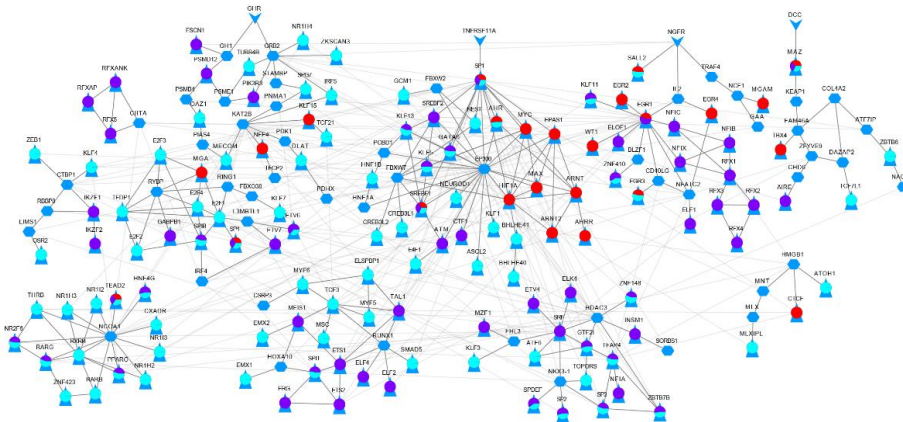


Figure 4.9: Condition specific network of the pairwise comparison condition 12d vs persister. The network was constructed by merging all augmented network results from Forest with μ 0.01. There are 188 nodes and 407 edges. Triangles with a pie chart represent TF, hexagons are Steiner nodes and V shape shows cell surface receptors. Each pie chart on TFs represents the histone mark that the TF is obtained from. The colours of the pie chart represent different histone markers; red for H3K27ac, purple for H3K27me3 and light blue for H3K4me3.

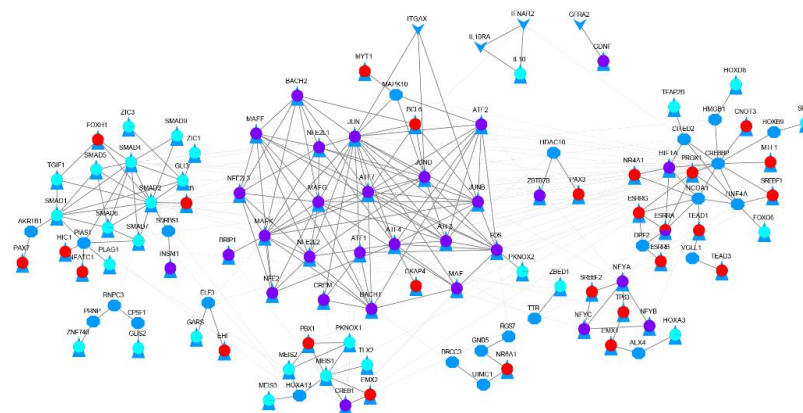


Figure 4.10: Condition specific network of the pairwise comparison condition naïve vs persister. The network was constructed by merging all augmented network results from Forest with μ 0.01. There are 111 nodes and 281 edges. Triangles with a pie chart represent TFs, hexagons are Steiner nodes and V shape shows cell surface receptors. Each pie chart on TFs represents the histone mark that the TF is obtained from. The colours of the pie chart represent different histone markers; red for H3K27ac, purple for H3K27me3 and light blue for H3K4me3.

4.5. Comparison of The Results of Overrepresentation Enrichment Analysis of Clusters in Each Condition Specific Network

One of the most important characteristics of Glioblastoma is its phenotypic heterogeneity that explains why developing a treatment for GBM is hard. From the aspects of system biology, pathway-based analysis can be meaningful to understand similarities and differences among condition specific networks. We first performed cluster analysis on the three condition specific networks; naïve vs 12d, 12d vs persister and naïve vs persister. Then we identified functions and pathways in each cluster which are overrepresented in a set of TFs. To conclude with the analysis, we compared commonalities and differences of enriched pathways (KEGG & Reactome pathways), biological processes and molecular functions across clusters in each condition specific network.

Biological process analysis reveals that cell cycle related functions are negatively regulated in the 12 vs persister condition network, while cell functions related to stem cell differentiation and proliferation are enriched in naïve vs 12d and naïve vs persister condition networks (Figure 4.14). Histone modification related biological process are active both in the naïve vs 12d and in the 12d vs persister networks while there is no any enrichment for this

category in the networks specific to naïve vs persister. Therefore, chromatin remodelling should be crucial for the GSCs subjected to 12 days dasatinib treatment. However, sensitive and insensitive glioblastoma stem cells have tendency to regulate biological process associated with stemness.



Figure 4.11: Pathway analysis across clusters in the naïve vs 12d network. The representation shows enriched KEGG and Reactome pathways in each cluster. The overrepresentation enrichment analysis was conducted by WebGestaltR. FDR value shows the significance of the related category while size of the points is associated with the number of genes involved in each category.

