UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

THAÍS SARRAF SABEDOT

**Molecular classification of adult diffuse gliomas based on DNA methylation
reveals subgroups of G-CIMP tumors associated with distinct clinical
features**

Ribeirão Preto

2018

THAÍS SARRAF SABEDOT

**Molecular classification of adult diffuse gliomas based on DNA methylation
reveals subgroups of G-CIMP tumors associated with distinct clinical
features**

Original version

Doctoral thesis submitted to the Ribeirão Preto Medical School – FMRP-USP, in partial fulfillment of the requirements to obtain a doctoral degree (PhD) in Science.

Area: Genetics

Advisor: Prof. Dr. Houtan Noushmehr

Ribeirão Preto

2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

I authorize the reproduction and total or partial dissemination of this study, by electronic or conventional forms, to research and teaching purposes, provided appropriately citation.

Ficha catalográfica elaborada pela Biblioteca Central da USP Ribeirão Preto
com os dados fornecidos pelo(a) autor(a)

S115m	<p>Sabedot, Thais Sarraf</p> <p>Molecular classification of adult diffuse gliomas based on DNA methylation reveals subgroups of G-CIMP tumors associated with distinct clinical features. Ribeirão Preto, 2018.</p> <p>121 p. : il. ; 30 cm</p> <p>Doctoral thesis (Doctorate Candidate - Program in Genetics) - Ribeirão Preto Medical School (FMRP/USP), USP, 2018.</p> <p>Area: Genetics</p> <p>Advisor: Noushmehr, Houtan</p> <p>1. DNA methylation. 2. Glioma. 3. G-CIMP 4. Epigenetics 5. Bioinformatics</p>
-------	--

Tese de autoria de Thaís Sarraf Sabedot, sob o título “**Molecular classification of adult diffuse gliomas based on DNA methylation reveals subgroups of G-CIMP tumors associated with distinct clinical features**”, apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Doutora em Ciências pelo Programa de Pós-graduação em Genética, aprovada em _____ de _____ de _____ pela comissão julgadora constituída pelos doutores:

Prof. Dr. _____

Instituição: _____

Presidente

Prof. Dr. _____

Instituição: _____

Prof. Dr. _____

Instituição: _____

Prof. Dr. _____

Instituição: _____

Prof. Dr. _____

Instituição: _____

*Dedicated to the memory of my biggest fan in the whole world, my beloved grandmother
Irene*

Acknowledgements

I would like to thank my parents and my family for loving me unconditionally and for making sure all my dreams come true.

I would also like to thank my best friends Amanda, Isabel, Natália, Tathi e Maritza. You have been by my side through all this journey and I can not thank you enough for your constant support.

I am grateful for all my lab mates, more importantly my friends, from OMICs laboratory whom have taught me so much over the years that it is hard to condense in one paragraph.

I would like to thank all employees and professors from the Department of Genetics at USP and from the Department of Neurosurgery at HFH, specially Nancy Takacs and Conni Kick.

Also, I would like to thank Ana Valeria Castro, Laila Poisson, Ana de Carvalho, Jim Snyder, George Divine for all the stimulating discussions and insightful comments.

In special, I am grateful for having Houtan Noushmehr as my mentor. There are no words to thank you enough for everything you did to me.

Finally, I would like to thank São Paulo Research Foundation (FAPESP) for the financial support and predoctoral fellowship (Processes # 2016/06488-3 and # 2016/12329-5).

Resumo

SABEDOT, TS. **Classificação molecular de gliomas difusos em adulto baseada em metilação do DNA revela subgrupos de tumores G-CIMP associados com aspectos clínicos distintos**. 2018. 121 f. Tese (Doutorado em Ciências) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, 2018.

Gliomas são tumores heterogêneos, o que contribui para seu alto grau de mortalidade, apesar de avanços na classificação e tratamento. Desde 2016, a incorporação do estado dos genes *IDH* e da integridade dos cromossomos 1p e 19q na classificação de gliomas fornece aplicações clínicas importantes para o diagnóstico e tratamento deste tumor; entretanto, a procura por assinaturas moleculares que possam refinar ainda mais os subtipos de glioma em subgrupos mais homogêneos é um esforço contínuo. Este estudo utilizou o maior número de amostras de gliomas adultos (n=932) até a atualidade, variando dos graus II ao IV, a fim de definir subgrupos de glioma utilizando assinaturas de metilação do DNA, independentemente de grau e histologia. No total, 7 subtipos foram identificados: Classic-like, Mesenchymal-like, LGm6-GBM, PA-like, Codels, G-CIMP-low and G-CIMP-high. A maior parte dos subgrupos com *IDH* tipo selvagem, isto é, Classic-like, Mesenchymal-like, LGm6-GBM, possuem padrão de baixa metilação do DNA e um pior risco prognóstico; características clínicas típicas de glioblastomas, o tipo mais agressivo de gliomas. Uma descoberta interessante foi a identificação do subgrupo PA-like dentre gliomas com *IDH* tipo selvagem, o qual compartilha aspectos genômicos similares a astrocitoma pilocítico, um glioma pediátrico benigno com bom quadro clínico entre gliomas com *IDH* tipo selvagem. Codels, os quais abragem pacientes com mutação em *IDH* e codeleção dos cromossomos 1p e 19, possuem o melhor prognóstico dentre os gliomas difusos em adultos. Uma descoberta importante em relação a gliomas com mutação em *IDH*, porém sem codeleção dos cromossomos 1p e 19q, foi a estratificação de gliomas com fenótipo metilador de ilhas CpG (G-CIMP) em G-CIMP-low, com níveis mais baixos de metilação do DNA e pior quadro clínico, e G-CIMP-high, com níveis mais altos de metilação do DNA e melhor risco prognóstico. Curiosamente, o grau de metilação do DNA (-low e -high) estava associado com alterações distintas em elementos regulatórios e modificações de histona aberrantes na região promotora de genes do ciclo celular. Estes achados consolidaram a importância clínica da epigenética, particularmente da metilação do DNA, em gliomas, como também levantou a possibilidade de que a sobrevida média ruim de G-CIMP-low pode ser associada a elementos regulatórios. Além disso, a hipótese de que *enhancers* ativos podem agir na regulação gênica de G-CIMP-low fornece mais evidências de que elementos regulatórios podem levar à maior agressividade e proliferação de G-CIMP-low. Este estudo visa 1) identificar e caracterizar subtipos de gliomas difusos em adultos baseados na metilação do DNA, e 2) avaliar a associação entre modificações de histona com um subtipo mais agressivo de G-CIMP.

Palavras-chaves: Metilação do DNA. Glioma. G-CIMP. Epigenética. Bioinformática.

Abstract

SABEDOT, TS. **Molecular classification of adult diffuse gliomas based on DNA methylation reveals subgroups of G-CIMP tumors associated with distinct clinical features.** 2018. 121 p. Thesis (PhD of Science) – Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, SP, 2018.

Gliomas are heterogeneous tumors which contribute to their high mortality despite advancements in classification and treatment. As of 2016, the incorporation of *IDH* status and the integrity of chromosomes 1p and 19q to glioma classification have provided important clinical application for diagnostics and treatment; however, the search for molecular signatures that further refine glioma subtypes into more homogeneous subgroups is an ongoing effort. This study used the largest sample cohort (n=932) of adult gliomas to date, ranging from grades II to IV, in order to define gliomas subgroups using DNA methylation signatures, independent of histopathological grading. In total, 7 subtypes were identified: Classic-like, Mesenchymal-like, LGm6-GBM, PA-like, Codels, G-CIMP-low and G-CIMP-high. Most *IDH*-wildtype subgroups, e.g. Classic-like, Mesenchymal-like and LGm6-GBM, had low DNA methylation pattern and a poor outcome, typical of glioblastomas, the most aggressive phenotype of gliomas. An interesting finding was the identification of the PA-like subgroup within *IDH*-wildtype samples, which shared similar genomic features with pilocytic astrocytoma, a rare pediatric benign glioma, with a good overall survival (OS) among *IDH*-wildtype gliomas. Codels, which comprise *IDH* mutant gliomas with codeletion of chromosomes 1p/19q have the best OS across all adult gliomas. An important finding regarding *IDH* mutant gliomas with no codeletion of chromosomes 1p/19q, was the further segregation of the Glioma-CpG Island Methylator Phenotype (G-CIMP) into G-CIMP-low, with lower levels of DNA methylation and worse OS, and G-CIMP-high, characterized by higher DNA methylation profile and better OS. Interestingly, the degree of G-CIMP methylation (-low and -high) was associated with distinct alterations in regulatory elements and aberrant histone modifications at promoter regions of cell cycle genes. These findings consolidated the clinical importance of epigenetics, particularly DNA methylation, in gliomas, as well as the possibility that aggressive OS in G-CIMP-low may be driven by regulatory elements. Moreover, our results suggest that active enhancers that might be acting in gene regulation in G-CIMP-low provide more evidence of the regulatory elements that might be driving aggressiveness and proliferation in G-CIMP-low. This study aims 1) to identify and characterize adult diffuse glioma DNA methylation subtypes, and 2) evaluate the association of histone modifications with a more aggressive G-CIMP subtype.

Keywords: DNA methylation. Glioma. G-CIMP. Epigenetics. Bioinformatics.

List of Figures

Figure 1 – Schematic flow chart of the study design from chapter 2.	30
Figure 2 – Venn diagram with overlapping CpGs between both Illumina platforms.	32
Figure 3 – Schematic description of the methodology to identify epigenetically regulated genes.	36
Figure 4 – Glioma DNA methylation subtypes.	41
Figure 5 – <i>IDH</i> -specific DNA methylation subtypes.	42
Figure 6 – G-CIMP characterization.	44
Figure 7 – <i>IDH</i> -wildtype glioma characterization.	47
Figure 8 – Epigenetically regulated genes that define glioma subtypes.	49
Figure 9 – Summary of glioma subtypes	53
Figure 10 – Comparison between G-CIMP-low and G-CIMP-high	54
Figure 11 – Effect of epigenetics in CIMP tumors.	58
Figure 12 – Schematic flow chart of the study design from chapter 3.	62
Figure 13 – Distribution of TCGA glioma samples across tissue source sites	63
Figure 14 – Quality control report.	65
Figure 15 – Read counts by step	66
Figure 16 – Boxplot of the number of peaks for each sample by ChIP-seq experiment.	70
Figure 17 – Histone modifications in the region surrounding TSS.	71
Figure 18 – Genomic view of <i>GAPDH</i> on chromosome 12	72
Figure 19 – Epigenomic changes associated with G-CIMP low.	75
Figure 20 – Gene expression changes associated with G-CIMP low.	76
Figure 21 – Pathway analysis of upregulated genes in G-CIMP-low. The x-axis represents the $-1*\log_{10}$ transformed p-value.	76
Figure 22 – Chromatin changes associated with transcription.	77
Figure 23 – <i>SIM2</i> activation by gain of H3K27ac at TSS.	78
Figure 24 – Distribution of H3K27ac marks outside TSS.	80
Figure 25 – Regulatory elements predicted by H3K27ac ChIP-seq and GeneHancer.	81
Figure 26 – <i>HOXA</i> cluster might be activated by nearby enhancers in G-CIMP-low	82
Figure 27 – Precision medicine.	89
Figure 28 – Glioma transcriptome subtypes	105

List of Tables

Table 1 – Common histone modifications associated with gene regulation	24
Table 2 – Non-tumor non-brain samples from TCGA used for the identification of epigenetically regulated genes in gliomas	35
Table 3 – Overview of clinical features arranged by established biomarkers	39
Table 4 – Number of samples selected using CISTROME	68
Table 5 – Clinical features of G-CIMP-low and G-CIMP-high profiled by ChIP-seq	69
Table 6 – Differential binding analysis of H3K27ac and H3K4me3 at TSS by gene	73
Table 7 – Epigenetically regulated genes	100

List of abbreviations and acronyms

AS	Astrocytoma
2HG	2-hydroxyglutarate
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
CBTRUS	Central brain tumor registry in the United States
ChIP-seq	Chromatin immunoprecipitation sequencing
CI	Confidence interval
CNS	Central neural system
CTCF	CCCTC-binding factor
DNA	Deoxyribonucleic acid
DNMT1	DNA Methyltransferase 1
EReg	Epigenetically regulated genes
FA	Female
GEO	NCBI Gene Expression Omnibus
HFH	Henry Ford Hospital
HM27	Illumina Infinium Human Methylation 27K
HM450	Illumina Infinium Human Methylation 450K
ICGC	International Cancer Genome Consortium
IDH	Isocitrate dehydrogenase
G-CIMP	Glioma-CpG island methylator phenotype
GBM	Glioblastoma
HOX	Homeobox

H3K4me3	Trimethylation of histone H3 at lysine 4
H3K27ac	Acetylation of histone H3 at lysine 27
LGG	Lower-grade glioma (grades II or III)
LQ	Lower-quartile
MA	Male
NA	Not available
OA	Oligoastrocytoma
OD	Oligodendroglioma
OS	Overall survival
PCA	Principal Component Analysis
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphisms
TAD	Topologically associated domains
TCGA	The Cancer Genome Atlas
TET	Ten-eleven translocation
TF	Transcription factor
TSS	Transcription start site
UQ	Upper-quartile
WGBS	Whole-genome bisulfite sequencing
WHO	World health organization

List of symbols

β	Beta
bp	base pairs
kb	kilobase
mg	milligrams
mut	mutant
seq	sequencing
wt	wildtype

Contents

1	Thesis outline	17
I	Introduction	18
2	Glioma overview	19
3	Epigenetic modifications and cancer	21
3.1	<i>DNA methylation</i>	21
3.2	<i>Chromatin remodeling</i>	23
II	Identification and characterization of adult diffuse glioma DNA methylation subtypes	25
4	Background	26
4.1	<i>Molecular subtypes of glioma</i>	26
5	Hypothesis	28
6	Objectives	29
6.1	<i>General objective</i>	29
6.2	<i>Specific objectives</i>	29
7	Materials and methods	30
7.1	<i>Study design</i>	30
7.2	<i>Sample selection and data acquisition</i>	30
7.3	<i>Preprocessing</i>	32
7.4	<i>Unsupervised analysis</i>	33
7.5	<i>Supervised analysis</i>	34
7.6	<i>Identification of epigenetically regulated genes</i>	34
7.7	<i>Validation and classification of new glioma samples</i>	36
7.8	<i>Sequence motif discovery</i>	37
8	Results	39

8.1	<i>Clinical features of patient cohort</i>	39
8.2	<i>Unsupervised analysis of gliomas reveals six DNA methylation groups</i>	40
8.3	<i>IDH-specific clustering shows overall concordance with glioma DNA methylation groups</i>	41
8.4	<i>Identification of G-CIMP subclassification elucidates differences in clinical outcome of IDH mutant gliomas</i>	43
8.5	<i>A subset of IDH-wildtype samples with normal-like features resembles pediatric glioma</i>	45
8.6	<i>Discovery of epigenetically regulated genes that can predict glioma molecular classification based on DNA methylated profile</i>	48
8.7	<i>Tool development to analyze TCGA DNA methylation data</i>	50
9	Publication	51
10	Discussion	52
11	Conclusion	55
III	Chromatin remodeling associated with G-CIMP tumors	56
12	Background	57
12.1	<i>Impact of epigenetics on gliomas</i>	57
12.2	<i>Novel G-CIMP subclassification</i>	57
13	Hypothesis	60
14	Objectives	61
14.1	<i>General objective</i>	61
14.2	<i>Specific objectives</i>	61
15	Material and methods	62
15.1	<i>Study design</i>	62
15.2	<i>Sample selection and data generation</i>	62
15.3	<i>Quality control and alignment</i>	64
15.4	<i>Preprocessing</i>	65

15.5	<i>Peak calling</i>	66
15.6	<i>Detection of differential binding in ChIP-seq data</i>	67
15.7	<i>Detection of differentially expressed genes</i>	67
15.8	<i>External datasets</i>	67
15.9	<i>Prediction of potential enhancer</i>	68
15.10	<i>Visualization</i>	68
16	Results	69
16.1	<i>Clinical features of patient cohort</i>	69
16.2	<i>Initial assessment</i>	69
16.3	<i>Enriched levels of histone modifications at transcription start sites</i> . .	72
16.4	<i>Regulatory elements associated with G-CIMP-low</i>	78
17	Discussion	83
18	Conclusion	85
IV	Synthesis and perspective	86
19	Synthesis	87
20	Perspective	88
21	Contributions to science	90
21.1	<i>Publications</i>	90
21.2	<i>Presentations</i>	91
21.2.1	Poster	91
21.2.2	Oral presentation	92
21.3	<i>Academic and professional honors</i>	92
21.4	<i>Professional memberships</i>	92
21.5	<i>Peer review</i>	92
	Bibliography	93
	APPENDIX A – Epigenetically regulated genes	100

ANNEX A – Exempt determination by the Ethics Committee of Ribeirão Preto Medical School	103
ANNEX B – Glioma gene expression subtypes	105
ANNEX C – Publication	106

1 Thesis outline

This thesis is divided into four chapters: first chapter introduces basic concepts of gliomas and epigenetics; the second describes the identification and characterization of adult diffuse glioma DNA methylation subtypes; the third aims to evaluate the association of histone modifications with a more aggressive G-CIMP subtype, and the fourth presents perspectives and synthesis of this study along with contributions to science.