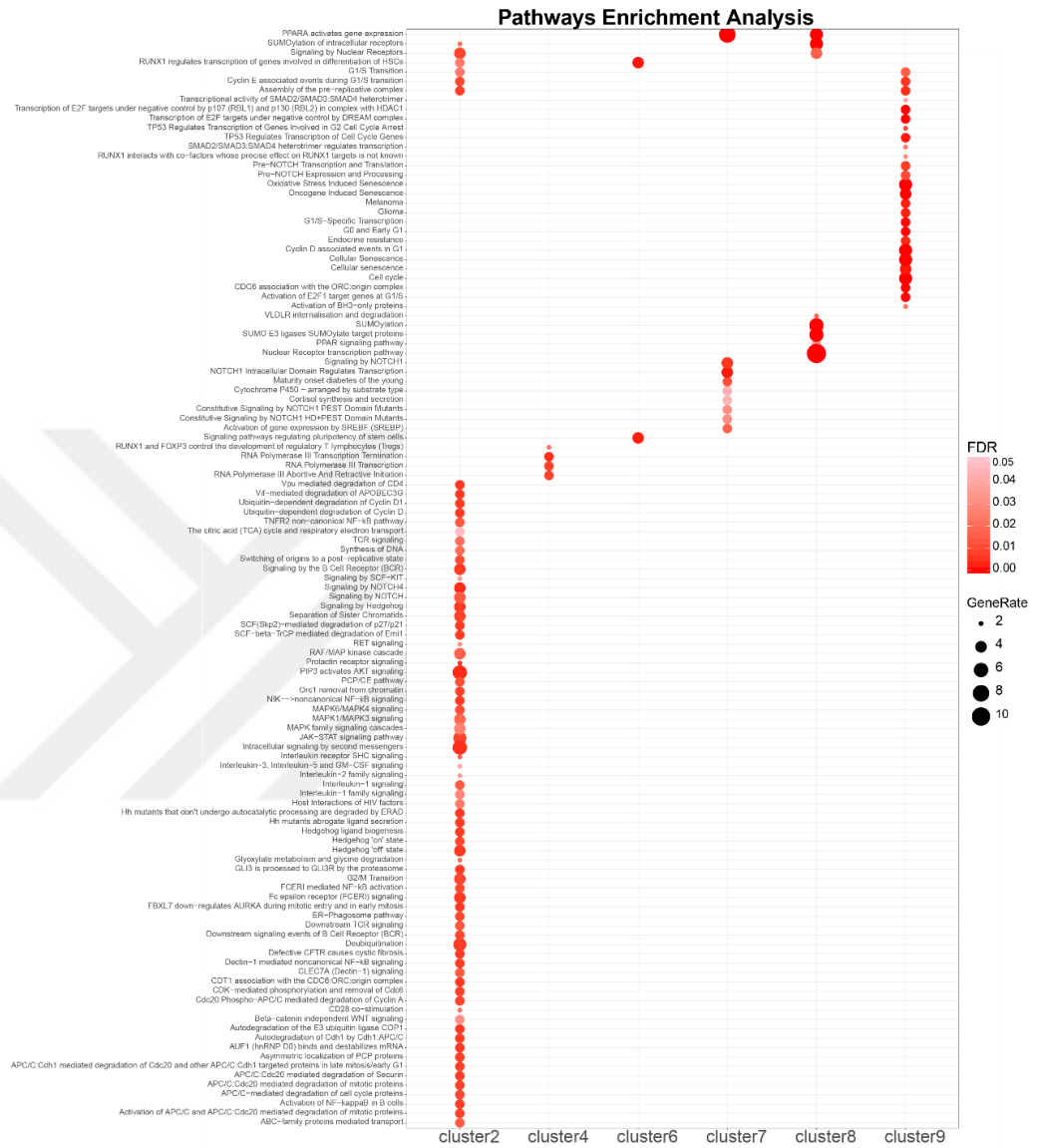


Figure 4.12: Pathway analysis across clusters in the 12d vs persister network. The representation shows enriched KEGG and Reactome pathways in each cluster. The overrepresentation enrichment analysis was conducted by WebGestaltR. FDR value shows the significance of the related category while size of the points is associated with the number of genes involved in each category.

Cell cycle- and Notch signalling-related pathways shows a distinctive enrichment trend when the 12d vs persister specific network is compared to other two networks; naïve vs 12d and naïve vs persister (Figure 4.16). Moreover, TGF- β activated SMAD proteins related pathways are differentially active in the 12d vs persister condition network which are negatively regulate cell cycle, cell growth functions within the cell.

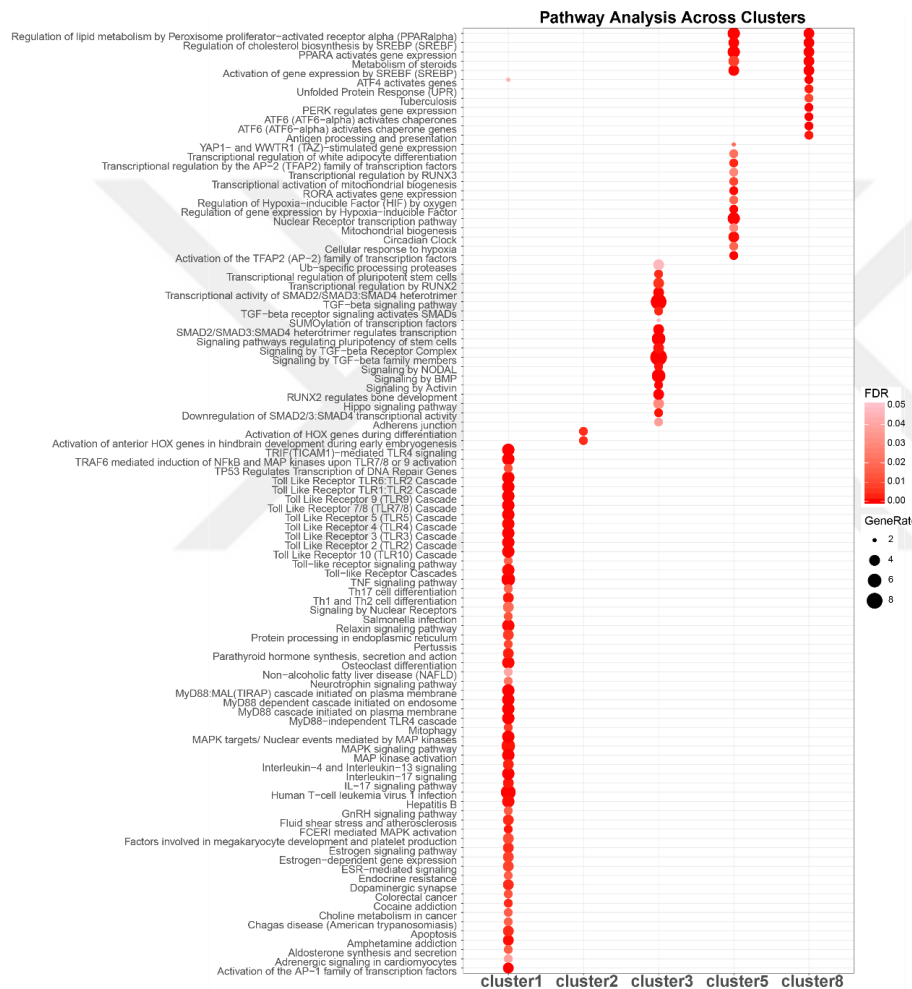


Figure 4.13: Pathway analysis across clusters in the naïve vs persister network. The representation shows enriched KEGG and Reactome pathways in each cluster. The overrepresentation enrichment analysis was conducted by WebGestaltR. FDR value shows the significance of the related category while size of the points is associated with the number of genes involved in each category.