Part I

Introduction

2 Glioma overview

Adult diffuse gliomas are central nervous system (CNS) tumors which arise from the malignant transformation of glial cells. Gliomas are relatively rare, with fewer than 200,000 US cases per year; however, high grade gliomas may present with a very aggressive behavior and a median survival of only about 15 months (STUPP et al., 2005; MALTA et al., 2017). Current treatment consists of surgical tumor resection followed by radiotherapy in addition to chemotherapy. Treating gliomas remains a challenge, partly due to their heterogeneity among the histological subtypes, and despite the approval of multiple new therapies in the past decade, the high mortality rate persists.

For many years, glioma classification was based on the type of cell of origin, i.e. astrocytes, oligodendrocytes and ependymal cells, and on tumor grading. Grade I gliomas are typically pediatric benign tumors which can be cured surgically; on the other hand, grade IV are very aggressive malignant tumors that may grow rapidly and are often fatal. A subset of lower grade gliomas (LGG, grades II-III) will progress to higher grade gliomas (grade IV, GBM); however, current glioma grading fails to predict which of them will relapse and/or progress. This can be partly attributed to gliomas' molecular heterogeneity which cannot be detected by histopathological grading (AHMED et al., 2014).

Over the last decade, large-scale profiling studies, such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), have accelerated the comprehensive understanding of the role of genomics and epigenomics in carcinogenesis, including diffuse gliomas. These efforts led to an update of the World Health Organization (WHO) glioma classification, which currently incorporates molecular parameters to the traditional histological classification (LOUIS et al., 2016). As a result of the integration of pheno- and genotypic features, gliomas are classified into two major categories: *IDH*-wildtype and *IDH*-mutant gliomas. *IDH*-wildtype gliomas are most frequently associated with primary (*de novo*) glioblastomas and account for 90% of grade IV cases, which predominate in older patients (median age at diagnosis of 62 years). On the other hand, *IDH*-mutant GBMs are usually defined as secondary glioblastoma; represent 10% of GBM cases; preferentially arise from precursor lower grade diffuse astrocytoma or lower grade anaplastic astrocytoma, compromise younger patients (median age at diagnosis of 44 years) and are located in the frontal brain (YAN et al., 2009; LOUIS et al., 2016).

Another molecular feature that further segregated *IDH*-mutant gliomas into discrete groups is the integrity status of chromosomes arms 1p and 19q. *IDH*-mutant glioma subgroups are classified into 1p/19q codeleted (*IDH*-mutant codelets) and euploid groups (*IDH*-mutant 1p/19q intact or non-codelets). Along with the traditional histology/grade-based glioma classification, *IDH* status, chromosome 1 and 19 integrity has also been incorporated to the WHO glioma classification.

Despite the advances in glioma classification, the search for molecular signatures that further narrow down glioma subtypes into more homogeneous molecular subgroups is still an ongoing effort. In this line, the impact of DNA methylation profile in the characterization of glioma subtypes was investigated. The authors found that *IDH*-mutant glioma manifested a genome-wide hypermethylation of CpG islands, characterizing the Glioma-CpG Island Methylator Phenotype (G-CIMP) (TURCAN et al., 2012). This subgroup had a better prognosis than *IDH*-mutant gliomas that did not present CIMP. It is known that *IDH* mutation extensively remodels tumor epigenome and possibly changes chromosome topology and genomic regulatory units (NOUSHMEHR et al., 2010).

Besides the prognostic value that DNA methylation showed in gliomas; previous studies also reported that DNA methylation also had a predictive value in gliomas. They reported that patients whose tumors presented methylation of the *MGMT* promoter had a better response to treatment compared to patients with an unmethylated promoter (HEGI et al., 2005). The results from these studies yielded valuable insights into tumor biology and progression that may ultimately be translated into more precise diagnosis as well as effective cancer treatment and preventive strategies.

3 Epigenetic modifications and cancer

Throughout the course of malignant transformation, cancer cells typically acquire several chromosomal abnormalities, nucleotide substitutions and epigenetic modifications that collectively play an important role in cancer biology and tumor progression. Epigenetic modifications are defined as heritable changes in gene regulation that do not affect the DNA sequence (ECCLESTON et al., 2007) and are involved in normal development and in pathological conditions. These events are responsible to determine cell phenotype and are related to environmental, chemical, nutritional and pharmaceutical agents (CORTESSIS et al., 2012). Understanding how these factors act in gene expression regulation is essential to correlate the adaptation within genetic code to different cell commitment. Two major epigenetic mechanisms known to play an important role during development from a zygote to an adult, which are characterized as phases of epigenetic reprogramming and differentiation (HEMBERGER; DEAN; REIK, 2009), are DNA methylation and remodeling of the chromatin structure. Methylation of DNA is involved in regulating gene expression as well as modifying the chromatin architecture and its interactions with proteins, which highlights the interplay between these two mechanisms reported, i.e.

3.1 DNA methylation

DNA methylation was the first recognized, and probably the most extensively studied, epigenetic modification in mammals. DNA methylation involves the covalent transfer of a methyl group to the C-5 position of the cytosine ring of DNA, which normally precedes a guanine (e.g. CpG dinucleotides in genomic DNA), creating 5mC. When this event occurs at gene promoter regions with a high density of CpGs (e.g. CpG island), commonly the gene is silenced, such as it occurs in the X chromosome inactivation (BIRD, 1986). However, aberrant DNA methylation patterns in promoter regions of genes responsible for fundamental mechanisms, such as apoptosis and cell cycle, may lead to the development of complex diseases such as cancer, Alzheimer's disease and diabetes (SCHINKE et al., 2010). 5mC may be hydroxylated by the *TET1/2* dioxygenases to generate 5hmC which, in turn, can be recognized by DNMT1, leading to replication-dependent passive demethylation. This complex mechanism is potentially required for

orchestrating the balance between DNA methylation and demethylation, ensuring the transcriptional control at specific genomic regions (WU; ZHANG, 2011; ESTELLER, 2008).

It is known that global changes in the epigenetic landscape are a hallmark of cancer. Moreover, the interaction between DNA methylation and regulatory elements, such as enhancers or silencers, has important functional consequences in gene regulation and this could be mediated by histone modifications, encoded by complex genomic signatures on looping chromatin (WHALEN; TRUTY; POLLARD, 2016). Therefore, characterizing the epigenetic interplay between histone modifications and DNA methylation could help to understand cell mechanisms related to tumor progression and gliomagenesis, that could be associated with disruption of enhancer elements and insulators. Important genes related to DNA methylation changes in gliomas were the *IDH* genes. *IDH1* and its mitochondrial homologue *IDH2* encode for isocitrate dehydrogenases 1 and 2, respectively, which generate reduced nicotinamide adenine dinucleotide phosphate (NADPH) from NADP+ by catalyzing the interconversion of isocitrate and α -ketoglutarate outside of the Krebs cycle. The *IDH1/2* mutation, first described in gliomas in 2008 (PARSONS et al., 2008), occurs at a single amino acid residue, is usually somatic, heterozygous and mutually exclusive. The mutated forms of *IDH* genes lead to the production of a potential oncometabolite called 2-hydroxyglutarate (2-HG) (DANG et al., 2009), associated with accumulation of DNA methylation at promoter regions due to abnormal interactions of 2-HG with the TET (Ten-eleven translocation methylcytosine dioxygenase) family of DNA hydroxylases, inducing gene silence modulated by epigenetics in some cases of gliomas (NOUSHMEHR et al., 2010). TET family enzymes have the biochemical ability to demethylate the DNA by converting 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC) (TAHILIANI et al., 2009).

Due to the importance of this epigenetic mechanism and available technologies that facilitate the interrogation of DNA methylation across the epigenome, DNA methylation has been widely studied in several cancer types, including glioblastomas (THE CANCER GENOME ATLAS RESEARCH NETWORK, 2008; BRENNAN et al., 2013). Recent studies (BERMAN et al., 2012; MEDVEDEVA et al., 2014) have shown that gain of DNA methylation can disrupt transcription factors binding sites motifs at promoter regions of coding genes in cancer, which may lead to epigenetic silencing by changing the chromatin structure.

3.2 Chromatin remodeling

Chromatin's main function is to package DNA molecules within the cell in order to prevent DNA damage and to control replication and gene expression. The fundamental unit of chromatin is the nucleosome, composed of an octamer of the four core histones (H3, H4, H2A, H2B) around which 147 base pairs of DNA are wrapped (RICHMOND; DAVEY, 2003). Epigenetic chemical posttranslational modifications that take place on histones, in particular methylation and acetylation, can alter the interaction of histones with DNA and nuclear proteins (KOUZARIDES, 2007). Histone modifications change the accessibility of chromatin either by recruiting and/or occluding non-histone effector proteins, resulting in activation or repression, respectively, depending upon which residues are modified and the type of modifications that occur (SHARMA; KELLY; JONES, 2010). Two of the most common studied histone modifications at promoter regions are the trimethylation of histone H3 at lysine 4 (H3K4me3) and the acetylation of histone H3 at lysine 27 (H3K27ac) (GUENTHER et al., 2007; KARLIĆ et al., 2010). Supposedly, both modifications at histone 3 (trimethylation at lysine 4 and acetylation at lysine 27) help to create an open chromatin structure which then recruits effectors that mediate a transcriptional state. However, disease states, such as cancer, could induce aberrant modification of histones at promoter regions, leading to uncontrolled activation of oncogenes (BANNISTER; KOUZARIDES, 2011). Another important function of H3K27ac is the identification of active regulatory elements, called enhancers, which are short (50-1,500bp) DNA sequences that may be bound by transcription factors to activate the likelihood of a particular gene to be transcribed (CREYGHTON et al., 2010). Identifying active enhancers in cancer could provide insights into tumorigenesis and new targets. Several enhancers were mapped in glioblastomas which suggest evidence of gene regulation with phenotypical implications (LIANG et al., 2002; ORZAN et al., 2011). Table 1 summarizes the most studied roles of H3K4me3 and H3K27ac associated with gene activation.

Histone modifications and DNA methylation interact with each other in different ways to determine gene expression status, chromatin organization and cellular identity (CEDAR; BERGMAN, 2009). Studies highlighted that DNA methylation and histone modifications are known to affect nuclear organization (GILBERT et al., 2007). Therefore,

the interplay of different epigenetic mechanisms is important to maintain the function of the genome.

Table 1 – Common histone modifications associated with gene regulation

Functional definition	H3K4me3	H3K27ac
Active TSS		
Active enhancer		
Bivalent/poised TSS		
Weak/Quiescent		

The organized chromatin structure, megabase-sized interaction domains, can be defined by loops called topologically associated domains (TADs). The domains are shown to be stable across cell types and conserved across species (DIXON et al., 2012). Moreover, the boundaries of TADs are enriched for insulator binding proteins CTCF, that controls the contact of distant enhancer and promoter elements and, consequently, regulate gene expression. For instance, a recent study showed that *IDH* mutation in gliomas was associated with hypermethylation at CTCF binding sites, which then led to *PDGFRA* activation by an enhancer interaction with this oncogene through chromatin reorganization (FLAVAHAN et al., 2016).

New technologies have arisen and are constantly evolving to interrogate molecular data. In particular, two techniques are relatively new and informative in accessing epigenetic information across the entire genome, including intergenic and unexplored regions: chromatin immunoprecipitation sequencing (ChIP-seq), for protein interactions with DNA including histone modifications and transcription factors, and whole-genome bisulfite sequencing (WGBS) for DNA methylation at single CpG sites (PLONGTHONGKUM; DIEP; ZHANG, 2014).

Part II

Identification and characterization of adult diffuse glioma DNA methylation subtypes

4 Background

4.1 Molecular subtypes of glioma

GBM was the first major cancer type studied in large cohorts by TCGA. The pilot study integrated genomic, epigenomic and clinical data to better understand the molecular basis of GBM associated with response to treatment (THE CANCER GENOME ATLAS RESEARCH NETWORK, 2008). Two years later, Verhaak and colleagues were able to classify GBM into 4 molecular subtypes defined by transcriptomic data: proneural, neural, classical and mesenchymal subtypes (VERHAAK et al., 2010) which was later refined into three (proneural, classical and mesenchymal) (WANG et al., 2017). Concurrently, Nounshmehr and collaborators identified the presence of a subset of GBM with hypermethylation of CpG islands enriched for mutation of the *IDH1* gene (NOUSHMEHR et al., 2010). This subgroup was named G-CIMP and it frequently occurred as a subgroup of proneural subtype of gliomas, in younger patients with favorable outcome when compared to G-CIMP negative tumors. The finding was validated *in vitro* in another cohort which also showed that *IDH1* mutation could establish the hypermethylator phenotype by remodeling the epigenome of the cell (TURCAN et al., 2012).

In another study, Sturm and colleagues integrated both pediatric and adult GBM in order to identify distinct GBM subgroups based on genome-wide DNA methylation profiles. Thereby, six subgroups of samples were described: IDH, K27, G34, RTK I (PDGFRA), RTK II (Classic) and Mesenchymal (STURM et al., 2012). Three of those were enriched for specific mutations: K27 and G34 presented mutation in *H3F3A*, whereas IDH is associated with *IDH1* mutated samples and global hypermethylation consistent with the previous G-CIMP group. The remaining subgroups correlated with previously described molecular subtypes of adult GBM (BRENNAN et al., 2013).

Likewise, TCGA provided relevant insights for the adult lower-grade glioma (LGG; gliomas grades II-III) field, describing mutations in *IDH1/2*, *TP53*, and *ATRX* and codeletion of chromosome arms 1p and 19q (1p/19q codeletion) as clinically relevant markers. LGG patients with an *IDH1/2* mutation and 1p/19q codeletion had the most favorable clinical outcomes. The great majority of LGG not carrying an *IDH* mutation mark, a distinguishable *IDH*-wildtype subgroup, had GBM-like clinical behavior with poor survival (THE CANCER GENOME ATLAS RESEARCH NETWORK, 2015).

Usually, LGG and GBM are characterized as two independent entities. However, mixing different glioma grades by molecular profiling might lead to more homogeneous classification associated with clinical features.

5 Hypothesis

DNA methylation signatures may define discrete adult diffuse gliomas subgroups independently of histopathological grading.

6 Objectives

6.1 *General objective*

To perform an integrated analysis of publicly available DNA methylation and genomic data from adult primary diffuse gliomas (LGG and GBM).

6.2 *Specific objectives*

The specific objectives of this study are:

- to acquire DNA methylation and genomic data publicly available from adult primary diffuse gliomas, independent of grade and histology
- to classify glioma subgroups according to DNA methylation profile and correlate the findings with LGG and GBM subgroups (THE CANCER GENOME ATLAS RESEARCH NETWORK, 2015; BRENNAN et al., 2013)
- to integrate epigenomic and genomic data in order to identify potential genes regulated by DNA methylation
- to identify prognostic biomarkers for each glioma subtype
- to validate the DNA methylation glioma subtypes and biomarkers in an independent glioma dataset
- to develop a package for integrative analysis of epigenomic and genomic TCGA data

7 Materials and methods

The data analysis was performed in the OMICs laboratory located at the Department of Genetics - FMRP, Ribeirão Preto, SP, Brazil and also at the Department of Neurosurgery Research of the Henry Ford Hospital, Detroit, MI, USA. The project received an exempt determination by the Ethics Committee of Ribeirão Preto Medical School A.

7.1 Study design

The methodology used in order to identify potential subgroups of gliomas based on molecular profiling included the following steps: sample selection and data acquisition, unsupervised analysis, identification of epigenetically regulated genes, validation and classification of new glioma samples and, finally, sequence motif discovery. A flow chart is represented in figure 1.

Figure 1 – Schematic flow chart of the study design from chapter 2.

7.2 Sample selection and data acquisition

A database query from the TCGA data portal retrieved 932 adult primary glioma samples, which excluded recurrent, non-tumor brain and control samples, with DNA methylation data available. For download, a R script was developed to access the data

portal and transfer the files to a local machine along with the corresponding clinical data. In total, there were 516 LGG and 416 GBM available samples.

About half of the GBM samples ($n=287$) was profiled using an older DNA methylation array platform, called Illumina HumanMethylation 27 platform (HM27), available during the early years of TCGA, which interrogates 27,578 CpG probes, mostly located at CpG islands and gene promoter regions. Due to becoming an outdated platform, Illumina updated the DNA methylation array with a more comprehensive genome-wide coverage platform, Infinium Human Methylation 450K (HM450), which interrogates methylation at 485,421 sites and captures non-CpG islands, intergenic regions and gene body. Thereby, the remaining GBM samples ($n=129$) and all LGG samples ($n=516$) were profiled using HM450. In order to ensure the largest number of samples, data of both platforms (HM27 and HM450) were merged, which resulted in 25,978 CpG probes and 932 glioma samples to be analyzed. Figure 2 represents a Venn diagram with CpG sites from each platform and common probes of the sets are represented by the areas of overlap among the circles.

Figure 2 – Venn diagram with overlapping CpGs between both Illumina platforms.

Sizes of the circles are proportional to the number of CpGs interrogated by each platform. A total of 459,599 CpG sites are exclusive to HM450, whereas 25,978 probes are shared between HM450 and HM27. Only 1,600 sites are unique to HM27.

Additional matching gene expression data, generated by RNA-sequencing, were retrieved for 636 TCGA glioma samples. Besides, DNA methylation data (HM27) from non-tumor brain tissue (n=77) was also obtained from a previously published dataset (GUINTIVANO; ARYEE; KAMINSKY, 2013) as control, under NCBI Gene Expression Omnibus (GEO) accession GSE41826.

7.3 Preprocessing

The DNA methylation data acquired from TCGA and GEO were previously preprocessed and normalized. Each CpG probe received a score called β -value, ranging

from 0 to 1. The closer to zero, the more copies of the CpG site in the sample were unmethylated, however, a value of one suggest that every copy of the locus was methylated. This score was calculated by the following formula: $\beta = (M/(M+U))$, where M was the methylated intensity signal and U was the unmethylated intensity for each probe. CpG probes overlapping known single nucleotide polymorphisms (SNP), non-unique mapping, repeats or not being statistically significantly different from background (p-value > 0.01) were labeled as "not available" (NA).

The gene expression data, aligned to human reference genome by TCGA, was normalized using within-lane normalization to adjust for GC-content effect or gene length, followed by upper-quantile to minimize effects related to distributional differences between lanes, according to the EDASeq protocol (RISSO et al., 2011). The RNA-seq data was then log2 transformed.

7.4 *Unsupervised analysis*

In order to group samples with similar DNA methylation profile, an unsupervised analysis was performed in an attempt to find molecular subtypes of gliomas. Filtering methods described below were applied to the data with the aim of reducing the number of probes to glioma-specific CpG sites. This step is fundamental due to the computational challenges of applying clustering methods in large datasets (SANDER, 2000). Methods to select tumor-specific regions were used as described (THE CANCER GENOME ATLAS RESEARCH NETWORK, 2014b) with slight alterations.

Initially, there were 25,978 CpG probes, which were filtered down to those within a 1,500 bp window surrounding known transcription start sites (TSS) and overlapping CpG islands defined by UCSC Genome Browser (n=9,600). Probes labeled as NA (n=1,815) and probes that were designed for sequences on X and Y chromosomes were also discarded (n=2). To correct for possible tumor purity bias across samples, the data was dichotomized according to the following criteria: CpG probes with β -values greater than 0.3 were classified as "methylated", otherwise, they were labeled as "unmethylated". To reduce sample heterogeneity, the remaining probes were filtered to the ones that were methylated in, at least, 10% of the tumor samples (n=2,146). Lastly, CpG sites which were methylated

in the non-tumor brain cohort (GUINTIVANO; ARYEE; KAMINSKY, 2013) were omitted from this analysis.

Ultimately, 1,300 CpG probes were defined as glioma-specific sites. Subsequently, an unsupervised hierarchical clustering was performed on the dichotomized data using Ward’s method for linkage and binary distance metric for clustering. The dendrogram was cut to determine the cluster assignments. For further visualization and downstream analysis, the original β -values were used. The probes were ordered in accordance with hierarchical clustering of the β -values using the complete agglomeration method and euclidean distance measurement.

Apart from combining all diffuse glioma samples together, a separate analysis segregating samples by *IDH* mutation status was also performed using the same approach as employed above. A total of 1,308 CpG probes specific for *IDH* mutant glioma samples (n=450) was identified, along with 914 CpG sites that define *IDH*-wildtype glioma samples (n=430). In addition, a Principal Component Analysis (PCA) was performed on genome-wide CpGs of all glioma samples, along with 77 non-tumor brain samples, towards verifying the correlation between cluster assignments.

7.5 *Supervised analysis*

In an attempt to identify biological and/or clinical features that may distinguish each molecular subtype, groups were compared using statistical tests such as Wilcoxon test, for nonparametric analysis, and Student’s t-test, for parametric analysis, followed by false discovery rate estimation using the Benjamini and Hochberg (BH) method (BENJAMINI; HOCHBERG, 1995). Data with $FDR \leq 0.05$ were called statistical significant.

7.6 *Identification of epigenetically regulated genes*

To estimate the effect of epigenetic differences on gene expression for each molecular subtype, a method to combine DNA methylation array with RNA-seq was developed. Each CpG probe was mapped to the nearest gene, resulting in 19,530 pairs with an unique ID. Since only glioma samples with both DNA methylation array and RNA-sequencing could be used in the analysis, there were 636 samples (123 GBM and 513 LGG) available

for this step. Due to the lack of a non-tumor brain dataset with both DNA methylation and RNA-sequencing profiled by the same platforms in TCGA, tumor adjacent normal non-brain tissue was used as control in this analysis: 110 samples from 11 different tissues. Table 2 shows all the selected tissues.

Table 2 – Non-tumor non-brain samples from TCGA used for the identification of epigenetically regulated genes in gliomas

Tissue
Bladder
Breast
Colon
Head and neck
Kidney renal clear cell
Kidney renal papillary cell
Liver
Lung
Prostate
Thyroid
Uterus

After combining the two datasets of both gliomas and non-tumor samples, ordered by CpG-gene ID, each sample was classified either as "methylated" (β -value > 0.3) or "unmethylated" (β -value ≤ 0.3) for each pair. Then, the mean gene expression was calculated for the methylated and unmethylated groups, separately. The CpG-gene pairs were filtered to the ones in which the mean expression of the methylated samples was lower than 1.28 standard deviations (bottom 10%) of the mean expression in the unmethylated group. Therefore, the CpG-gene pairs in which more than 80% of the glioma samples in the methylated group had expression levels lower than the mean expression in the unmethylated samples were called as being epigenetically regulated. This method was used as described by the TCGA analysis working group in 2014 (THE CANCER GENOME ATLAS RESEARCH NETWORK, 2014a) and showed in figure 3.

Epigenetic Silencing Calling Method

1. Organize samples as either methylated ($\beta \geq 0.3$) or unmethylated ($\beta < 0.3$) for each probe.

Probe selection

2. Select probes for which the mean expression in the methylated group is lower than 1.28 standard deviations (bottom 10%) of the mean expression in the unmethylated group
3. Select probes from step #2 in which >80% of the samples in the methylated group have expression levels lower than the mean expression in the unmethylated group

Silencing call for each sample

4. Label individual sample as epigenetically silenced for a specific probe selected from step 3 if:
 - a. it belongs to methylated group and
 - b. expression of the gene is lower than mean of unmethylated group of samples

At the gene level, a sample has to be called silenced at more than half of the probes

Figure 3 – Schematic description of the methodology to identify epigenetically regulated genes.

On the left, main steps are highlighted as a text. On the right, a figure represents the threshold to classify samples as epigenetically silenced by each gene. Source: Hui Shen and Toshi Hinoue.

To identify sets of CpG-gene pairs enriched in each glioma molecular subtype, a Fisher-test was used to verify statistical significance comparing the DNA methylation clusters with non-tumor samples. The data were counted and organized into a contingency table using 50% as threshold. A p-value detection was calculated for each comparison and then adjusted for false discovery rate estimation. Each CpG-gene set was then called as Epigenetically Regulated gene (EReg).

7.7 Validation and classification of new glioma samples

The molecular signatures identified in this project were tested on heterogeneous sets of gliomas, publicly available for download in the GEO database, in order to assess

the consistency and ability to replicate the same findings in an independent dataset at both molecular and clinical level. Four large sets of glioma samples with DNA methylation array data have been published in the literature: Sturm et al. (2012) (GSE36278) consists of adult and pediatric glioblastomas, Turcan et al. (2012) (GSE30339), which comprise gliomas with *IDH1* mutation, Lambert et al. (2013) (GSE44684), with grade I gliomas and, finally, Mur et al. (2013) (GSE61160), with gliomas carrying codeletion of chromosome arms 1p and 19q. In total, 324 glioma samples were used for validation.

The method used to classify the validation set into glioma molecular subtypes identified in this project, which can also be applied to any other glioma sample with DNA methylation array data, requires a random-forest machine-learning model to be performed using the set of CpG probes that defines the target molecular subtype. For *IDH*-wildtype glioma samples, the non-TCGA cohort was classified using 914 CpGs that define *IDH*wt gliomas. The model was trained initially using a random set of 80% of TCGA samples and tested on the remaining 20%. With an accuracy of more than 85%, the model was applied to the external dataset to predict cluster assignments. For *IDH* mutant glioma samples, a two-step random forest machine-learning method was required to predict each assigning for the non-TCGA dataset: first, the model was trained on 80% of TCGA *IDH* mutant glioma samples using 1,308 CpG probes that define *IDH* mutant specific groups. In this step, samples predicted to be *IDH*mut-K3 were labeled as Codel. TCGA samples identified as *IDH*mut-K1 and *IDH*mut-K2 were then trained on 90 CpGs to discriminate between G-CIMP-low and G-CIMP-high. Both models were finally tested on the remaining 20% samples and got an accuracy of more than 85%.

7.8 Sequence motif discovery

With the purpose of capturing candidate DNA motif sequences, which might be potential transcription factor binding sites that regulate each glioma subtype, a method to search for sequence motif by analyzing genomic positions was applied. The tool, called HOMER (version 4.4) (HEINZ et al., 2010), performs a differential motif discovery analysis by comparing genomic coordinates inputted by the user with background sequences, which could be provided by the user as well or the tool will select random sequences from the reference genome matched with GC% content. Then, HOMER will auto normalize

the target sequences and the background by reducing bias caused by lower-order oligo sequences. HOMER looks for enrichment of known motifs from its database, which is composed by published data from functional experiments. Lastly, HOMER will also search for *de novo* motifs by selecting the most enriched oligos, compared with background, and calculating the motif enrichment using cumulative binomial distribution. Differential CpG probes between glioma molecular subtypes were classified according to genomic location: UCSC defined CpG islands, CpG shores flanking CpG island ± 2 kb and open seas, CpGs isolated in the genome. Then, each category was inputted to HOMER with up to two mismatches allowed.

8 Results

8.1 Clinical features of patient cohort

The dataset of gliomas from TCGA comprises 216 grade II, 241 grade III and 416 grade IV samples. Due to *IDH* status and the integrity of chromosome arms 1p and 19q being the most relevant clinically discoveries related to molecular features for the glioma field, table 3 shows the overall characteristics of this cohort organized by establish biomarkers.

Table 3 – Overview of clinical features arranged by established biomarkers

	<i>IDH</i> -wildtype	<i>IDH</i> mutant	<i>IDH</i> mutant 1p/19q code1
Histology			
Glioblastoma	317	27	1
Astrocytoma	52	112	4
Oligoastrocytoma	15	69	30
Oligodendroglioma	19	37	117
Unknown	15	33	18
Grade			
Grade II	19	114	19
Grade III	67	104	70
Grade IV	317	27	1
Unknown	15	33	18
Age			
Median (LQ-UQ)	59 (51-68)	38 (30-44)	46 (35-54)
Unknown	15	33	18
Survival (months)			
Median (CI)	14 (12-16)	80 (63-100)	95 (78-Inf)
Unknown	14	32	18
<i>MGMT</i> promoter			
Methylated	170	242	169
Unmethylated	248	36	1

As reported by others (THE CANCER GENOME ATLAS RESEARCH NETWORK, 2015; ECKEL-PASSOW et al., 2015), the majority of glioblastomas are *IDH*-wildtype, oligodendrogliomas are mostly *IDH* mutant 1p/19q codeleted and *IDH* mutants are frequently associated with lower-grade gliomas. Likewise, *IDH*-wildtype glioma patients are usually older at diagnosis than *IDH* mutants, have the worst prognosis and commonly

the *MGMT* promoter is unmethylated. However, it's possible to observe that these 3 major groups are still heterogeneous and further stratification could improve glioma classification.

8.2 *Unsupervised analysis of gliomas reveals six DNA methylation groups*

An unsupervised analysis performed using hierarchical clustering across 932 glioma samples on 1,300 glioma-specific CpGs showed 6 distinct DNA methylation subtypes, labeled LGm1-6, showed on figure 4A. Likewise, supplemental figure B presents 4 glioma transcriptome subtypes, labeled LGr1-4. *IDH* status divided the samples into 2 major groups: *IDH* mutant, which comprises LGm1, LGm2 and LGm3 carrying DNA hypermethylation compared to both non-tumor brain and other glioma samples; and *IDH*-wildtype: LGm4, LGm5, LGm6. Additionally, LGm3 is also enriched for samples harboring codeletion of 1p and 19q chromosome arms.

In concordance with LGm clusters, a PCA of 932 gliomas and 77 non-tumor brain across all CpG probes showed a robust molecular classification using DNA methylation data, as seen in figure 4B. Interestingly, one of the *IDH* mutant subgroups, LGm1, seems to present an intermediate phenotype between *IDH* mutant and *IDH*-wildtype, and this feature reflected in a survival difference, showed by figure 4C. On the other hand, LGm6 samples clustered closer to non-tumor brain samples than other glioma subtypes.

Figure 4 – Glioma DNA methylation subtypes.

(A) Heatmap with DNA methylation data across 932 gliomas on 1,300 CpG probes. Columns indicate samples and rows, CpG sites. Non-tumor brain samples are represented on the left. (B) PCA of 932 gliomas and 77 non-tumor brain samples across genome-wide CpGs (HM450+HM27 probe set). (C) Kaplan-Meier survival curves showing TCGA glioma subtypes. Ticks represent censored values.

8.3 *IDH-specific clustering shows overall concordance with glioma DNA methylation groups*

An unsupervised analysis was performed separating samples by *IDH* status. By using the same methodology, there were 1,308 CpG probes able to divide 450 *IDH* mutant samples into 3 subgroups, called *IDH*-mutant K1-3. 1,162 CpGs (89%) overlapped with

the 1,300 glioma-specific CpGs. On the other hand, 914 CpG probes specific to 430 *IDH*-wildtype samples were able to distinguish between 3 subgroups, called *IDH*-wildtype K1-3. Also, 853 CpG probes (66%) overlapped with the 1,300 glioma-specific CpGs. Interestingly, *IDH*-specific subgroups showed profound agreement with LGm clusters, as seen on figure 5.

A

B

Figure 5 – *IDH*-specific DNA methylation subtypes.

(A) Heatmap with DNA methylation data across 450 *IDH*-mutant gliomas on 1,308 CpG probes. Columns indicate samples and rows, CpG sites. (B) Heatmap with DNA methylation data across 430 *IDH*-wildtype gliomas on 914 CpG probes.

8.4 *Identification of G-CIMP subclassification elucidates differences in clinical outcome of IDH mutant gliomas*

Based on table 3, the expected median overall survival for *IDH* mutant glioma samples without 1p and 19q codeletion is 80 months (CI 63-100). However, the *IDH* mutant K1 group has a median overall survival of only 38.7 months (CI 32-Inf). Along with that, it was noticed that *IDH* mutant K1 is also globally hypomethylated compared to the other *IDH* mutant non-codel samples, as seen in figure 6A.

Figure 6 – G-CIMP characterization.

(A) Boxplot of mean DNA methylation scores (genome-wide) by *IDH* mutant non-coding glioma subtypes. Significant difference is highlighted with *** (p -value $< 2.2 \times 10^{-16}$). (B) Heatmap of 131 differentially methylated probes between *IDH*mut-K1 and *IDH*mut-K2. (C) Kaplan-Meier survival curves of refined *IDH* mutant classification. (D) Genomic alterations across *IDH* mutant subgroups. Red represents amplification and blue, deletion. (E) Distribution of differentially methylated probes across CpG location followed by motif analysis. (F) DNA methylation levels between longitudinally matched G-CIMP samples.

A supervised analysis between *IDH* mutant K1 and K2 resulted in 131 differentially methylated probes and revealed a subgroup of 25 samples within *IDH* mutant K1 that were hypomethylated (figure 6B). This event could not be captured during the unsupervised analysis because only 19 out of 131 probes overlap with the *IDH* mutant-specific CpGs ($n=1,308$). Along with hypomethylation, these 25 patients also have a significant worse

prognosis (figure 6C), compared to other *IDH* mutant glioma samples. These 25 samples were then called G-CIMP-low, due to a lower DNA methylation profile; the remaining *IDH* mutant K1 samples, along with *IDH* mutant K2, were labeled as G-CIMP-high (n=249), due to higher DNA methylation levels; finally, *IDH* mutant K3 was called Codel (n=174), because it consists, mostly, of *IDH* mutant-codel LGGs. By investigating genomic features of these 3 *IDH* mutant subtypes, cell cycle alterations were identified in 15 out of 18 G-CIMP-low samples (figure 6D). These alterations are commonly found in glioblastomas, which are the most aggressive glioma phenotype. A supervised analysis between G-CIMP-low (*IDH* mutant K1 with low DNA methylation) and the remaining *IDH* mutant K1 samples, using only HM450 data due to the possibility of capturing more differences, identified 633 significantly hypomethylated CpGs in G-CIMP-low. Dividing this set of probes into CpG location and performing a sequence motif analysis allowed the discovery of an association between hypomethylation in intergenic regions (open sea CpGs) and the TGTT motif signature (figure 6E), known to be related with *SOX* and *OLIG2* transcription factor (TF) families (LODATO et al., 2013). Both these transcription factors were previously described as part of a core set of neurodevelopmental factors fundamental to GBM propagation (SUVÀ et al., 2014), which, again, associates G-CIMP-low with a more aggressive phenotype.