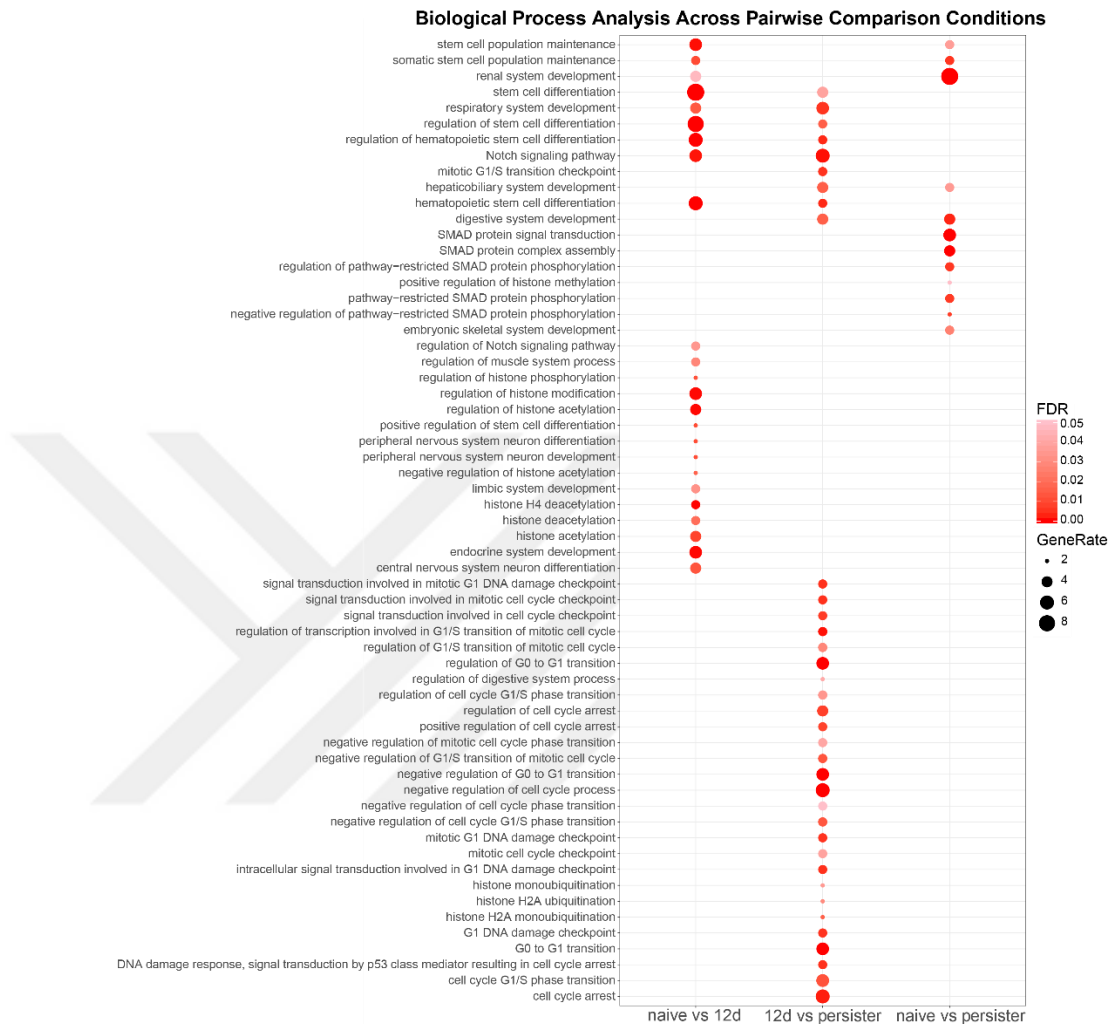


Figure 4.14: Biological process analysis across pairwise comparison condition networks. The representation shows enriched biological processes in each network; naïve vs 12d, 12d vs persister and naïve vs persister. The overrepresentation enrichment analysis was conducted by WebGestaltR. FDR value shows the significance of related category while size of the points is associated with the number of genes involved in each category.



Figure 4.15: Molecular function analysis across pairwise comparison condition networks. The representation shows enriched molecular functions in each network; naïve vs 12d, 12d vs persister and naïve vs persister. The overrepresentation enrichment analysis was conducted by WebGestaltR. FDR value shows the significance of related category while size of the points is associated with the number of genes involved in each category.