Intrigued by the possible relation between G-CIMP-low and G-CIMP-high, a set of longitudinally matched G-CIMP samples from TCGA (n=23 samples from 10 patients) was downloaded and, by analyzing the DNA methylation pattern at differentially methylated probes between G-CIMP-low and G-CIMP-high (n=90 CpGs), it was observed that 4 out of 10 patients showed a demethylation pattern upon glioma relapse, providing evidence for disease progression from G-CIMP-high to G-CIMP-low (figure 6F).

8.5 A subset of *IDH*-wildtype samples with normal-like features resembles pediatric glioma

The *IDH*-wildtype unsupervised clustering revealed 3 subgroups: *IDH*wt-K1, with an enrichment for samples previously classified as classical, was named as Classic-like, *IDH*wt-K2, with the majority of samples belonging to the mesenchymal class (VERHAAK et al., 2010), was called Mesenchymal-like and, finally, *IDH*wt-K3, with mixed subtypes and lower DNA methylation levels comparable with non-tumor brain samples (figure 7A).

However, a supervised analysis within *IDH*wt-K3 by tumor grade was not able to identify differentially methylated CpGs.

Additional glioma samples from already published datasets were downloaded and classified into the *IDH*wt K1-3 subtypes. 221 *IDH*-wildtype gliomas with DNA methylation array data were identified: 114 mixed adult and pediatric glioblastomas (STURM et al., 2012), 32 *IDH*wt LGGs (TURCAN et al., 2012), 61 grade I pilocytic astrocytomas (LAMBERT et al., 2013) and 14 oligodendroglial tumors (MUR et al., 2013) (figure 7B). As expected, RTK II "Classic" and Mesenchymal samples (STURM et al., 2012) were, predominantly, predicted as Classic-like (*IDH*wt-K1) and Mesenchymal-like (*IDH*wt-K2), respectively. CIMP- samples from both studies (TURCAN et al., 2012; MUR et al., 2013) were mostly distributed across Classic-like and Mesenchymal-like. On the other hand, all pilocytic astrocytoma samples were classified as *IDH*wt-K3 (figure 7B). Besides, by dividing the TCGA *IDH*wt-K3 subtype by tumor grade (LGG and GBM), there was a major overall survival difference: *IDH*wt-K3 LGGs showed a *IDH* mutant-like survival curve whereas *IDH*wt-K3 GBMs had a typical *IDH*-wildtype survival curve with a poor clinical outcome (figure 7C). These findings led to the separation of *IDH*wt-K3 into 2 subgroups: PA-like, which are the LGGs classified as *IDH*wt-K3, and LGm6-GBM, which comprise GBMs classified as *IDH*wt-K3.

Finally, in order to further characterize this newly identified group of *IDH*-wildtype gliomas with better prognosis, known genomic alterations related to pilocytic astrocytomas were investigated, such as *BRAF*, *NF1*, *NTRK1*, *NTRK2*, *FGFR1*, and *FGFR2* (JONES et al., 2013). Interestingly, 52% of PA-like samples had alterations in at least one of these genes, but, overall, no other described GBM or *IDH*-wildtype glioma common altered pathways were mutated (figure 7D) .

Figure 7 – *IDH*-wildtype glioma characterization.

(A) Heatmap of 914 CpGs that define *IDH*-wildtype glioma subgroups based on DNA methylation data. On the left, non-tumor brain samples, in the middle 430 *IDH*wt TCGA samples and on the right, 221 external *IDH*wt glioma samples. (B) Distribution of previous published DNA methylation subtypes across the TCGA *IDH*-wildtype subgroups predicted by a random forest model. (C) Kaplan-Meier survival curves of refined TCGA *IDH*-wildtype groups. (D) Genomic alterations across *IDH*-wildtype subgroups. Red represents amplification, blue, deletion, purple, in-frame indel and pink, truncating mutation.

As one last attempt to explain such unexpected clinical and molecular features of PA-like samples, Dr. Daniel Brat, a neuropathologist who has been working with TCGA for many years, collaborated with this study and agreed to review the histology of these 26 samples. By his report, 23 of the 26 cases were indeed grade II or III diffuse gliomas. The remaining 3 cases were reclassified as grade I pilocytic astrocytoma.

8.6 *Discovery of epigenetically regulated genes that can predict glioma molecular classification based on DNA methylated profile*

Since alterations in the epigenome could affect gene expression, RNA-sequencing data was integrated with DNA methylation array in search for gene signatures to define each glioma molecular subtype. For the *IDH* mutant cohort, there were 17 G-CIMP-low, 234 G-CIMP-high and 173 Codel samples available with both DNA methylation and gene expression data. Additionally, 3 ERegs were identified: EReg1 with a set of 15 genes that define G-CIMP-low by being hypomethylated and upregulated, EReg2 with 15 genes that define *IDH* mutant gliomas and, finally, EReg3, also with 15 genes epigenetically silenced, which define Codels.

The random forest prediction model was applied to the validation set of 103 *IDH* mutant samples: 22 mixed adult and pediatric glioblastomas (STURM et al., 2012), 49 *IDH* mutant LGGs (TURCAN et al., 2012) and 32 oligodendroglial tumors (MUR et al., 2013). Interestingly, 9 G-CIMP-low were identified and 6 of them belong to the *IDH* subgroup from Sturm et al. (2012). 22 G-CIMP-high were captured and enriched for CIMP+ groups from Turcan et al. (2012) and Mur et al. (2013). As expected, 72 predicted Codels were mostly enriched for CD-CIMP+, which are *IDH* mutant glioma samples carrying a codeletion of 1p and 19q chromosome arms, from Mur et al. (2013). The 45 EReg signatures that define *IDH* mutant were recapitulated using the validation set, as seen in figure 8A and B, and can also be found on table 7.

Figure 8 – Epigenetically regulated genes that define glioma subtypes.

(A) Heatmap of 45 ERegs that define G-CIMP-low, G-CIMP-high and Codel subtypes. On the left, 110 non-tumor samples were plotted using DNA methylation data. In the middle, DNA methylation heatmap of 932 TCGA samples. On the right, the mean RNA-sequencing data for matching gene was plotted by group. (B) Heatmap of DNA methylation data from validation set, using the same CpG probes that define EReg1-2-3. (C) Heatmap of 27 ERegs that define Classic-like and LGM6 (PA-like + LGM6-GBM) subtypes. On the left, 110 non-tumor samples were plotted using DNA methylation data. In the middle, DNA methylation heatmap of 932 TCGA samples. On the right, the mean RNA-sequencing data for matching gene was plotted by group. (D) Heatmap of DNA methylation data from validation set, using the same CpG probes that define EReg4-5.

For the *IDH*-wildtype cohort, there were 69 Classic-like, 98 Mesenchymal-like, 12 LGM6-GBM and 25 PA-like samples available with both DNA methylation and gene expression data. Only 2 ERegs were identified: EReg4, with a set of 15 hypermethylated and downregulated genes that define Classic-like, and EReg5, with 12 hypomethylated

CpGs that define the LGm6 (PA-like + LGm6-GBM) group. For EReg5, there were no epigenetically regulated genes that could distinguish LGm6. So a different approach was performed: DNA methylation levels of LGm6 samples (n=77) were compared with 140 randomly selected samples from the 855 remaining TCGA glioma samples (n=140) and 12 hypomethylated CpGs were identified ($\text{FDR} < 1 \times 10^{-21}$) and mapped to the nearest gene.

The same approach was used to validate EReg4 and EReg5 signatures that define Classic-like and LGm6 (PA-like + LGm6-GBM), respectively, using 221 non-TCGA *IDH*-wildtype gliomas, which were classified into Classic-like, Mesenchymal-like, LGm6-GBM and PA-like, as previously described. As expected, the DNA methylation profile was successfully reproduced in the validation cohort, as seen in figure 8C and D. There were no significant methylation differences which could serve as biomarkers for the Mesenchymal-like subtype neither to distinguish PA-like and LGm6-GBM.

8.7 Tool development to analyze TCGA DNA methylation data

The R code generated by this methodology to analyze TCGA data was organized into functions and compiled as part of a R/Bioconductor package called TCGAbiolinks (COLAPRICO et al., 2015). The package aims to search, download and prepare relevant TCGA molecular data for analysis, not only using glioma samples, but a large variety of tumor types. Also, it's possible to reproduce the pipeline described in this study in order to regenerate figures and validate the results. In particular, codes to perform unsupervised analysis (function `TCGAanalyze_Clustering`), supervised analysis (`TCGAanalyze_DMR`), to plot heatmaps (`TCGAvisualize_Heatmap`), to create volcano plots for DNA methylation data (`TCGAvisualize_volcano`), to integrate DNA methylation with gene expression data (`TCGAvisualize_starburst`) and, finally, a function to calculate and plot the mean methylation boxplot for each subgroup (`TCGAvisualize_meanMethylation`) were incorporated to TCGAbiolinks.

9 Publication

Sections 8.1-6 of chapter 2 were part of a manuscript published by Cell on January, 2016 by Michele Ceccarelli, Floris P. Barthel, Tathiane M. Malta, **Thais S. Sabedot** et al., under the title "Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma", which is attached to the end of this document as Annex C.

Additionally, section 8.7 of chapter 2 was incorporated to a manuscript published by Nucleic Acids Research on December, 2015 by Antonio Colaprico, Tiago C. Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, **Thais S. Sabedot**, Tathiane M. Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi and Houtan Noushmehr, under the title "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data".

10 Discussion

The sample cohort used in this study with 932 glioma samples profiled by DNA methylation array was the largest multi-platform glioma epigenomic analysis performed to date. Figure 9 provides an overview of the molecular subtypes based on DNA methylation data associated with diffuse gliomas. Interestingly, histology and classic tumor grading were not enough to predict cluster assignments. However, classification based on DNA methylation resulted in subgroups with reduced heterogeneity. A comprehensive and integrative analysis of diffuse gliomas in order to better refine the molecular profiling is essential as we lean towards genome-based clinical stratification of patients.

The segregation of G-CIMP tumors into G-CIMP-low and G-CIMP-high brought new insights regarding the importance of epigenetics in gliomas. A summary of the most important findings that define G-CIMP-low and G-CIMP-high are shown in figure 10. Lower DNA methylation presented by G-CIMP-low is associated with a more aggressive phenotype. Understanding what is causing demethylation in G-CIMP-low could eventually lead to new drug targets, which may prevent tumor progression. Another interesting aspect is the hypothesis that G-CIMP-high might emerge as G-CIMP-low glioma at recurrence. By investigating DNA methylation levels between primary G-CIMP-high and matched recurrence, it was noted a pattern of demethylation after relapse in a subset of samples, which resembles the G-CIMP-low phenotype, suggesting then a progression from G-CIMP-high to G-CIMP-low. A recent study showed that the level of DNA methylation in a set of CpG probes might predictive the risk of primary G-CIMP-high to progress to G-CIMP-low upon disease recurrence (SOUZA et al., 2017). However, further investigation on progression within G-CIMP tumors needs to be addressed in order to provide evidence that loss of DNA methylation in G-CIMP-high lead to G-CIMP-low phenotype.

Likewise, the discovery of a subset of *IDH*-wildtype glioma samples with favorable outcome (PA-like) was an important finding that could eventually lead to different treatment strategies for these patients. The molecular profile of PA-like shared similarities to pilocytic astrocytoma, which are usually benign tumors with extremely rare progression to higher grades (HAYOSTEK et al., 1993). These findings suggest that a different mechanism might drive gliomagenesis of PA-like tumors.

Figure 9 – Summary of glioma subtypes

Figure 10 – Comparison between G-CIMP-low and G-CIMP-high

Source: (MALTA et al., 2017)

In summary, the recent incorporation of molecular features, such as *IDH* status and chromosome 1 and 19 integrity, in WHO glioma classification (LOUIS et al., 2016) shows the importance of biomarker discovery to improve disease understanding. Refined molecular profiling methods allows deeper investigation of diseases, which eventually may lead the path to precision medicine.

11 Conclusion

By combining multiple molecular data, this study was able to classify diffuse glioma into more homogeneous subgroups depicting distinct clinical outcome more accurately than previous classifications. Moreover, this study also provided new prognostic biomarkers based on DNA methylation data that could potentially be used clinically to guide therapeutic schemes for each patient. The publication shows the importance of this study to the scientific community.

Part III

Chromatin remodeling associated with G-CIMP tumors

12 Background

12.1 *Impact of epigenetics on gliomas*

In order to solve the influence of *IDH* mutation behind the epigenetic of gliomas, Turcan et al. (2012) showed that *IDH* mutation was capable of remodeling the methylome of immortalized primary human astrocytes cells. Using a self-organizing map analysis, they demonstrated progressive changes in the methylome of mutant *IDH1* astrocytes compared to *IDH1*-wildtype, both in genes that underwent *de novo* methylation as well as in genes that originally possessed low levels of methylation and subsequently acquired high levels of methylation. Besides the astrocyte model, this study also evaluated primary LGG tumors and confirmed the G-CIMP phenotype previously described in GBM (NOUSHMEHR et al., 2010) in a manner dependent on the presence of *IDH* mutation. The inhibition of the TET2 protein, responsible for demethylation by producing 5hmC, was also observed after *IDH1* mutation. Together, these observations unveiled mechanisms of how *IDH* mutation alters specific histone marks, induces DNA hypermethylation and reshapes the methylome, showing the interplay between genetic and epigenetic changes and the establishment of CIMP phenotype in gliomas (TURCAN et al., 2012). However, there are no published data with implications of G-CIMP subclassification (G-CIMP-low and G-CIMP-high) into the epigenetics of gliomas and how this affects patient clinical outcome.

12.2 *Novel G-CIMP subclassification*

CpG Island Methylator Phenotype (CIMP) subtype was first described in 1999, by Toyota et al. (1999) in a study that aimed to understand the global methylation patterns of CpG islands in colorectal cancer. Since then, CIMP has been identified in several different tumors, such as breast, bladder, lung, ovarian, etc. as reviewed by Miller, Sánchez-Vega e Elnitski (2016). Despite the lack of consensus on how CIMP phenotypes are related among different types of cancer, it is clear that the concept of CIMP has evolved recently. Perhaps the most striking finding is that CIMP is no longer restricted only to methylation changes at CpG islands (figure 11). As seen in gliomas, the high methylation levels of DNA outside CpG islands seem to play a definitive role in the biology of this phenotype.

And, more importantly, implications regarding other epigenomic mechanisms, like histone modifications, are also involved in G-CIMP subclassification.

Figure 11 – Effect of epigenetics in CIMP tumors.

Hypothetical G-CIMP+ locus (bottom) that is mediated by alteration in CpG methylation at functional elements, whereas non-tumor cell (top) and G-CIMP- cell (middle) shows no significant differences in DNA methylation. (MALTA et al., 2017)

Glioma-CpG Island Methylator Phenotype (G-CIMP) was first described as *IDH*-mutant gliomas manifesting a genome-wide hypermethylation of CpG islands (NOUSH-MEHR et al., 2010). Additionally, G-CIMP was also associated with a favorable clinical outcome compared to *IDH*-mutant gliomas that did not present CIMP. Importantly, advances in techniques to interrogate DNA methylation data, such as new methylation arrays (HM450), allowed the investigation of DNA methylation levels outside CpG islands, which resulted in the stratification of G-CIMP tumors into 2 subgroups with distinct clinical features: G-CIMP-low tumors characterized by lower DNA methylation levels, abnormalities in cell cycle genes and worse clinical outcome; and G-CIMP-high tumors presenting with higher DNA methylation levels and better prognosis (CECCARELLI et al., 2016). Differences in DNA methylation between G-CIMP-low and G-CIMP-high

were shown to be located at potential regulatory elements, which may be binding sites of transcription factors (*SOX* and *OLIG2*) related to proliferation in cancer (SUVÀ et al., 2014). It is also known that DNA methylation may alter local histone conformation (LEWIS; BIRD, 1991) suggesting that lower DNA methylation levels in G-CIMP-low tumors are not a random event and could indicate modulation of the chromatin structure leading to changes in the transcription machinery.

Remarkably, evidences supporting a potential tumor progression from G-CIMP-high to G-CIMP-low yielded valuable insights into malignant transformation captured by DNA methylation changes within *IDH*-mutant gliomas. Longitudinal analyses revealed that a subset of G-CIMP-high cases progressed to G-CIMP-low, and the molecular changes occurred at candidate noncoding functional elements, suggesting a potential master regulator affecting aggressive glioma progression upon recurrence (CECCARELLI et al., 2016). Possibly, master regulators or functional genomic elements (e.g., enhancers or silencers) drive glioma progression upon recurrence from G-CIMP-high to -low. Therefore, the identification of candidate histone modifications and transcription factors could elucidate differences in subgroups of tumors at primary and recurrent stages by capitalizing on next-generation sequencing and relationships between transcription factor binding, histone modification, DNA methylation and gene regulation.

13 Hypothesis

Histone modifications, such as H3K4me3 and H3K27ac, differentiates the epigenome of G-CIMP-low and G-CIMP-high tumors.

14 Objectives

14.1 *General objective*

To investigate the impact of H3K4me3 and H3K27ac in the epigenome of G-CIMP low and high tumors.

14.2 *Specific objectives*

The specific objectives of this study are:

- to select G-CIMP-low and G-CIMP-high samples from the tumor bank at Hermelin Brain Tumor Center (Henry Ford Hospital)
- to profile H3K4me3 and H3K27ac by ChIP-sequencing
- to preprocess and map H3K4me3 and H3K27ac data of G-CIMP-low and G-CIMP-high samples
- to detect differentially bound peaks between G-CIMP-low and G-CIMP-high
- to evaluate H3K4me3 and H3K27ac histone marks at biomarkers that can distinguish G-CIMP-low from G-CIMP-high (EReg1)
- to integrate ChIP-seq data with RNA-sequencing from G-CIMP tumors
- to interrogate potential active enhancers and predict target genes in G-CIMP tumors
- to identify master regulators of G-CIMP subtypes

15 Material and methods

15.1 Study design

The methodology used to analyze the epigenome of G-CIMP-low and G-CIMP-high included the following steps: sample selection and data generation, quality control and alignment, preprocessing, peak calling, detection of differential binding in ChIP-seq data, external datasets and, finally, prediction of potential enhancers.. A flowchart is represented in figure 12.

Figure 12 – Schematic flow chart of the study design from chapter 3.

15.2 Sample selection and data generation

A great proportion of TCGA glioma samples were collected at Henry Ford Hospital (22%, n=243), as seen in figure 13, including 45 G-CIMP samples. Therefore, a collaboration with Dr. Laila Poisson, biostatistician, and Dr. Steven Kalkanis, neurosurgeon, research members from Henry Ford Hospital involved in contributing to TCGA, was established with our group in order to proceed with further molecular characterization of G-CIMP tumors.

Figure 13 – Distribution of TCGA glioma samples across tissue source sites

Following the discovery of ERegs, it was fundamental to investigate other types of epigenetic modifications that could help explain the differences between glioma molecular subtypes. Thus, 10 fresh-frozen G-CIMP tumor samples were collected from Hermelin Brain Tumor Center (HBTC) tumor bank at Henry Ford Hospital (HFH), with extensive clinical follow-up, including treatment and imaging information, to profile two important histone modifications: H3K27ac and H3K4me3. These samples were selected based on data availability at TCGA, such as DNA-sequencing, RNA-sequencing and whole exome-sequencing, high tumor purity evaluated by neuropathologists and sufficient amount of tissue.

In collaboration with Dr. Ana DeCarvalho and her research team at HFH, about 100-400 mg of fresh frozen tissues by sample were cut. Then, tissue samples were sent to Active Motif, a company that offers epigenetics-related products and services, for the following steps, according to their protocol: samples were cross-linked for 10 minutes by

adding fresh formaldehyde directly to the culture medium at a final concentration of 1%. Using 10X (1.15M) glycine for 5 minutes at room temperature to quench the reaction. Chromatin from fixed cells were sonicated using a Bioruptor Pico (Diagenode, Cat # B01060001) with 30 seconds on/30 seconds off cycles to produce fragments between 200 and 500 base pairs. For immunoprecipitation, 100 grams of sonicated chromatin were used and 10 grams (10%) were saved as an input control. To probe for active enhancers, samples were incubated at 4C overnight with an H3K27ac antibody (Active Motif, Cat # 39133) or an IgG control (Sigma, Cat # R9133). As a secondary, protein A/G magnetic beads (Pierce, Cat # 88802) were added to the samples prior to an additional incubation for 2 hours at 4C. The beads were then washed with a series of salt buffers before elution. The immunoprecipitated and input control DNA were purified using A QIAprep Spin Miniprep Kit (Qiagen, Cat # 27104). Finally, the samples were single-end sequenced with read lengths of 75 bp each and an average coverage of $\sim 100\times$.

15.3 Quality control and alignment

First, FastQC (version 0.11.5) was used to do quality control checks by each sample on the raw sequence data, followed by MultiQC (version 1.4) (EWELS et al., 2016) to combine all reports into a single report by experiment. All the samples, from both ChIP-specific antibodies to H3K27ac and H3K4me3, showed average Phred scores above 30, low level of duplication and no adapter sequences content, as shown in figure 14. The software used to map the sequence files to the most recent reference genome (hg38) was bwa-mem (version 0.7.15). The output of this tool is a SAM file.

Figure 14 – Quality control report.

Mean quality scores are plotted by sample (top), sequence duplication level plot (middle) and percentage of sequences with adapter content (bottom). Plots represent H3K4me3 ChIP-seq (left) and H3K27ac (right).

15.4 Preprocessing

After the alignment, SAM files were converted to BAM, filtered to only include reads with mapping quality greater than 30 and, finally, sorted using samtools (version 1.3.1) (LI et al., 2009). Duplicated reads were tagged and removed using picard MarkDuplicates

tool (version 2.7.1). Figure 15 shows the total number of reads before the alignment, after filtering and removing duplicated reads.

Figure 15 – Read counts by step

Number of raw unaligned reads (top), read counts after qualify filtering (middle) and reads retained after removal of duplicates (bottom).

15.5 Peak calling

After mapping reads, peaks were called, by sample, to identify regions of ChIP enrichment ($\text{FDR} \leq 0.01$) in gliomas compared to control (input) using MACS2 (version 2.1.1) (ZHANG et al., 2008). The output of this tool contains the genomic location of each peak, followed by the absolute peak summit position, pileup height, fold enrichment over the control (input), log10 transformed p-value and FDR.

15.6 *Detection of differential binding in ChIP-seq data*

In order to identify differentially bound peaks between G-CIMP-low and G-CIMP-high samples, R/Bioconductor package DiffBind (ROSS-INNES et al., 2012) was used to analyze the dataset. This tool allows the user to input peak calling files from MACS2 and is composed by several steps. First, DiffBind reads in the files and associated metadata and then detect common peaks across all the samples towards creating a single set of binding site intervals.

Afterwards, DiffBind counts the number of reads that overlaps each binding site interval, by sample, using sequence read files. To do the differential analysis, DiffBind divides the samples by group according to the metadata provided by the user and then compares the groups by performing differential binding affinity analysis using DESeq2, by default. Finally, each peak is assigned with a fold-change, p-value, and FDR representing the confidence in which they are differentially bound. Differentially bound peaks were then assigned to discrete categories based on genomic position, using gencode (version 22) (HARROW et al., 2006) as reference for gene location: promoter (1,500 bp window surrounding known TSS) or intergenic (non-promoter) regions.

15.7 *Detection of differentially expressed genes*

RNA-sequencing data for 16 G-CIMP-low and 233 G-CIMP-high, aligned to hg38 and annotated by gencode (version 22), were downloaded from TCGA and processed using TCGAbiolinks (COLAPRICO et al., 2015). A supervised differential gene expression analysis was performed using Student's t-test between G-CIMP-low and G-CIMP-high. Pathway analysis was performed by Metascape (TRIPATHI et al., 2015).

15.8 *External datasets*

To further refine the list of differential binding peaks to highly confident targets, an genomic database, CISTROME, which is a publicly curated data portal of ChIP-seq across thousands of different tissues and cell types (LIU et al., 2011), was used to provide external datasets as validation. Table 4 refers to samples selected using CISTROME.

Table 4 – Number of samples selected using CISTROME

	H3K27ac	H3K4me3
GBM <i>IDH</i> wt	(n=3) (SUVÅ et al., 2014)	(n=3) (LIU et al., 2015)
Non-tumor brain	(n=3) (BERNSTEIN et al., 2010)	(n=3) (BERNSTEIN et al., 2010)
Normal pancreas	(n=1) (BERNSTEIN et al., 2010)	
Normal blood	(n=1) (PEETERS et al., 2015)	

15.9 Prediction of potential enhancer

The existence of H3K27ac mark outside TSS could indicate the presence of an active enhancer (CREYGHTON et al., 2010). Therefore, differentially bound H3K27ac peaks located at intergenic regions were then integrate with GeneHancer, a rich publicly curated database which catalogs all known genomic enhancers, candidate TF binding sites and their putative targeted genes (FISHILEVICH et al., 2017), to provide evidence of potential enhancers that might act in G-CIMP tumors. More than 285,000 enhancers are ranked and scored based on available functional experiments provided on these database.

Finally, inferred target genes regulated by the predicted enhancers were extracted and mapped to the RNA-sequencing data from G-CIMP tumors. The percentage of target genes that were up- or downregulated was calculated and combined into a master table.

15.10 Visualization

To visualize the interplay between ChIP-seq and RNA-sequencing, the data was uploaded to Genome Browser as a custom track in order to allow the integration with ENCODE and Roadmap data (KAROLCHIK et al., 2003).

16 Results

16.1 Clinical features of patient cohort

Since major DNA methylation changes between G-CIMP-low and G-CIMP-high were described both at promoter and potential regulatory regions, which possibly resulted in deregulation of nearby genes, a set of 4 G-CIMP-low and 6 G-CIMP-high were chosen to be profiled for H3K27ac and H3K4me3 by ChIP-sequencing in an attempt to identify active TSS and active enhancers related to these 2 glioma phenotypes. Table 5 shows the clinical features of the samples. Also, all of them are part of TCGA and the fresh-frozen tissues were stored at the tumor bank from the Hermelin Brain Tumor Center at Henry Ford Hospital.

Table 5 – Clinical features of G-CIMP-low and G-CIMP-high profiled by ChIP-seq

Sample-ID	Histology	Grade	Age	Gender	OS (months)	Subtype
TCGA-DU-6408	OD	III	23	FA	114	G-CIMP-high
TCGA-DU-6542	OA	III	25	MA	7.7	G-CIMP-high
TCGA-DU-7007	AS	II	33	MA	62.9	G-CIMP-high
TCGA-DU-7008	OD	II	41	FA	156.1	G-CIMP-high
TCGA-DU-7304	OA	III	43	MA	23.2	G-CIMP-high
TCGA-06-0221	GBM	IV	31	MA	19.8	G-CIMP-high
TCGA-DU-7010	AS	III	58	FA	14.9	G-CIMP-low
TCGA-06-0129	GBM	IV	30	MA	33.6	G-CIMP-low
TCGA-06-1805	GBM	IV	28	FA	4.2	G-CIMP-low
TCGA-06-2570	GBM	IV	21	FA	9.3	G-CIMP-low

16.2 Initial assessment

After data alignment, preprocessing and peak calling, the average number of peaks retained by each G-CIMP-low and G-CIMP-high sample is shown on figure 16. As expected, the number of H3K27ac peaks is higher than H3K4me3, possibly due to H3K27ac being known to be an active enhancer histone modification marker (HEINTZMAN et al., 2009).

By centering the peaks around all known TSS (n=60,483), which include coding and non-coding genes, based on gencode annotation, and calculating the mean by group (G-CIMP-low, G-CIMP-high, GBM-IDHwt, non-tumor brain and control), both H3K27ac and H3K4me3 clearly overlapped in promoters, confirming their putative role as a modification commonly associated with active transcription (figure 17).

Figure 16 – Boxplot of the number of peaks for each sample by ChIP-seq experiment. Each dot represents a sample. Circles are G-CIMP-high samples and triangles, G-CIMP-low. The y-axis shows the number of peaks.

Figure 17 – Histone modifications in the region surrounding TSS.

(A) Smoothed means of H3K4me3 peaks ± 2 kb around TSS. (B) Smoothed means of H3K27ac peaks ± 2 kb around TSS.

H3K4me3 and H3K27ac occupancy was also observed at the promoter of a house-keeping gene *GAPDH*, as seen in figure 18, in all the samples, which provides more evidence of the quality of the data.

Figure 18 – Genomic view of *GAPDH* on chromosome 12

Each track represents ChIP-seq values of each tissue type. The top 5 tracks represents H3K27ac peaks for G-CIMP-high, G-CIMP-low, GBM *IDH*wt, non-tumor brain and pancreas/blood. The next 3 tracks are H3K4me3 data between G-CIMP-high, G-CIMP-low and GBM *IDH*wt. ChIP inputs are also shown and derived from each tissue type independently. The last track indicates transcript annotation for this locus and gene expression across 53 human tissues, which brain is colored yellow.

16.3 Enriched levels of histone modifications at transcription start sites

A differential binding analysis of H3K4me3 and H3K27ac was performed between G-CIMP-low and G-CIMP-high and then mapped to known TSS. Additionally, a differential gene expression analysis of RNA-sequencing data between G-CIMP-low and G-CIMP-high

was also performed. The number of genes found by each analysis are represented in table 6. Peaks with $FDR \leq 0.05$ were considered differentially bound. 39,943 genes did not have a H3K4me3 mark on TSS detected by ChIP-seq, followed by 37,917 genes that did not have a H3K27ac mark at TSS for both G-CIMP-low and G-CIMP-high. Figure 19 shows a volcano plot with differential binding analysis by gene.

Table 6 – Differential binding analysis of H3K27ac and H3K4me3 at TSS by gene

	H3K4me3	H3K27ac
Gain	193	414
Loss	105	389
No difference	20,242	21,762

Gain means gain of H3K4me3 or H3K27ac in G-CIMP-low compared to G-CIMP-high. Loss means loss of H3K4me3 or H3K27ac in G-CIMP-low compared to G-CIMP-high. No difference means a $FDR > 0.05$.

Interestingly, the set of 15 paired CpG-gene biomarkers that define G-CIMP-low (EReg1) based on DNA methylation array and gene expression, do not show differential binding of either H3K4me3 or H3K27ac at TSS, except for *CD248* that showed gain of H3K27ac at TSS by G-CIMP-low compared to G-CIMP-high. This might indicate that another mechanism, such as the activation of a nearby enhancer, might be regulating the set of genes defined as EReg1 in G-CIMP-low tumors.

The differential gene expression analysis of RNA-sequencing data revealed 7,468 downregulated genes and 4,115 upregulated genes in G-CIMP-low compared to G-CIMP-high, as seen in figure 20. In addition, a gene expression pathway analysis of genes upregulated in G-CIMP-low showed an enrichment of cell cycle related genes (figure 21).

Next, in order to correlate histone marks at TSS with gene expression, both H3K4me3 and H3K27ac ChIP-seq data were integrated with RNA-sequencing (figure 22). *SLC14A1*, for example, was found to be highly expressed in mature astrocytes of non-tumor cells (ZHANG et al., 2016) and it is silent in G-CIMP-low. On the other hand, *CDK4* is active in G-CIMP-low and was shown to be an important cell cycle related gene, which its activation have been showed to be associated with poor survival in GBM (KIM et al., 2010). Also, *HILS1* was reported to be a pseudogene that functions as a chromatin modifier (YAN et al., 2003). Another interesting example is the gene *SIM2* that was poised for activation in G-CIMP-high and became active in G-CIMP-low (figure 23). This gene

was reported to regulate cell invasion in glioblastoma samples leading to a progressive state (SU et al., 2014).

Figure 19 – Epigenomic changes associated with G-CIMP low.

(A) Differential binding of H3K4me3 peaks at TSS. Each dot is a gene. (B) Differential binding of H3K27ac peaks at TSS. Orange means gain of H3K4me3 or H3K27ac by G-CIMP-low and blue means loss of H3K4me3 or H3K27ac by G-CIMP-low.

Figure 20 – Gene expression changes associated with G-CIMP low.

Differential expression of genes. Each dot is a gene. Orange means upregulation of the gene by G-CIMP-low and blue means downregulation.

Figure 21 – Pathway analysis of upregulated genes in G-CIMP-low. The x-axis represents the $-1 \cdot \log_{10}$ transformed p-value.

Figure 22 – Chromatin changes associated with transcription.

Scatter plot of all H3K27ac and H3K4me3 peaks overlapping known TSS. Significant gains or losses of H3K27ac (y-axis) and H3K4me3 (x-axis) are highlighted. Each dot indicates a gene and the shape indicates expression difference between G-CIMP-low vs. G-CIMP-high. Circles are downregulated and squares are upregulated. If a gene is enriched for both H3K27ac and H3K4me3, this implies active TSS and the associated gene expression is defined as upregulated (highlighted in purple, upper right corner). If a gene is depleted for both H3K4me3 and H3K27ac, this implies weak or quiescent expression of the associated gene (highlighted in blue, lower left corner). Additional information about bivalent or poised expression is also obtained from this plot.

Taken together, the differential binding analysis of H3K27ac and H3K4me3 suggests that major histone modifications occur in G-CIMP-low samples, compared to G-CIMP-high, which could help explain the more aggressive phenotype of G-CIMP-low with the activation of cell cycle genes mediated by chromatin remodeling at TSS.

Figure 23 – *SIM2* activation by gain of H3K27ac at TSS.

Each track represents ChIP-seq values of each tissue type. The top 5 tracks represents H3K27ac peaks for G-CIMP-high, G-CIMP-low, GBM *IDHwt*, non-tumor brain and pancreas/blood. The next 3 tracks are H3K4me3 data between G-CIMP-high, G-CIMP-low and GBM *IDHwt*. ChIP inputs are also shown and derived from each tissue type independently. The last track indicates transcript annotation for this locus. *SIM2* TSS is highlighted in blue.

16.4 Regulatory elements associated with G-CIMP-low

Another interesting feature of H3K27ac mark is the possibility of identifying active enhancers at intergenic regions. To investigate that, differential binding H3K27ac peaks were mapped to the following genomic locations: exon, intron and intergenic. Figure 24 shows the number of peaks found divided by gain in G-CIMP-low and loss in G-CIMP-low based on fold-change. All the data outside the TSS was considered as intergenic for the following steps.