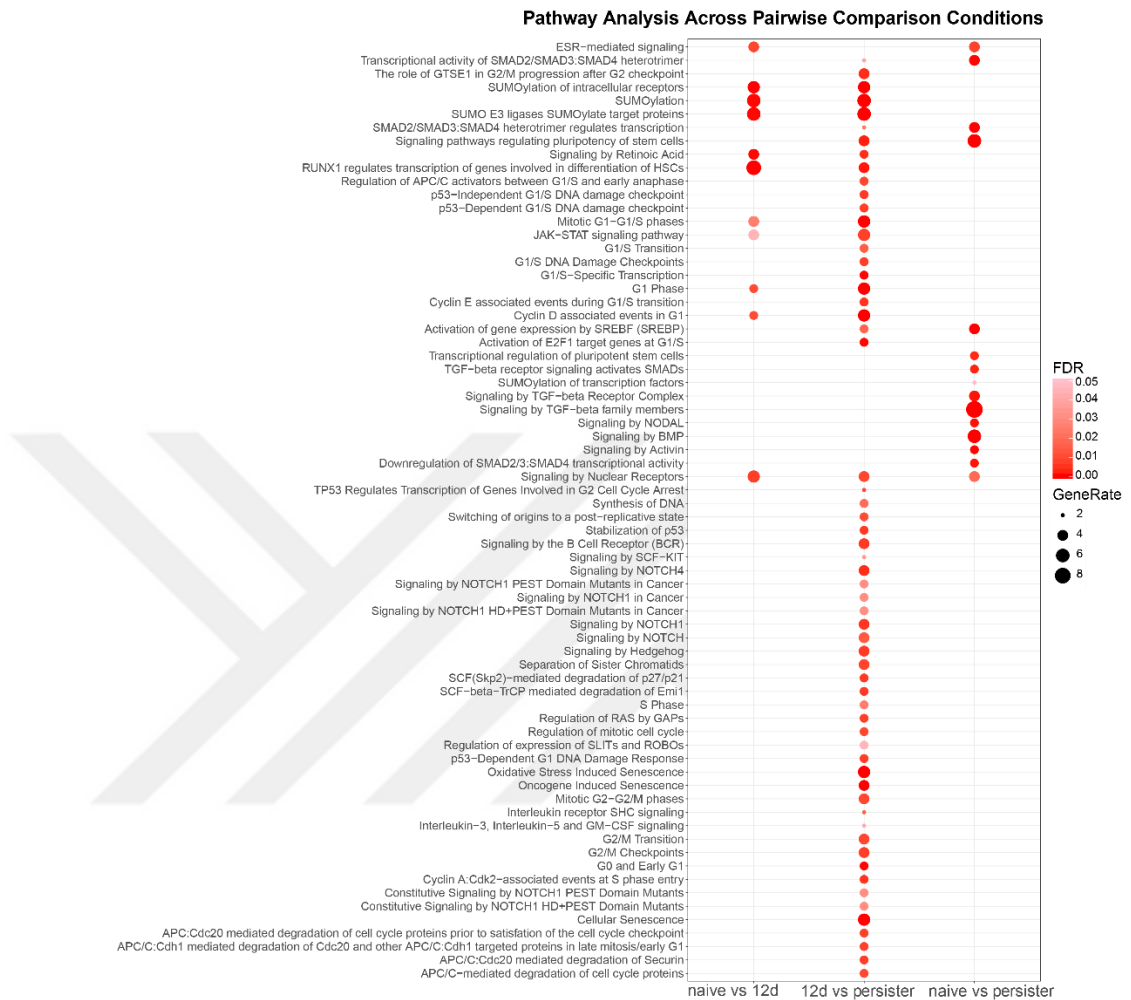


Figure 4.16: Pathway analysis across pairwise comparison conditions. The representation shows enriched KEGG and Reactome pathways in each network; naïve vs 12d, 12d vs persister and naïve vs persister. The overrepresentation enrichment analysis was conducted by WebGestaltR. FDR value shows the significance of related category while size of the points is associated with the number of genes involved in each category.



CHAPTER 5

5. DISCUSSION AND CONCLUSION

In this study, by using gene expression changes and chromatin accessibility data we model reversible transition of GSCs from proliferative to slow-cycling, persister states under the effect of dasatinib treatment in such a way that this extensive network modelling led us to map transcriptomic and epigenetic changes to human interactome. By taking the existence of specific three conditions into account, specifically naïve, 12d-dasatinib treated and persister states, the networks for each pairwise comparison condition were constructed.

After conducting the transcriptomic analysis, we found that Notch signalling related genes show augmented level of expression in the GSCs which are in persister state (Figure 4.1). The relation between Notch-related genes and cell differentiation and resistance in cancer was also emphasized by Borggreffe and his colleagues (Borggreffe & Oswald, 2009). In addition, we also figured out that two histone demethylases genes, *kdm5b* and *kdm6b*, which demethylases H3K4 and H3K27 respectively, are highly expressed in persister GSCs (Figure 4.1). These findings are particularly important in terms of biological mechanisms underlying GSCs resistance; specially, the main reason is that epigenetic mechanisms are better candidate to explain the reversible mechanisms in GSCs than genetic mechanisms. The significantly differential expression pattern of these histone modification genes reveals that different chromatin structural states may be responsible for the regulation of gene expression which causes to generate the phenotypic resistance. Furthermore, in support of the fact that the resistance of GSCs shows the dependence on chromatin reorganization, Bernstein et al. reached a similar conclusion by stressing it in their work (Liau et al., 2017). Another significant result is that GSCs in their naïve state are closer to those in persister state than 12d-dasatinib treated GSCs when the expression pattern of significantly altered genes of the states are compared correspondingly (Figure 4.2). These results enable us to make some relations, especially such as the connection between intra-tumoral heterogeneity of GBM and therapeutic resistance (Inda et al., 2010). Pre-existence of sub-clones resistant to dasatinib treatment may be responsible for the closeness between the gene expression patterns of naïve and persister GSCs; this can be explained by the fact that before

treatment, naïve state not only includes the sub-clones which are resistance against dasatinib but also those are sensitive to it. Under the treatment, it can be observed that the sensitive cells are killed by the drug while resistant cells keep their formation in persister state without losing their stability.

In this study, we used Garnet module of the Omics Integrator which correlates transcriptomic changes with the transcription factor binding affinity. It searches TF binding motifs on the accessible regions of DNA obtained from histone marker specific epigenomic data and identifies significantly active TFs. In spite of using the single omic data type, the Garnet module integrates two types of omics data (A.A. et al., 2006). Our results are compatible with the intra-tumoral heterogeneity of GBM, we found that different transcription factors which are main regulators of gene expression, are responsible for the changes between pairwise comparison conditions (Figure 4.6). In the network construction, we revealed not only the existence of direct interactions between these significantly active TFs, but also the distribution of histone modifications across TF binding motifs. In this way, we are able to interpret the network topology in a way that how the pathways, which are enriched in each network, and which are affected as a result of the epigenetic regulation or specifically, under the effect of histone modifications.

We also found that MAPK signalling pathway is enriched in a large cluster in the network that represents the alterations between naive and persister GSCs, and the TFs are enriched in the ones obtained from H3K27me3 markers, leading to a heterochromatin-like chromatin structure, tightly packed DNA regions (Figure 4.13). Controversially, the binding motifs of TFs in other networks are located at the open chromatin regions where H3K27ac and H3K4me3 markers are enriched (Figure 4.8 and 4.9). Although previous studies emphasized that aberrant activation of MAPK signalling pathway is closely related with cell invasion and proliferation in GBM (Wilson & Filipp, 2018), by taking advantage of integrating histone modification data into TFs network, we can conclude that MAPK signalling may be the key pathway that plays a significant role in the reversible transition of GSCs from proliferative to slow-cycling state.