Next, the data was integrated with GeneHancer to identify any previously defined enhancers, thereby filtering the list to robust master regulator candidates. GeneHancer database contains more than 240,000 enhancers and 7,651 of them mapped with the differential binding H3K27ac peaks of G-CIMP-low and G-CIMP-high. GeneHancer database also provides target genes which were shown to be activated by each enhancer. With that,

the target genes were extracted from the database and combined with the G-CIMP-low and G-CIMP-high RNA-seq gene expression data. Finally, the number of upregulated target genes nearby an active enhancer was divided by the number of target genes that showed either no difference or that were downregulated, generating then a percentage of potential genes regulated by each enhancer. The same approach was applied to downregulated genes and enhancers that lost activity in G-CIMP-low. The potential list of enhancers was compiled into a master table and represented in figure 25.

Figure 24 – Distribution of H3K27ac marks outside TSS.

H3K27ac peaks outside TSS were classified according to genomic features such as exon, intron or intergenic.

Figure 25 – Regulatory elements predicted by H3K27ac ChIP-seq and GeneHancer.

Each dot represents a potential enhancer. The size of the circle represents the percentage of target genes nearby the enhancer that is either upregulated (red) or downregulated (green) in G-CIMP-low compared to G-CIMP-high.

Interestingly, the Homeobox (HOX) A cluster of genes, which was shown to be activated in glioma initiating cells (GALLO et al., 2013), was upregulated in G-CIMP-low compared to G-CIMP-high and 3 nearby enhancers were potentially active in G-CIMP-low, by gain of H3K27ac mark at intergenic regions and confirmed with GeneHancer data (figure 26).

Figure 26 – *HOXA* cluster might be activated by nearby enhancers in G-CIMP-low

Each track represents ChIP-seq values of each tissue type. The top 5 tracks represents H3K27ac peaks for G-CIMP-high, G-CIMP-low, GBM *IDH*wt, non-tumor brain and pancreas/blood. The next 3 tracks are H3K4me3 data between G-CIMP-high, G-CIMP-low and GBM *IDH*wt. ChIP inputs are also shown and derived from each tissue type independently. The last track indicates gene annotation for this locus. Genes highlighted in red are significantly upregulated in G-CIMP-low compared to G-CIMP-high. Highlighted in blue are potential active enhancers enriched in G-CIMP-low.

A DNA motif scan analysis performed using 2,467 potential active enhancers revealed an enrichment for the homeobox motif (geometric test $p=1e-5$, fold enrichment=1.35), which might suggest that the *HOX* genes play an important role as master regulators in G-CIMP-low tumors.

17 Discussion

For the first time, the landscape of epigenomic modifications stratifying G-CIMP tumors into G-CIMP-low and G-CIMP-high was described. Even though G-CIMP-low and G-CIMP-high are both gliomas harboring *IDH* mutation without 1p and 19q codeletion, their epigenomes carry major differences. G-CIMP-low samples were previously characterized with an enrichment of alterations in cell cycle genes and, interestingly, by investigating the chromatin state of the promoter regions of these genes, it was found that the epigenome of G-CIMP-low presented modifications of H3K4me3 and H3K27ac histone marks for some of these genes. This finding corroborates the idea that gene activation may be correlated with epigenomic modifications. Besides, it was possible to newly identify a large number of genes with alterations in G-CIMP-low compared to G-CIMP-high which could serve as additional biomarkers to distinguish G-CIMP subtypes.

The integration of H3K27ac and H3K4me3 (profiled by ChIP-sequencing) with gene expression RNA-sequencing provided insights into the tumorigenic events that drive aggressiveness in G-CIMP-low patients. For instance, *SLC14A1* was shown to be downregulated in G-CIMP-low followed by loss of H3K4me3 and H3K27ac at TSS compared to G-CIMP-high. This gene encodes a protein related to transmembrane transporter activity of water and urea, and was shown to be active both in mature astrocyte cells (ZHANG et al., 2016) and normal brain (LUCIEN et al., 2005). On the other hand, *CDK4*, a gene whose function is important for cell cycle G1 phase progression (SERRANO; HANNON; BEACH, 1993), presented histone modifications at TSS along with upregulation in G-CIMP-low. This gene was reported to be a biomarker for poor survival in GBM (KIM et al., 2010) and, currently, there are 2 approved drugs that inhibit *CDK4* to treat breast cancer (FINN et al., 2009). This finding suggests potential drug targets for G-CIMP-low patients that can be incorporated as effective therapeutic strategies tailored to glioma subtypes. In addition to that, a pseudogene called *HILS1*, that might function as a chromatin modifier, was identified as active with both ChIP-seq and RNA-seq data from G-CIMP-low (YAN et al., 2003). This gene is important during mammalian spermiogenesis and might regulate gene transcription. Another gene already described in the literature as associated with tumorigenesis in GBM was *SIM2* (SU et al., 2014). Interestingly, *SIM2* was poised for activation in G-CIMP high and became active in G-CIMP low, which might suggest that

the recruitment of this gene is important to G-CIMP-low. Additionally, non-coding genes, such as *RP11-1109M24.5* and *RP11-680F20.10*, seem to be highly active in G-CIMP-low compared to G-CIMP-high, and they have never been described in the literature of gliomas before. The identification of new biomarkers could be informative to provide therapeutic targets or to complement standard clinical diagnostics.

Exploratory analysis of G-CIMP tumors provided evidence that the chromatin of G-CIMP-low might have been reorganized in order to control transcription of genes, which is strongly linked to histone modifications at regulatory elements. The discovery of potential active enhancers nearby active genes, such as the *HOXA* cluster, shows the importance of investigating intergenic regions. Notably, these functional elements might regulate transcription without causing genomic alterations, that could only be captured by epigenomic profiling; which highlights, once again, the importance of epigenetics in gliomagenesis and progression. *HOX* genes encode transcription factors which play an important role in proper embryologic development (GAUNT; SHARPE; DUBOULE, 1988). In particular, *HOXA* cluster genes, located in chromosome 7, were reported to be fundamental to differentiation and development of multiple organisms (RINGROSE; PARO, 2004). Interestingly, *HOXA* genes were shown to be upregulated and predictive of poor clinical outcome in glioblastomas (COSTA et al., 2010; MURAT et al., 2008). Additionally, the presence of 3 potential enhancers within the *HOXA* cluster, followed by an enrichment for the homeobox motif signature found in active enhancers, yielded insights that *HOX* genes might be master regulators of G-CIMP-low.

To validate the hypothesis that epigenomic modifications are potential drivers of G-CIMP-low, further functional studies must be conducted. In an appropriate *IDH*-mutant cell-culture system, techniques, such as chromosome conformation capture to define the interactions between regulatory elements and target genes, or targeted CRISPR/Cas-9 to delete the genomic elements, can be performed using a G-CIMP-low and G-CIMP-high model in order to provide more evidence regarding the relation between epigenetics and gene regulation in gliomas.

18 Conclusion

The assessment of H3K4me3 and H3K27ac through ChIP-sequencing enabled mapping histone modifications at promoter regions between G-CIMP-low and G-CIMP-high. When integrated with gene expression RNA-sequencing, most of the gains of both H3K4me3 and H3K27ac at TSS showed positive correlation with upregulation of the gene in G-CIMP-low compared to G-CIMP-high. The same could be seen with the majority of losses of H3K4me3 and H3K27ac at TSS followed by downregulation of the gene in G-CIMP-low. Several genes upregulated in G-CIMP-low have been described as associated with proliferation in glioblastomas, which links G-CIMP-low to aggressiveness.

Interestingly, paired CpG sites and genes that were described as biomarkers to discriminate G-CIMP-low from G-CIMP-high (EReg1) did not show histone modifications associated with H3K4me3 and H3K27ac.

Additionally, potential active enhancers were identified by H3K27ac-marked inter-genic regions in G-CIMP-low compared to G-CIMP-high. Also, an *in silico* validation with an enhancer database (GeneHancer) showed predicted target genes regulated by these enhancers and, surprisingly, it was found that possibly these enhancers are acting in the regulation of important group of genes. Potential master regulators of G-CIMP-low were the *HOX* cluster of genes, fundamental to normal development and linked to glioma cancer stem cells (TABUSE et al., 2011). The identification of master regulators could elucidate mechanisms that drive glioma progression from G-CIMP-high to G-CIMP-low with follow-up studies.

Part IV

Synthesis and perspective

19 Synthesis

IDH mutation and the integrity of chromosomes 1p and 19q were previously reported as relevant biomarkers to gliomas. However, as seen in chapter 2, this study was able to segregate even further glioma subtyping using DNA methylation patterns. Among *IDH*-wildtypes, a subset of patients harbored a silent genomic landscape and displayed similarity to pilocytic astrocytoma. This phenotype, PA-like, provided relative survival advantage in relation to other *IDH*-wildtypes. Within *IDH* mutants, two novel subgroups were identified with significant clinical implications: G-CIMP-low, with lower levels of DNA methylation and a poor outcome, and G-CIMP-high, characterized by higher DNA methylation and good overall survival. Interestingly, the findings showed that there might be a progression from G-CIMP-high to G-CIMP-low. However, epigenetic mechanisms associated with DNA demethylation are yet to be determined. As shown in chapter 3, by investigating the chromatin state of G-CIMP+ tumors, important features that define G-CIMP-low and G-CIMP-high were identified. Among them, it was found that histone modifications associated with active promoter of cell cycle genes characterize G-CIMP-low along with potential enhancer elements that regulate gene activation. Taken together, the epigenomic landscape of gliomas yielded valuable insights into tumorigenic events that might drive glioma phenotype with opportunity for further targeted therapeutic strategies.

20 Perspective

In order to investigate the hypothesis that G-CIMP-low might progress from primary G-CIMP-high tumors, the epigenome of longitudinal matched glioma samples (primary and multiple recurrent) will be evaluated. Fresh-frozen glioma tissues from the Hermelin Brain Tumor Center at Henry Ford Hospital will be obtained, along with clinical records, to generate methylome data to identify G-CIMP-high and G-CIMP-low molecular phenotypes by applying previously defined ERegs. Then, cases predicted to be G-CIMP-low and G-CIMP-high will be selected to generate chromatin immunoprecipitation-, whole-genome bisulfite-, and RNA- sequencing data, and an integrated omic analysis to expand EReg signatures will be performed. Also, by expanding the number of samples, possibly more epigenomic differences will be identified between G-CIMP-high and G-CIMP-low. Finally, the molecular signatures at primary diagnosis that are specifically associated with the progression to G-CIMP-low upon recurrence will be analyzed.

While genomic profiling studies continue to accelerate the discovery of novel molecular targets, often functional validation of these targets is overlooked. An established pipeline will be incorporated for functional follow-up of noncoding targets using targeted genome editing - a powerful and highly efficient approach for testing the effects of removing regulatory elements or modulating target gene expression. Moreover, these tools will be implemented in cell line models that represent biologically relevant models of glioma, which is critical given the known cell-type specificity of noncoding elements. Several independent approaches will be applied in to order to confirm the role of these elements in G-CIMP-low, such as integrating our data with published epigenomics data and chromosome conformation capture, shRNA and CRISPR-cas9.

The identification and functional validation of master regulators associated with glioma progression to G-CIMP-low could lead the discovery of new drug targets and, also, provide opportunities for refined clinical trials. Ideally, in the future, patients could have personalized tumor care by integrating molecular research, surgical technologies and precision medicine in order to provide better treatment options (figure 27). This goal can only be achieved with the interplay between basic science of tumor biology and clinical medicine.

Figure 27 – Precision medicine.

21 Contributions to science

21.1 Publications

- THE CANCER GENOME ATLAS RESEARCH NETWORK. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, v. 455, n. 7216, p. 1061-1068, 2015.

This was my first contribution to The Cancer Genome Atlas as a member of the analysis working group. My role was to investigate the DNA methylation of lower-grade gliomas, specially in an unsupervised fashion. In this study, we were able to integrate different molecular data to finally classify these tumors in three major subgroups, which contributed to the revision of glioma classification by the World Health Organization, using, for the first time, molecular parameters in addition to histology to define tumor entities. Our group was responsible for writing the methodology, organizing results and figures associated with DNA methylation analysis.

- CECCARELLI, M.*; BARTHEL, F.P.*; MALTA, T.M.*; **SABEDOT; T.S.*** et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *Cell*, v. 164, n. 3, p. 550-536, 2016.

This study was my main publication during my PhD. Prof. Dr. Houtan Noushmehr, Dr. Tathiane Malta and me led the DNA methylation analysis of LGG and GBM samples. Our group was able to identify 7 subgroups of gliomas along with biomarkers that can distinguish each subtype. We were highly involved during the manuscript preparation, figure designs and further discussions.

- COLAPRICO, A.; SILVA, T.C.; OLSEN, C.; GAROFANO, D.; **SABEDOT; T.S.**; MALTA, T.M.; PAGNOTTA, S.M.; CASTIGLIONI, I.; CECCARELLI, M.; BONTEMPI, G.; NOUSHMEHR, H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, v. 44, n. 8, p. 71, 2015.

During the development of TCGAbiolinks, I contributed with R scripts to analyze DNA methylation and to integrate DNA methylation with gene expression data.

- MALTA, T.M.; SOUZA, C.F.; **SABEDOT; T.S.**; SILVA, T.C.; MOSELLA, M.Q.S.; KALKANIS, S.N.; SNYDER, J.; CASTRO, A.V.B.C.; NOUSHMEHR, H. Glioma

CpG Island Methylator Phenotype (G-CIMP): Biological and Clinical Implications. *Neuro-Oncology*, 2017.

In this review, my role was to help describe new findings regarding the epigenome annotation of G-CIMP-low and G-CIMP-high.

- MALL, R.; CERULO, L.; KUNJI, K.; BENSMAIL, H.; **SABEDOT; T.S.**; NOUSHMEHR, H.; IAVARONE, A.; CECCARELLI, M. RGBM: regularized gradient boosting machines for identification of the transcriptional regulators of discrete glioma subtypes. *Nucleic Acids Research*, 2018.

My contribution to the development of RGBM was discussing the validation section in which the tool was used to identify master regulators of the 7 molecular subtypes of gliomas.

21.2 Presentations

21.2.1 Poster

- SABEDOT, T. S.; MALTA, T. M., THE CANCER GENOME ATLAS RESEARCH NETWORK; NOUSHMEHR, H. (2015, August). Identification and characterization of prognostic biomarkers for adult glioma. II Simpósio Internacional de Medicina Personalizada organized by Hospital Albert Einstein, São Paulo, SP, Brazil.
- SABEDOT, T. S.; MALTA, T. M., DIHN, H.; CECCARELLI, M.; BERMAN, B.; NOUSHMEHR, H. (2016, November). Epigenetic alterations at intergenic regions associated with progression in a subset of *IDH* mutant gliomas. Annual meeting of the Society for Neuro-Oncology, Scottsdale, AZ, USA.
- SABEDOT, T. S.; NOUSHMEHR, H. (2017, May). Omics profiling of G-CIMP tumors reveals dynamic modification of chromatin associated with regulatory regions. Research Symposium organized by Henry Ford Hospital, Detroit, MI, USA.
- SABEDOT, T. S.; POISSON, L., DeCarvalho, A.; NOUSHMEHR, H. (2017, November). Depletion of 5-hydroxymethylcytosine in aggressive G-CIMP subtype. Annual meeting of the Society for Neuro-Oncology, San Francisco, CA, USA.

21.2.2 Oral presentation

- SABEDOT, T. S.; NOUSHMEHR, H. (2016, March). Whole-genome bisulfite sequencing of a subset of *IDH* mutant glioma reveals changes at intergenic regions. Update on Neuro-Oncology organized by the Society for Neuro-Oncology Latin America, Rio de Janeiro, RJ, Brazil.

21.3 Academic and professional honors

- Honorable Mention at Society for Neuro-Oncology Latin America (SNOLA) 2016 – Updates in Neuro-Oncology for the oral presentation. March, 2016. Rio de Janeiro, RJ, Brazil.
- Prêmio Pró-Reitoria de Pesquisa 2016 regarding high-impact publication (CECCA-RELLI et al., 2016) provided by USP. August, 2016. Ribeirão Preto, SP, Brazil

21.4 Professional memberships

- 2017- Member, Society for Neuro-Oncology (SNO)
- 2018- Member, American Association for Cancer Research (AACR)

21.5 Peer review

In February of 2018, I was invited to review a manuscript submitted to *Genetics and Molecular Biology*.

Bibliography

- AHMED, R. et al. Malignant gliomas: current perspectives in diagnosis, treatment, and early response assessment using advanced quantitative imaging methods. *Cancer management and research*, Dove Press, v. 6, p. 149, 2014. Citado na página 19.
- BANNISTER, A. J.; KOUZARIDES, T. Regulation of chromatin by histone modifications. *Cell research*, Nature Publishing Group, v. 21, n. 3, p. 381, 2011. Citado na página 23.
- BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, JSTOR, p. 289–300, 1995. Citado na página 34.
- BERMAN, B. P. et al. Regions of focal dna hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature genetics*, Nature Research, v. 44, n. 1, p. 40–46, 2012. Citado na página 22.
- BERNSTEIN, B. E. et al. The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, Nature Publishing Group, v. 28, n. 10, p. 1045, 2010. Citado na página 68.
- BIRD, A. P. CpG-rich islands and the function of dna methylation. *Nature*, Springer, v. 321, n. 6067, p. 209–213, 1986. Citado na página 21.
- BRENNAN, C. W. et al. The somatic genomic landscape of glioblastoma. *Cell*, Elsevier, v. 155, n. 2, p. 462–477, 2013. Citado 3 vezes nas páginas 22, 26, and 29.
- CECCARELLI, M. et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, Elsevier, v. 164, n. 3, p. 550–563, 2016. Citado 4 vezes nas páginas 58, 59, 92, and 105.
- CEDAR, H.; BERGMAN, Y. Linking dna methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, Nature Publishing Group, v. 10, n. 5, p. 295–304, 2009. Citado na página 23.
- COLAPRICO, A. et al. Tcgabioblinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, Oxford University Press, v. 44, n. 8, p. e71–e71, 2015. Citado 2 vezes nas páginas 50 and 67.
- CORTESSIS, V. K. et al. Environmental epigenetics: prospects for studying epigenetic mediation of exposure–response relationships. *Human genetics*, Springer, v. 131, n. 10, p. 1565–1589, 2012. Citado na página 21.
- COSTA, B. M. et al. Reversing hoxa9 oncogene activation by pi3k inhibition: epigenetic mechanism and prognostic significance in human glioblastoma. *Cancer research*, AACR, v. 70, n. 2, p. 453–462, 2010. Citado na página 84.
- CREYGHTON, M. P. et al. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 107, n. 50, p. 21931–21936, 2010. Citado 2 vezes nas páginas 23 and 68.
- DANG, L. et al. Cancer-associated idh1 mutations produce 2-hydroxyglutarate. *Nature*, Nature Publishing Group, v. 462, n. 7274, p. 739–744, 2009. Citado na página 22.

DIXON, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, Nature Research, v. 485, n. 7398, p. 376–380, 2012. Citado na página 24.

ECCLESTON, A. et al. Epigenetics. *Nature*, Nature Publishing Group, v. 447, n. 7143, p. 395–395, 2007. Citado na página 21.

ECKEL-PASSOW, J. E. et al. Glioma groups based on 1p/19q, idh, and tert promoter mutations in tumors. *New England Journal of Medicine*, Mass Medical Soc, v. 372, n. 26, p. 2499–2508, 2015. Citado na página 39.

ESTELLER, M. Epigenetics in cancer. *n Engl j Med*, Mass Medical Soc, v. 2008, n. 358, p. 1148–1159, 2008. Citado na página 22.

EWELS, P. et al. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, Oxford University Press, v. 32, n. 19, p. 3047–3048, 2016. Citado na página 64.

FINN, R. S. et al. Pd 0332991, a selective cyclin d kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro. *Breast Cancer Research*, BioMed Central, v. 11, n. 5, p. R77, 2009. Citado na página 83.

FISHILEVICH, S. et al. Genehancer: genome-wide integration of enhancers and target genes in genecards. *Database*, Oxford University Press, v. 2017, 2017. Citado na página 68.

FLAVAHAN, W. A. et al. Insulator dysfunction and oncogene activation in idh mutant gliomas. *Nature*, Nature Publishing Group, v. 529, n. 7584, p. 110, 2016. Citado na página 24.

GALLO, M. et al. A tumorigenic mll-homeobox network in human glioblastoma stem cells. *Cancer research*, AACR, v. 73, n. 1, p. 417–427, 2013. Citado na página 81.

GAUNT, S. J.; SHARPE, P. T.; DUBOULE, D. Spatially restricted domains of homeo-gene transcripts in mouse embryos: relation to a segmented body plan. *Development*, The Company of Biologists Ltd, v. 104, n. Supplement, p. 169–179, 1988. Citado na página 84.

GILBERT, N. et al. Dna methylation affects nuclear organization, histone modifications, and linker histone binding but not chromatin compaction. *J Cell Biol*, Rockefeller University Press, v. 177, n. 3, p. 401–411, 2007. Citado na página 23.

GUENTHER, M. G. et al. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, Elsevier, v. 130, n. 1, p. 77–88, 2007. Citado na página 23.

GUINTIVANO, J.; ARYEE, M. J.; KAMINSKY, Z. A. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*, Taylor & Francis, v. 8, n. 3, p. 290–302, 2013. Citado 2 vezes nas páginas 32 and 34.

HARROW, J. et al. Gencode: producing a reference annotation for encode. *Genome biology*, BioMed Central, v. 7, n. 1, p. S4, 2006. Citado na página 67.

- HAYOSTEK, C. J. et al. Astrocytomas of the cerebellum: a comparative clinicopathologic study of pilocytic and diffuse astrocytomas. *Cancer*, Wiley Online Library, v. 72, n. 3, p. 856–869, 1993. Citado na página 52.
- HEGI, M. E. et al. Mgmt gene silencing and benefit from temozolomide in glioblastoma. *New England Journal of Medicine*, Mass Medical Soc, v. 352, n. 10, p. 997–1003, 2005. Citado na página 20.
- HEINTZMAN, N. D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, Nature Publishing Group, v. 459, n. 7243, p. 108, 2009. Citado na página 69.
- HEINZ, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, Elsevier, v. 38, n. 4, p. 576–589, 2010. Citado na página 37.
- HEMBERGER, M.; DEAN, W.; REIK, W. Epigenetic dynamics of stem cells and cell lineage commitment: digging waddington’s canal. *Nature reviews Molecular cell biology*, Nature Publishing Group, v. 10, n. 8, p. 526–537, 2009. Citado na página 21.
- JONES, D. T. et al. Recurrent somatic alterations of fgfr1 and ntrk2 in pilocytic astrocytoma. *Nature genetics*, Nature Publishing Group, v. 45, n. 8, p. 927, 2013. Citado na página 46.
- KARLIĆ, R. et al. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 107, n. 7, p. 2926–2931, 2010. Citado na página 23.
- KAROLCHIK, D. et al. The ucsc genome browser database. *Nucleic acids research*, Oxford University Press, v. 31, n. 1, p. 51–54, 2003. Citado na página 68.
- KIM, H. et al. Integrative genome analysis reveals an oncomir/oncogene cluster regulating glioblastoma survivorship. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 107, n. 5, p. 2183–2188, 2010. Citado 2 vezes nas páginas 73 and 83.
- KOUZARIDES, T. Chromatin modifications and their function. *Cell*, Elsevier, v. 128, n. 4, p. 693–705, 2007. Citado na página 23.
- LAMBERT, S. R. et al. Differential expression and methylation of brain developmental genes define location-specific subsets of pilocytic astrocytoma. *Acta neuropathologica*, Springer, v. 126, n. 2, p. 291–301, 2013. Citado 2 vezes nas páginas 37 and 46.
- LEWIS, J.; BIRD, A. Dna methylation and chromatin structure. *FEBS letters*, Elsevier, v. 285, n. 2, p. 155–159, 1991. Citado na página 59.
- LI, H. et al. The sequence alignment/map format and samtools. *Bioinformatics*, Oxford University Press, v. 25, n. 16, p. 2078–2079, 2009. Citado na página 65.
- LIANG, Y. et al. Activation of vascular endothelial growth factor a transcription in tumorigenic glioblastoma cell lines by an enhancer with cell type-specific dnase i accessibility. *Journal of Biological Chemistry*, ASBMB, v. 277, n. 22, p. 20087–20094, 2002. Citado na página 23.

LIU, F. et al. Egfr mutation promotes glioblastoma through epigenome and transcription factor network remodeling. *Molecular cell*, Elsevier, v. 60, n. 2, p. 307–318, 2015. Citado na página 68.

LIU, T. et al. Cistrome: an integrative platform for transcriptional regulation studies. *Genome biology*, BioMed Central, v. 12, n. 8, p. R83, 2011. Citado na página 67.

LODATO, M. A. et al. Sox2 co-occupies distal enhancer elements with distinct pou factors in escs and npcs to specify cell state. *PLoS genetics*, Public Library of Science, v. 9, n. 2, p. e1003288, 2013. Citado na página 45.

LOUIS, D. N. et al. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, Springer, v. 131, n. 6, p. 803–820, 2016. Citado 2 vezes nas páginas 19 and 54.

LUCIEN, N. et al. Ut-b1 urea transporter is expressed along the urinary and gastrointestinal tracts of the mouse. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, Am Physiological Soc, v. 288, n. 4, p. R1046–R1056, 2005. Citado na página 83.

MALTA, T. M. et al. Glioma cpg island methylator phenotype (g-cimp): biological and clinical implications. *Neuro-oncology*, 2017. Citado 3 vezes nas páginas 19, 54, and 58.

MEDVEDEVA, Y. A. et al. Effects of cytosine methylation on transcription factor binding sites. *BMC genomics*, BioMed Central, v. 15, n. 1, p. 119, 2014. Citado na página 22.

MILLER, B. F.; SÁNCHEZ-VEGA, F.; ELNITSKI, L. The emergence of pan-cancer cimp and its elusive interpretation. *Biomolecules*, Multidisciplinary Digital Publishing Institute, v. 6, n. 4, p. 45, 2016. Citado na página 57.

MUR, P. et al. Codeletion of 1p and 19q determines distinct gene methylation and expression profiles in idh-mutated oligodendroglial tumors. *Acta neuropathologica*, Springer, v. 126, n. 2, p. 277–289, 2013. Citado 3 vezes nas páginas 37, 46, and 48.

MURAT, A. et al. Stem cell-related “self-renewal” signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. *Journal of clinical oncology*, American Society of Clinical Oncology, v. 26, n. 18, p. 3015–3024, 2008. Citado na página 84.

NOUSHMEHR, H. et al. Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma. *Cancer cell*, Elsevier, v. 17, n. 5, p. 510–522, 2010. Citado 5 vezes nas páginas 20, 22, 26, 57, and 58.

ORZAN, F. et al. Enhancer of zeste 2 (ezh2) is up-regulated in malignant gliomas and in glioma stem-like cells. *Neuropathology and applied neurobiology*, Wiley Online Library, v. 37, n. 4, p. 381–394, 2011. Citado na página 23.

PARSONS, D. W. et al. An integrated genomic analysis of human glioblastoma multiforme. *Science*, American Association for the Advancement of Science, v. 321, n. 5897, p. 1807–1812, 2008. Citado na página 22.

PEETERS, J. G. et al. Inhibition of super-enhancer activity in autoinflammatory site-derived t cells reduces disease-associated gene expression. *Cell reports*, Elsevier, v. 12, n. 12, p. 1986–1996, 2015. Citado na página 68.

- PLONGTHONGKUM, N.; DIEP, D. H.; ZHANG, K. Advances in the profiling of dna modifications: cytosine methylation and beyond. *Nature Reviews Genetics*, Nature Research, v. 15, n. 10, p. 647–661, 2014. Citado na página 24.
- RICHMOND, T. J.; DAVEY, C. A. The structure of dna in the nucleosome core. *Nature*, Nature Publishing Group, v. 423, n. 6936, p. 145, 2003. Citado na página 23.
- RINGROSE, L.; PARO, R. Epigenetic regulation of cellular memory by the polycomb and trithorax group proteins. *Annu. Rev. Genet.*, Annual Reviews, v. 38, p. 413–443, 2004. Citado na página 84.
- RISSO, D. et al. Gc-content normalization for rna-seq data. *BMC bioinformatics*, BioMed Central, v. 12, n. 1, p. 480, 2011. Citado na página 33.
- ROSS-INNES, C. S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, Nature Publishing Group, v. 481, n. 7381, p. 389, 2012. Citado na página 67.
- SANDER, C. Genomic medicine and the future of health care. *Science*, American Association for the Advancement of Science, v. 287, n. 5460, p. 1977–1978, 2000. Citado na página 33.
- SCHINKE, C. et al. Aberrant dna methylation in malignant melanoma. *Melanoma research*, NIH Public Access, v. 20, n. 4, p. 253, 2010. Citado na página 21.
- SERRANO, M.; HANNON, G. J.; BEACH, D. A new regulatory motif in cell-cycle control causing specific inhibition of cyclin d/cdk4. *nature*, Nature Publishing Group, v. 366, n. 6456, p. 704, 1993. Citado na página 83.
- SHARMA, S.; KELLY, T. K.; JONES, P. A. Epigenetics in cancer. *Carcinogenesis*, Oxford University Press, v. 31, n. 1, p. 27–36, 2010. Citado na página 23.
- SOUZA, C. F. de et al. Distinct epigenetic shift in a subset of glioma cpG island methylator phenotype (g-cimp) during tumor recurrence. *bioRxiv*, Cold Spring Harbor Laboratory, p. 156646, 2017. Citado na página 52.
- STUPP, R. et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, Mass Medical Soc, v. 352, n. 10, p. 987–996, 2005. Citado na página 19.
- STURM, D. et al. Hotspot mutations in h3f3a and idh1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer cell*, Elsevier, v. 22, n. 4, p. 425–437, 2012. Citado 4 vezes nas páginas 26, 37, 46, and 48.
- SU, Y. et al. Targeting sim2-s decreases glioma cell invasion through mesenchymal–epithelial transition. *Journal of cellular biochemistry*, Wiley Online Library, v. 115, n. 11, p. 1900–1907, 2014. Citado 2 vezes nas páginas 74 and 83.
- SUVÀ, M. L. et al. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell*, Elsevier, v. 157, n. 3, p. 580–594, 2014. Citado 3 vezes nas páginas 45, 59, and 68.

TABUSE, M. et al. Functional analysis of hoxd9 in human gliomas and glioma cancer stem cells. *Molecular cancer*, BioMed Central, v. 10, n. 1, p. 60, 2011. Citado na página 85.

TAHILIANI, M. et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. *Science*, American Association for the Advancement of Science, v. 324, n. 5929, p. 930–935, 2009. Citado na página 22.

THE CANCER GENOME ATLAS RESEARCH NETWORK. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, Nature Publishing Group, v. 455, n. 7216, p. 1061–1068, 2008. Citado 2 vezes nas páginas 22 and 26.

THE CANCER GENOME ATLAS RESEARCH NETWORK. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, Nature Publishing Group, v. 513, n. 7517, p. 202, 2014. Citado na página 35.

THE CANCER GENOME ATLAS RESEARCH NETWORK. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, Nature Publishing Group, v. 507, n. 7492, p. 315, 2014. Citado na página 33.

THE CANCER GENOME ATLAS RESEARCH NETWORK. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med*, Mass Medical Soc, v. 2015, n. 372, p. 2481–2498, 2015. Citado 3 vezes nas páginas 26, 29, and 39.

TOYOTA, M. et al. CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 96, n. 15, p. 8681–8686, 1999. Citado na página 57.

TRIPATHI, S. et al. Meta-and orthogonal integration of influenza “omics” data defines a role for ubr4 in virus budding. *Cell host & microbe*, Elsevier, v. 18, n. 6, p. 723–735, 2015. Citado na página 67.

TURCAN, S. et al. Idh1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature*, Nature Research, v. 483, n. 7390, p. 479–483, 2012. Citado 6 vezes nas páginas 20, 26, 37, 46, 48, and 57.

VERHAAK, R. G. et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, Elsevier, v. 17, n. 1, p. 98–110, 2010. Citado 2 vezes nas páginas 26 and 45.

WANG, Q. et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer cell*, Elsevier, v. 32, n. 1, p. 42–56, 2017. Citado na página 26.

WHALEN, S.; TRUTY, R. M.; POLLARD, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature genetics*, Nature Publishing Group, v. 48, n. 5, p. 488, 2016. Citado na página 22.

WU, H.; ZHANG, Y. Mechanisms and functions of tet protein-mediated 5-methylcytosine oxidation. *Genes & development*, Cold Spring Harbor Lab, v. 25, n. 23, p. 2436–2452, 2011. Citado na página 22.

YAN, H. et al. Idh1 and idh2 mutations in gliomas. *New England Journal of Medicine*, Mass Medical Soc, v. 360, n. 8, p. 765–773, 2009. Citado na página 19.

YAN, W. et al. Hils1 is a spermatid-specific linker histone h1-like protein implicated in chromatin remodeling during mammalian spermiogenesis. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 100, n. 18, p. 10546–10551, 2003. Citado 2 vezes nas páginas 73 and 83.

ZHANG, Y. et al. Model-based analysis of chip-seq (macs). *Genome biology*, BioMed Central, v. 9, n. 9, p. R137, 2008. Citado na página 66.

ZHANG, Y. et al. Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron*, Elsevier, v. 89, n. 1, p. 37–53, 2016. Citado 2 vezes nas páginas 73 and 83.