In summary, much more effort on epigenetic heterogeneity may serve well to understand better why the GBM is more resistant to current therapies in comparison with the other cancer types. In addition to these significant results, it can be stated that regulation of histone modifications is more active in the enhancer regions. It can be concluded that considerable amount of work is needed to identify the roles of enhancer and silencer regions on the regulation gene expression. This may allow us to possess a comprehensive understanding on GBM resistance mechanisms and also more precise and accurate personalized treatment.

REFERENCES

- A.A., M., I., N., K., B., C., W., G., S., R.D., F., & A., C. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(SUPPL.1). <https://doi.org/10.1186/1471-2105-7-S1-S7>
- Alberta, J. A., Maire, C. L., Golebiowski, D., Kane, M. F., Huillard, E., Rowitch, D. H., ... Harrington, E. P. (2011). The Central Nervous System-Restricted Transcription Factor Olig2 Opposes p53 Responses to Genotoxic Damage in Neural Progenitors and Malignant Glioma. *Cancer Cell*, 19(3), 359–371. <https://doi.org/10.1016/j.ccr.2011.01.035>
- Cisek, K., Krochmal, M., Klein, J., & Mischak, H. (2016). The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease. *Nephrology Dialysis Transplantation*, 31(12), 2003–2011. <https://doi.org/10.1093/ndt/gfv364>
- Das, J., Chen, P., Norris, D., Padmanabha, R., Lin, J., Moquin, R. V., ... Barrish, J. C. (2006). 2-Aminothiazole as a Novel Kinase Inhibitor Template. Structure–Activity Relationship Studies toward the Discovery of *N*-(2-Chloro-6-methylphenyl)-2-[[6-[4-(2-hydroxyethyl)-1-piperazinyl]-2-methyl-4-pyrimidinyl]amino]-1,3-thiazole-5-carboxamide (Dasatinib, BMS-354825) as a Potent *pan*-Src Kinase Inhibitor. *Journal of Medicinal Chemistry*, 49(23), 6819–6832. <https://doi.org/10.1021/jm060727j>
- de Tayrac, M., Le, S., Aubry, M., Mosser, J., & Husson, F. (2009). Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics*, 10(1), 32. <https://doi.org/10.1186/1471-2164-10-32>
- Dirks, R. A. M., Stunnenberg, H. G., & Marks, H. (2016). Genome-wide epigenomic profiling for biomarker discovery. *Clinical Epigenetics*, 8(1), 1–17. <https://doi.org/10.1186/s13148-016-0284-4>
- Eder, K., & Kalman, B. (2014). Molecular Heterogeneity of Glioblastoma and its Clinical

- Relevance. *Pathology and Oncology Research*, 20(4), 777–787. <https://doi.org/10.1007/s12253-014-9833-3>
- Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25), 14863–14868. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9843981>
- Fan, X., Khaki, L., Zhu, T. S., Soules, M. E., Talsma, C. E., Gul, N., ... Eberhart, C. G. (2009). Notch Pathway Blockade Depletes CD133-Positive Glioblastoma Cells and Inhibits Growth of Tumor Neurospheres and Xenografts. *Stem Cells*, 28(1), N/A–N/A. <https://doi.org/10.1002/stem.254>
- Flensburg, C., Kinkel, S. A., Keniry, A., Blewitt, M. E., & Oshlack, A. (2014). A comparison of control samples for ChIP-seq of histone modifications. *Frontiers in Genetics*, 5(SEP), 1–8. <https://doi.org/10.3389/fgene.2014.00329>
- Floratos, A., Smith, K., Ji, Z., Watkinson, J., & Califano, A. (2010). geWorkbench: an open source platform for integrative genomics. *Bioinformatics*, 26(14), 1779–1780. <https://doi.org/10.1093/bioinformatics/btq282>
- Friedman, N. (2004). Inferring Cellular Networks Using Probabilistic Graphical Models. *Science*, 303(5659), 799–805. <https://doi.org/10.1126/science.1094068>
- Galli, R., Binda, E., Orfanelli, U., Cipelletti, B., Gritti, A., De Vitis, S., ... Vescovi, A. (2004). Isolation and Characterization of Tumorigenic, Stem-like Neural Precursors from Human Glioblastoma. *Cancer Research*, 64(19), 7011–7021. <https://doi.org/10.1158/0008-5472.CAN-04-1364>
- Haider, S., & Pal, R. (n.d.). Integrated Analysis of Transcriptomic and Proteomic Data. Retrieved from <https://www.ingentaconnect.com/content/ben/cg/2013/00000014/00000002/art00003>
- Halfon, M. S., Carmena, A., Gisselbrecht, S., Sackerson, C. M., Jiménez, F., Baylies, M. K., & Michelson, A. M. (2000). Ras Pathway Specificity Is Determined by the Integration of Multiple Signal-Activated and Tissue-Restricted Transcription Factors. *Cell*, 103(1), 63–74. [https://doi.org/10.1016/S0092-8674\(00\)00105-7](https://doi.org/10.1016/S0092-8674(00)00105-7)
- Huang, Z., Cheng, L., Guryanova, O. A., Wu, Q., & Bao, S. (2010). Cancer stem cells in glioblastoma-molecular signaling and therapeutic targeting. *Protein and Cell*, 1(7), 638–655. <https://doi.org/10.1007/s13238-010-0078-y>
- Huse, J. T., Holland, E., & DeAngelis, L. M. (2013). Glioblastoma: Molecular Analysis and Clinical Implications. *Annual Review of Medicine*, 64(1), 59–70. <https://doi.org/10.1146/annurev-med-100711-143028>

- Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., ... Hood, L. (2001). Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science*, 292(5518), 929–934. <https://doi.org/10.1126/science.292.5518.929>
- Ikushima, H., Todo, T., Ino, Y., Takahashi, M., Miyazawa, K., & Miyazono, K. (2009). Autocrine TGF- β Signaling Maintains Tumorigenicity of Glioma-Initiating Cells through Sry-Related HMG-Box Factors. *Cell Stem Cell*, 5(5), 504–514. <https://doi.org/10.1016/j.stem.2009.08.018>
- Inda, M.-M., Bonavia, R., Mukasa, A., Narita, Y., Sah, D. W. Y., Vandenberg, S., ... Furnari, F. (2010). Tumor heterogeneity is an active process maintained by a mutant EGFR-induced cytokine circuit in glioblastoma. *Genes & Development*, 24(16), 1731–1745. <https://doi.org/10.1101/gad.1890510>
- Jun, H. J., Acquaviva, J., Chi, D., Lessard, J., Zhu, H., Woolfenden, S., ... Charest, A. (2012). Acquired MET expression confers resistance to EGFR inhibition in a mouse model of glioblastoma multiforme. *Oncogene*, 31(25), 3039–3050. <https://doi.org/10.1038/onc.2011.474>
- justin d. lathia. (2015). Cancer stem cells in glioblastoma Justin. *Genes and Development*, 26(10), 758. <https://doi.org/10.1101/gad.261982.115.tumors>
- Kadarmideen, H. N., von Rohr, P., & Janss, L. L. G. (2006). From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. *Mammalian Genome*, 17(6), 548–564. <https://doi.org/10.1007/s00335-005-0169-x>
- Kalkan, R. (2015). Glioblastoma stem cells as a new therapeutic target for glioblastoma. *Clinical Medicine Insights: Oncology*, 9, 95–103. <https://doi.org/10.4137/CMO.S30271>
- Lachmann, A., Giorgi, F. M., Lopez, G., & Califano, A. (2016). ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics (Oxford, England)*, 32(14), 2233–2235. <https://doi.org/10.1093/bioinformatics/btw216>
- Laks, D. R., Masterman-smith, M., Visnyei, K., Angenieux, B., Lazareff, J. A., Mischel, P. S., ... Horvath, S. (2011). NIH Public Access, 27(4), 980–987. <https://doi.org/10.1002/stem.15.Neurosphere>
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3). <https://doi.org/10.1186/gb-2009-10-3-r25>
- Lathia, J. D., Mack, S. C., Mulkearns-Hubert, E. E., Valentim, C. L. L., & Rich, J. N.

- (2015). Cancer stem cells in glioblastoma. *Genes & Development*, 29(12), 1203–1217. <https://doi.org/10.1101/gad.261982.115>
- Lefebvre, C., Rajbhandari, P., Alvarez, M. J., Bandaru, P., Lim, W. K., Sato, M., ... Califano, A. (2010). A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular Systems Biology*, 6, 377. <https://doi.org/10.1038/msb.2010.31>
- Lemée, J.-M., Clavreul, A., Aubry, M., Com, E., de Tayrac, M., Mosser, J., & Menei, P. (2018). Integration of transcriptome and proteome profiles in glioblastoma: looking for the missing link. *BMC Molecular Biology*, 19(1), 13. <https://doi.org/10.1186/s12867-018-0115-6>
- Leng, N., Li, Y., McIntosh, B. E., Nguyen, B. K., Duffin, B., Tian, S., ... Kendziorski, C. (2015). EBSeq-HMM: A Bayesian approach for identifying gene-expression changes in ordered RNA-seq experiments. *Bioinformatics*, 31(16), 2614–2622. <https://doi.org/10.1093/bioinformatics/btv193>
- Li, B., & Dewey, C. N. (2014). RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *Bioinformatics: The Impact of Accurate Quantification on Proteomic and Genetic Analysis and Research*, 41–74. <https://doi.org/10.1201/b16589>
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5), 589–595. <https://doi.org/10.1093/bioinformatics/btp698>
- Li, Y., Shao, T., Jiang, C., Bai, J., Wang, Z., Zhang, J., ... Li, X. (2015). Construction and analysis of dynamic transcription factor regulatory networks in the progression of glioma. *Scientific Reports*, 5(1), 15953. <https://doi.org/10.1038/srep15953>
- Liang, K., & Keleş, S. (2012). Normalization of ChIP-seq data with control. *BMC Bioinformatics*, 13(1). <https://doi.org/10.1186/1471-2105-13-199>
- Liau, B. B., Sievers, C., Donohue, L. K., Gillespie, S. M., Flavahan, W. A., Miller, T. E., ... Bernstein, B. E. (2017). Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance. *Cell Stem Cell*, 20(2), 233–246.e7. <https://doi.org/10.1016/j.stem.2016.11.003>
- Mahabir, R., Tanino, M., Elmansuri, A., Wang, L., Kimura, T., Itoh, T., ... Tanaka, S. (2014). Sustained elevation of Snail promotes glial-mesenchymal transition after irradiation in malignant glioma. *Neuro-Oncology*, 16(5), 671–685. <https://doi.org/10.1093/neuonc/not239>
- Min, S., Lee, B., & Yoon, S. (2016). Deep learning in bioinformatics. *Briefings in Bioinformatics*, 18(5), bbw068. <https://doi.org/10.1093/bib/bbw068>

- Misra, B. B., Langefeld, C., Olivier, M., & Cox, L. A. (2018). Integrated omics: tools, advances and future approaches. *Journal of Molecular Endocrinology*, (2016), R21–R45. <https://doi.org/10.1530/jme-18-0055>
- Molina, J. R., Hayashi, Y., Stephens, C., & Georgescu, M.-M. (2010). Invasive Glioblastoma Cells Acquire Stemness and Increased Akt Activation. *Neoplasia*, 12(6), 453-IN5. <https://doi.org/10.1593/NEO.10126>
- Montano, N., Cenci, T., Martini, M., D'Alessandris, Q. G., Pelacchi, F., Ricci-Vitiani, L., ... Pallini, R. (2011). Expression of EGFRvIII in Glioblastoma: Prognostic Significance Revisited. *Neoplasia*, 13(12), 1113-IN6. <https://doi.org/10.1593/NEO.111338>
- Morokoff, A., Ng, W., Gogos, A., & Kaye, A. H. (2015). Molecular subtypes, stem cells and heterogeneity: Implications for personalised therapy in glioma. *Journal of Clinical Neuroscience*, 22(8), 1219–1226. <https://doi.org/10.1016/j.jocn.2015.02.008>
- Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., & Stamatoyannopoulos, J. A. (2012). Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. *Cell*, 150(6), 1274–1286. <https://doi.org/10.1016/j.cell.2012.04.040>
- Nibbe, R. K., Koyutürk, M., & Chance, M. R. (2010). An Integrative -omics Approach to Identify Functional Sub-Networks in Human Colorectal Cancer. *PLoS Computational Biology*, 6(1), e1000639. <https://doi.org/10.1371/journal.pcbi.1000639>
- Parkhomenko, E., Tritchler, D., & Beyene, J. (2009). Sparse Canonical Correlation Analysis with Application to Genomic Data Integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 1–34. <https://doi.org/10.2202/1544-6115.1406>
- Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Shawn, M., Wakimoto, H., ... Suvà, M. L. (2014). NIH Public Access. *Science*, 344(6190), 1396–1401. <https://doi.org/10.1126/science.1254257>. Single-cell
- Pavel, A. B., Sonkin, D., & Reddy, A. (2016). Integrative modeling of multi-omics data to identify cancer drivers and infer patient-specific gene activity. *BMC Systems Biology*, 10(1), 16. <https://doi.org/10.1186/s12918-016-0260-9>
- Phillips, H. S., Kharbanda, S., Chen, R., Forrest, W. F., Soriano, R. H., Wu, T. D., ... Aldape, K. (n.d.). Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. <https://doi.org/10.1016/j.ccr.2006.02.019>

- Rheinbay, E., Suvà, M. L., Gillespie, S. M., Wakimoto, H., Patel, A. P., Shahid, M., ... Bernstein, B. E. (2013). An Aberrant Transcription Factor Network Essential for Wnt Signaling and Stem Cell Maintenance in Glioblastoma. *Cell Reports*, 3(5), 1567–1579. <https://doi.org/10.1016/j.celrep.2013.04.021>
- Saito, N., Fu, J., Zheng, S., Yao, J., Wang, S., Liu, D. D., ... Koul, D. (2014). A High Notch Pathway Activation Predicts Response to γ Secretase Inhibitors in Proneural Subtype of Glioma Tumor-Initiating Cells. *STEM CELLS*, 32(1), 301–312. <https://doi.org/10.1002/stem.1528>
- Sandberg, C. J., Altschuler, G., Jeong, J., Strømme, K. K., Stangeland, B., Murrell, W., ... Langmoen, I. A. (2013). Comparison of glioma stem cells to neural stem cells from the adult human brain identifies dysregulated Wnt- signaling and a fingerprint associated with clinical outcome. *Experimental Cell Research*, 319(14), 2230–2243. <https://doi.org/10.1016/J.YEXCR.2013.06.004>
- Schulte, A., Gunther, H. S., Martens, T., Zapf, S., Riethdorf, S., Wulfing, C., ... Lamszus, K. (2012). Glioblastoma Stem-like Cell Lines with Either Maintenance or Loss of High-Level EGFR Amplification, Generated via Modulation of Ligand Concentration. *Clinical Cancer Research*, 18(7), 1901–1913. <https://doi.org/10.1158/1078-0432.CCR-11-3084>
- Sosa, M. S., Bragado, P., & Aguirre-Ghiso, J. A. (2014). Mechanisms of disseminated cancer cell dormancy: An awakening field. *Nature Reviews Cancer*, 14(9), 611–622. <https://doi.org/10.1038/nrc3793>
- Steinhauser, S., Kurzawa, N., Eils, R., & Herrmann, C. (2016). A comprehensive comparison of tools for differential ChIP-seq analysis. *Briefings in Bioinformatics*, 17(November 2015), bbv110. <https://doi.org/10.1093/bib/bbv110>
- Stergachis, A. B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A. P., Zhang, M., ... Stamatoyannopoulos, J. A. (2014). Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, 515(7527), 365–370. <https://doi.org/10.1038/nature13972>
- Suravajhala, P., Kogelman, L. J. A., & Kadarmideen, H. N. (2016). Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare. *Genetics Selection Evolution*, 48(1), 38. <https://doi.org/10.1186/s12711-016-0217-x>
- Talpaz, M., Shah, N. P., Kantarjian, H., Donato, N., Nicoll, J., Paquette, R., ... Sawyers, C. L. (2006). Dasatinib in Imatinib-Resistant Philadelphia Chromosome-Positive Leukemias. *New England Journal of Medicine*, 354(24), 2531–2541. <https://doi.org/10.1056/NEJMoa055229>
- Tanaka, S., Louis, D. N., Curry, W. T., Batchelor, T. T., & Dietrich, J. (2013). Diagnostic

- and therapeutic avenues for glioblastoma: No longer a dead end? *Nature Reviews Clinical Oncology*, 10(1), 14–26. <https://doi.org/10.1038/nrclinonc.2012.204>
- Toedt, G., Barbus, S., Wolter, M., Felsberg, J., Tews, B., Blond, F., ... Radlwimmer, B. (2011). Molecular signatures classify astrocytic gliomas by *IDH1* mutation status. *International Journal of Cancer*, 128(5), 1095–1103. <https://doi.org/10.1002/ijc.25448>
- Tuncbag, N., Gosline, S. J. C., Kedaigle, A., Soltis, A. R., Gitter, A., & Fraenkel, E. (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLOS Computational Biology*, 12(4), e1004879. <https://doi.org/10.1371/journal.pcbi.1004879>
- Tuncbag, N., McCallum, S., Huang, S.-S. C., & Fraenkel, E. (2012). SteinerNet: a web server for integrating ‘omic’ data to discover hidden components of response pathways. *Nucleic Acids Research*, 40(Web Server issue), W505-9. <https://doi.org/10.1093/nar/gks445>
- Tuncbag, N., Milani, P., Pokorny, J. L., Johnson, H., Sio, T. T., Dalin, S., ... Fraenkel, E. (2016). Network Modeling Identifies Patient-specific Pathways in Glioblastoma. *Scientific Reports*, 6(1), 28668. <https://doi.org/10.1038/srep28668>
- Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., ... Cancer Genome Atlas Research Network, T. (n.d.). Cancer Cell Article Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17, 98–110. <https://doi.org/10.1016/j.ccr.2009.12.020>
- Wang, J., Duncan, D., Shi, Z., & Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Research*, 41(Web Server issue), 77–83. <https://doi.org/10.1093/nar/gkt439>
- Wang, J., Vasaiakar, S., Shi, Z., Greer, M., & Zhang, B. (2017). WebGestalt 2017: A more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Research*, 45(W1), W130–W137. <https://doi.org/10.1093/nar/gkx356>
- Wortman, I., Rossetti, N., Riggi, N., Rivera, M. N., Suvà, M. L., Wakimoto, H., ... Rozenblatt-Rosen, O. (2014). Reconstructing and Reprogramming the Tumor-Propagating Potential of Glioblastoma Stem-like Cells. *Cell*, 157(3), 580–594. <https://doi.org/10.1016/j.cell.2014.02.030>
- Xu, D. (2008). *Applications of fuzzy logic in bioinformatics*. Imperial College Press.
- Zhang, B., Kirov, S., & Snoddy, J. (2005). WebGestalt: An integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*,

33(SUPPL. 2), 741–748. <https://doi.org/10.1093/nar/gki475>

Zhao, Z., Meng, F., Wang, W., Wang, Z., Zhang, C., & Jiang, T. (2017). Comprehensive RNA-seq transcriptomic profiling in the malignant progression of gliomas. *Scientific Data*, 4, 1–7. <https://doi.org/10.1038/sdata.2017.24>



APPENDICES

APPENDIX A

STATISTICS OF FOREST NETWORKS

Table A. The table shows the number of terminals, number of terminal nodes in the final forest network and total number of node and edge in the final forest network (# of terminals, # of terminal in the optimal forest network, # of node/edge in the optimal forest network, respectively) at the end of each Forest run with different parameters. Columns represent for each condition specific network.

			naïve vs 12d			12d vs persister			naïve vs persister		
w	beta	mu	0.50	1	0.00	0.50	1	0.00	0.50	1	0.00
# of terminals			145			134			83		
# of terminals in optimal forest network			145			134			83		
# node/edge in optimal forest network			228 / 227			202 / 201			139 / 137		
w	beta	mu	0.50	1	0.01	0.50	1	0.01	0.50	1	0.01
# of terminals			145			134			83		
# of terminals in optimal forest network			145			131			83		

# node/edge in optimal forest network			186 /183			168 /165			108 / 105		
w	beta	mu	0.50	2	0.00	0.50	2	0.00	0.50	2	0.00
# of terminals			145			134			83		
# of terminals in optimal forest network			145			134			83		
# node/edge in optimal forest network			227 /226			206 /205			139 / 137		
w	beta	mu	0.50	2	0.01	0.50	2	0.01	0.50	2	0.01
# of terminals			145			134			83		
# of terminals in optimal forest network			0			134			83		
# node/edge in optimal forest network			0 / 0			169 /166			108 / 105		
w	beta	mu	0.50	3	0.00	0.50	3	0.00	0.50	3	0.00
# of terminals			145			134			83		
# of terminals in optimal forest network			145			134			83		
# node/edge in optimal forest network			228 / 227			203 / 202			139 / 137		
w	beta	mu	0.50	3	0.01	0.50	3	0.01	0.50	3	0.01
# of terminals			145			134			83		

# of terminals in optimal forest network			0			134			83		
# node/edge in optimal forest network			0 / 0			169 / 166			108 / 105		
w	beta	mu	0.50	4	0.00	0.50	4	0.00	0.50	4	0.00
# of terminals			145			134			83		
# of terminals in optimal forest network			145			134			83		
# node/edge in optimal forest network			226 / 225			202 / 201			139 / 137		
w	beta	mu	0.50	4	0.01	0.50	4	0.01	0.50	4	0.01
# of terminals			145			134			83		
# of terminals in optimal forest network			0			134			83		
# node/edge in optimal forest network			0 / 0			169 / 166			108 / 105		
w	beta	mu	0.50	5	0.00	0.50	5	0.00	0.50	5	0.00
# of terminals			145			134			83		
# of terminals in optimal forest network			145			134			83		
# node/edge in optimal forest network			226 / 225			202 / 201			140 / 138		
w	beta	mu	0.50	5	0.01	0.50	5	0.01	0.50	5	0.01

# of terminals			145			134			83		
# of terminals in optimal forest network			0			134			83		
# node/edge in optimal forest network			0 / 0			169 / 166			108 / 105		
w	beta	mu	1.00	1	0.00	0.50	1	0.00	0.50	1	0.00
# of terminals			145			134			83		
# of terminals in optimal forest network			145			134			83		
# node/edge in optimal forest network			227 / 226			200 / 199			137 / 136		
w	beta	mu	1.00	1	0.01	1.00	1	0.01	1.00	1	0.01
# of terminals			145			134			83		
# of terminals in optimal forest network			145			132			83		
# node/edge in optimal forest network			186 / 185			167 / 166			108 / 106		
w	beta	mu	1.00	2	0.00	1.00	2	0.00	1.00	2	0.00
# of terminals			145			134			83		
# of terminals in optimal forest network			145			134			83		
# node/edge in optimal forest network			227 / 226			202 / 201			137 / 136		

w	beta	mu	1.00	2	0.01	1.00	2	0.01	1.00	2	0.01
# of terminals			145			134			83		
# of terminals in optimal forest network			145			0			83		
# node/edge in optimal forest network			186 / 185			0 / 0			108 / 105		
w	beta	mu	1.00	3	0.00	1.00	3	0.00	1.00	3	0.00
# of terminals			145			134			83		
# of terminals in optimal forest network			145			0			83		
# node/edge in optimal forest network			186 / 185			0 / 0			108 / 106		
w	beta	mu	1.00	3	0.01	1.00	3	0.01	1.00	3	0.01
# of terminals			145			134			83		
# of terminals in optimal forest network			145			0			83		
# node/edge in optimal forest network			186 / 185			0 / 0			108 / 106		
w	beta	mu	1.00	4	0.00	1.00	4	0.00	1.00	4	0.00
# of terminals			145			134			83		
# of terminals in optimal forest network			145			134			83		

# node/edge in optimal forest network			217 / 216			201 / 200			138 / 137		
w	beta	mu	1.00	4	0.01	1.00	4	0.01	1.00	4	0.01
# of terminals			145			134			83		
# of terminals in optimal forest network			145			0			83		
# node/edge in optimal forest network			186 / 185			0 / 0			108 / 106		
w	beta	mu	1.00	5	0.00	1.00	5	0.00	1.00	5	0.00
# of terminals			145			134			83		
# of terminals in optimal forest network			145			134			83		
# node/edge in optimal forest network			216 / 215			202 / 201			134 / 133		
w	beta	mu	1.00	5	0.01	1.00	5	0.01	1.00	5	0.01
# of terminals			145			134			83		
# of terminals in optimal forest network			145			0			83		
# node/edge in optimal forest network			186 / 185			0 / 0			108 / 106		