APPENDIX A – Epigenetically regulated genes

Table 7 – Epigenetically regulated genes

EReg	CpG probe ID	Gene symbol	FDR
EReg1	cg00350296	CD248	0.0000000000133540043440993
	cg08361238	PTGFRN	0.0000000000000356210660291463
	cg03752628	PTGFRN	0.00000000000000498448793460387
	cg03922337	RRAS2	0.000000000070743554650448
	cg22892904	CBX2	0.0000000000223197272770447
	cg06454226	NDC80	0.000000000010338338537328
	cg07537523	BCAT1	0.00000000000710672476377348
	cg07598199	MYOZ2	0.000000000000204262799556049
	cg09604203	MYBL2	0.0000000000285168173036121
	cg10994126	PAPPA2	0.000000000000361778532371138
	cg22038738	PLAT	0.00000000000080045885390789
	cg12091331	PLAT	0.00000000000330904150176737
	cg14209518	NNMT	0.00000000000710672476377348
	cg25665697	LCTL	0.00000000000793302142740386
	cg15739581	GALNT3	0.0000000000014556323796812
EReg2	cg03694261	C11orf63	1.51040711579468e-59
	cg06630737	DRAXIN	1.67964363842944e-57
	cg21905630	GSX2	5.08613792017437e-57
	cg17686885	TOM1L1	2.77323003379022e-56
	cg20867633	GOLT1A	4.40010965535954e-56
	cg19965511	HCRT	1.21383356809506e-55
	cg22778981	ERBB2	3.82611923974791e-55
	cg27005179	ERBB2	3.82611923974791e-55
	cg16065186	ERBB2	9.12303870859647e-55
	cg14737977	ITGA3	1.84580903406971e-54
	cg18877506	PDPN	3.86289742058657e-54
	cg14865868	TRIB2	3.86289742058657e-54
Continued on next page			

Table 7 – continued from previous page

EReg	CpG probe ID	Gene symbol	FDR
	cg08260891	CTSA	1.30828567667321e-53
	cg17279839	RARRES2	1.53401674095458e-53
	cg20430816	FAM47E-STBD1	1.43222619885647e-52
EReg3	cg23376526	SLC27A3	1.38318474183323e-54
	cg07440414	TCF7	1.50879072056654e-53
	cg24719575	REST	5.77964516862353e-51
	cg27226949	EFNB2	9.67540917281903e-48
	cg19560758	ERRFI1	1.26067875120497e-47
	cg18071006	ULBP3	6.33914232335845e-36
	cg24576735	MARVELD1	5.77392605817606e-34
	cg01047414	PHYH	8.23614451099233e-33
	cg14795968	ACADL	9.21085446123328e-31
	cg05714479	LATS2	3.29908634882998e-30
	cg23833452	RAB32	8.00796129928111e-30
	cg20362772	PDLIM1	5.01561979939267e-29
	cg04388983	P2RY2	9.14277122616047e-29
	cg08164315	PTGER4	1.40365157263921e-28
	cg05060662	H2AFJ	1.91362247189223e-22
EReg4	cg26186727	NETO1	1.83319759798848e-38
	cg15239123	SMOC1	4.24227895365108e-36
	cg01817029	TRHDE	1.30871762336436e-35
	cg25192419	DOCK5	8.53744664618179e-34
	cg10644361	MIPOL1	1.79316578268758e-31
	cg00891541	SMPD3	4.95085388001245e-31
	cg11459714	KLK10	5.95588728887204e-30
	cg01049530	BMP3	1.07250262893881e-28
	cg00059225	GLRA1	3.40214174357111e-28
	cg08572611	ACTL6B	1.58537294647913e-26
	cg11051843	SEMA3C	5.54860376736528e-25
Continued on next page			

Table 7 – continued from previous page

EReg	CpG probe ID	Gene symbol	FDR
	cg01939681	GABRA6	9.66013614666015e-24
	cg25823578	PTGER2	3.78364605984119e-23
	cg21092462	KCNH1	9.46764893948347e-23
	cg01404615	DKK2	1.87328673878138e-21
EReg5	cg09088834	NINL	1.32276056213239e-29
	cg13878010	ADCY5	1.32276056213239e-29
	cg15439862	DSC3	1.89440665460972e-26
	cg09649610	GNG4	4.67671475009571e-26
	cg24251035	ZC3HAV1L	1.45735898988748e-25
	cg23555120	NUAK1	7.52579320331632e-25
	cg22040627	SLC13A5	9.72477659705395e-25
	cg23131007	ZNF280D	1.95066808424769e-23
	cg16313343	NOT AVAILABLE	3.82863183541272e-23
	cg24056567	PSD3	6.81142132034085e-22
	cg00489401	FLT4	6.92160782174642e-22
	cg10521852	LPAR2	8.73010537445412e-22

**ANNEX A – Exempt determination by the Ethics Committee of Ribeirão
Preto Medical School**

Faculdade de Medicina de Ribeirão Preto
Universidade de São Paulo

Ilma. Senhora Profa. Dra. Marcia Guimarães Villanova,
(Coordenadora do Comitê de Ética em Pesquisa do HCFMRP-USP)

Venho, por meio desta, solicitar dispensa de apreciação ética do projeto de pesquisa de Doutorado intitulado **“Análise Integrativa do Epigenoma de Gliomas com Fenótipo Metilador de Ilhas CpG(G-CIMP) Alto e Baixo: Caracterização e Desenvolvimento”** ao Comitê de Ética em Pesquisa do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo (HCFMRP-USP). A solicitação de dispensa justifica-se uma vez que esta pesquisa apresenta caráter retrospectivo e utiliza somente dados genômicos de amostras humanas que estão publicamente disponíveis à comunidade científica no banco de dados internacional *The Cancer Genome Atlas* ou TCGA (disponível em <https://portal.gdc.cancer.gov/>), com livre acesso a todos os usuários para baixar e analisar os dados sem necessidade de cadastro ou identificação prévios. Este projeto de Doutorado será desenvolvido no Departamento de Genética da FMRP-USP pela aluna de Pós-Graduação Thais Sarraf Sabedot e sob orientação do Professor da Faculdade de Medicina de Ribeirão Preto Dr. Houtan Noushmehr. As amostras estão de acordo com a ética e a legislação norte-americana vigente, tendo em vista que o consórcio TCGA, responsável pelo banco de dados, mantém em sigilo todas as informações que possam levar à identificação dos pacientes. O projeto constituirá de 932 amostras de tumores cerebrais, já publicadas na literatura (Ceccarelli et al., 2016; Noushmehr et al., 2010; Brennan et al., 2013; The Cancer Genome Atlas Research Network, 2015), incluindo dados clínicos e moleculares com ênfase para metilação de DNA e expressão gênica.

Ribeirão Preto, 16 de janeiro de 2018.

Thais Sarraf Sabedot

Não há necessidade de submissão ao
Comitê de Ética em Pesquisa:

Dra. Marcia Guimarães Villanova
Coordenadora do Comitê de Ética em Pesquisa
do HC e FMRP-USP
05/02/2018

ANNEX B – Glioma gene expression subtypes

TCGA diffuse glioma gene expression subtypes were identified by Dr. Michele Ceccarelli, Dr. Stefano M. Pagnotta and Dr. Antonio Iavarone using 667 RNA-seq profiles (513 LGG and 154 GBM) on 2,275 genes as described (CECCARELLI et al., 2016).

Figure 28 – Glioma transcriptome subtypes

ANNEX C – Publication

Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma

Graphical Abstract

Authors

Michele Ceccarelli, Floris P. Barthel, Tathiane M. Malta, ..., Houtan Noushmehr, Antonio Iavarone, Roel G.W. Verhaak

Correspondence

houtan@usp.br (H.N.),
ai2102@columbia.edu (A.I.),
rverhaak@mdanderson.org (R.G.W.V.)

In Brief

Integration of a large sample size of glioma tumors with multidimensional 'omic characterization and clinical annotation provides insights into molecular classification, telomere maintenance mechanisms, progression from low to high grade disease, driver mutations, and therapeutic options.

Highlights

- Comprehensive molecular profiling of 1,122 adult diffuse grade II, III, and IV gliomas
- Telomere length and telomere maintenance defined by somatic alterations
- DNA methylation profiling reveals subtypes of IDH mutant and IDH-wild-type glioma
- Integrated molecular analysis of progression from low-grade to high-grade disease

Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma

Michele Ceccarelli,^{1,2,24} Floris P. Barthel,^{3,4,24} Tathiane M. Malta,^{5,6,24} Thais S. Sabedot,^{5,6,24} Sofie R. Salama,⁷ Bradley A. Murray,⁸ Olena Morozova,⁷ Yulia Newton,⁷ Amie Radenbaugh,⁷ Stefano M. Pagnotta,^{2,9} Samreen Anjum,¹ Jiguang Wang,¹⁰ Ganiraju Manyam,³ Pietro Zoppoli,¹⁰ Shiyun Ling,³ Arjun A. Rao,⁷ Mia Grifford,⁷ Andrew D. Cherniack,⁸ Hailei Zhang,⁸ Laila Poisson,¹¹ Carlos Gilberto Carlotti, Jr.,^{5,6} Daniela Pretti da Cunha Tirapelli,^{5,6} Arvind Rao,³ Tom Mikkelsen,¹¹ Ching C. Lau,^{12,13} W.K. Alfred Yung,³ Raul Rabadan,¹⁰ Jason Huse,¹⁴ Daniel J. Brat,¹⁵ Norman L. Lehman,¹⁶ Jill S. Barnholtz-Sloan,¹⁷ Siyuan Zheng,³ Kenneth Hess,³ Ganesh Rao,³ Matthew Meyerson,^{8,18} Rameen Beroukhi,^{8,18,19} Lee Cooper,¹⁵ Rehan Akbani,³ Margaret Wrensch,²⁰ David Haussler,⁷ Kenneth D. Aldape,²¹ Peter W. Laird,²² David H. Gutmann,²³ TCGA Research Network, Houtan Noushmehr,^{5,6,25,*} Antonio Iavarone,^{10,25,*} and Roel G.W. Verhaak^{3,25,*}

¹Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha P.O. box 5825, Qatar

²Department of Science and Technology, University of Sannio, Benevento 82100, Italy

³Department of Genomic Medicine, Department of Bioinformatics and Computational Biology, Department of Biostatistics, Department of Neuro-Oncology, Department of Neurosurgery, Department of Pathology, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

⁴Oncology Graduate School Amsterdam, Department of Pathology, VU University Medical Center, 1081 HV Amsterdam, the Netherlands

⁵Department of Genetics (CISBi/NAP), Department of Surgery and Anatomy, Ribeirão Preto Medical School, University of São Paulo, Monte Alegre, Ribeirão Preto-SP CEP: 14049-900, Brazil

⁶Center for Integrative Systems Biology (CISBi, NAP/USP), Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, São Paulo 14049-900, Brazil

⁷UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

⁸The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA

⁹BIOGEM Istituto di Ricerche Genetiche "G. Salvatore," Campo Reale, 83031 Ariano Irpino, Italy

¹⁰Department of Neurology, Department of Pathology, Institute for Cancer Genetics, Department of Systems Biology and Biomedical Informatics, Columbia University Medical Center, New York, NY 10032, USA

¹¹Henry Ford Hospital, Detroit, MI 48202, USA

¹²Texas Children's Hospital, Houston, TX 77030, USA

¹³Baylor College of Medicine, Houston, TX 77030, USA

¹⁴Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

¹⁵Winship Cancer Institute, Emory University, Atlanta, GA 30322, USA

¹⁶Department of Pathology, The Ohio State University, Columbus, OH 43210, USA

¹⁷Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH 44106, USA

¹⁸Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

¹⁹Department of Medicine, Harvard Medical School, Boston, MA 02215, USA

²⁰Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA 94158, USA

²¹Princess Margaret Cancer Centre, Toronto, ON M5G 2M9, Canada

²²Van Andel Research Institute, Grand Rapids, MI 49503, USA

²³School of Medicine, Washington University, St. Louis, MO 63110, USA

²⁴Co-first author

²⁵Co-senior author

*Correspondence: houtan@usp.br (H.N.), ai2102@columbia.edu (A.I.), rverhaak@mdanderson.org (R.G.W.V.)
<http://dx.doi.org/10.1016/j.cell.2015.12.028>

SUMMARY

Therapy development for adult diffuse glioma is hindered by incomplete knowledge of somatic glioma driving alterations and suboptimal disease classification. We defined the complete set of genes associated with 1,122 diffuse grade II-III-IV gliomas from The Cancer Genome Atlas and used molecular profiles to improve disease classification, identify molecular correlations, and provide insights into

the progression from low- to high-grade disease. Whole-genome sequencing data analysis determined that *ATRX* but not *TERT* promoter mutations are associated with increased telomere length. Recent advances in glioma classification based on *IDH* mutation and 1p/19q co-deletion status were recapitulated through analysis of DNA methylation profiles, which identified clinically relevant molecular subsets. A subtype of *IDH* mutant glioma was associated with DNA demethylation and poor outcome;

a group of IDH-wild-type diffuse glioma showed molecular similarity to pilocytic astrocytoma and relatively favorable survival. Understanding of cohesive disease groups may aid improved clinical outcomes.

INTRODUCTION

Diffuse gliomas represent 80% of malignant brain tumors (Schwartzbaum et al., 2006). Adult diffuse gliomas are classified and graded according to histological criteria (oligodendroglioma, oligoastrocytoma, astrocytoma, and glioblastoma; grade II to IV). Although histopathologic classification is well established and is the basis of the World Health Organization (WHO) classification of CNS tumors (Louis et al., 2007), it suffers from high intra- and inter-observer variability, particularly among grade II-III tumors (van den Bent, 2010). Recent molecular characterization studies have benefited from the availability of the datasets generated by The Cancer Genome Atlas (TCGA) (Brennan et al., 2013; Eckel-Passow et al., 2015; Frattini et al., 2013; Kim et al., 2015; Suzuki et al., 2015; Cancer Genome Atlas Research Network et al., 2015) and have related genetic, gene expression, and DNA methylation signatures with prognosis (Noushmehr et al., 2010; Sturm et al., 2012; Verhaak et al., 2010). For example, mutations in the isocitrate dehydrogenase genes 1 and 2 (*IDH1/IDH2*) define a distinct subset of glioblastoma (GBM) with a hypermethylation phenotype (G-CIMP) with favorable outcome (Noushmehr et al., 2010; Yan et al., 2009). Conversely, the absence of *IDH* mutations in LGG marks a distinct IDH-wild-type subgroup characterized by poor, GBM-like prognosis (Eckel-Passow et al., 2015; Cancer Genome Atlas Research Network et al., 2015). Recent work by us and others has proposed classification of glioma into *IDH* wild-type cases, *IDH* mutant group additionally carrying codeletion of chromosome arm 1p and 19q (*IDH* mutant-codel) and samples with euploid 1p/19q (*IDH* mutant-non-codel), regardless of grade and histology (Eckel-Passow et al., 2015; Cancer Genome Atlas Research Network et al., 2015). Mutation of the *TERT* promoter, which has been reported with high frequency across glioma, may be an additional defining feature. Current analyses have not yet clarified the relationships between LGGs and GBMs that share common genetic hallmarks like *IDH* mutation or *TERT* promoter mutation status. An improved understanding of these relationships will be necessary as we evolve toward an objective genome-based clinical classification.

To address the above issues, we assembled a dataset comprising all TCGA newly diagnosed diffuse glioma consisting of 1,122 patients and comprehensively analyzed using sequencing and array-based molecular profiling approaches. We have addressed crucial technical challenges in analyzing this comprehensive dataset, including the integration of multiple platforms and data sources (e.g., multiple methylation and gene expression platforms). We identified new diffuse glioma subgroups with distinct molecular and clinical features and shed light on the mechanisms driving progression of lower grade glioma (LGG) (WHO grades II and III) into full-blown GBM (WHO grade IV).

RESULTS

Patient Cohort Characteristics

The TCGA LGG and GBM cohorts consist of 516 and 606 patients, respectively. Independent analysis of the GBM dataset was previously described, as was analysis of 290 LGG samples (Brennan et al., 2013; Cancer Genome Atlas Research Network et al., 2015). 226 LGG samples were added to our current cohort (Table 1). Clinical data, including age, tumor grade, tumor histology, and survival, were available for 93% (1,046/1,122) of cases (Table S1). The majority of samples were grade IV tumors (n = 590, 56%), whereas 216 (21%) and 241 (23%) were grade II and III tumors, respectively. Similarly, 590 (56%) samples were classified as GBM, 174 (17%) as oligodendroglioma, 169 (16%) as astrocytoma, and 114 (11%) as oligoastrocytoma.

Among the data sources considered in our analysis were gene expression (n = 1,045), DNA copy number (n = 1,084), DNA methylation (n = 932), exome sequencing (n = 820), and protein expression (n = 473). Multiple and overlapping characterization assays were employed (Table S1). All data files that were used in our analysis can be found at https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015/.

Identification of Novel Glioma-Associated Genomic Alterations

To establish the set of genomic alterations that drive gliomagenesis, we called point mutations and indels on the exomes of 513 LGG and 307 GBM using the Mutect, Indelocator, Varscan2, and RADIA algorithms and considered all mutations identified by at least two callers. Significantly mutated genes (SMGs) were determined using MutSigCV. This led to the identification of 75 SMGs, 10 of which had been previously reported in GBM (Brennan et al., 2013), 12 of which had been reported in LGG (Cancer Genome Atlas Research Network et al., 2015), and 8 of which had been identified in both GBM and LGG studies. 45 SMGs have not been previously associated with glioma and ranged in mutation frequency from 0.5% to 2.6% (Table S2A). We used GISTIC2 to analyze the DNA copy number profiles of 1,084 samples, including 513 LGG and 571 GBM, and identified 162 significantly altered DNA copy number segments (Table S2B). We employed PRADA and deFuse to detect 1,144 gene fusion events in the RNA-seq profiles available for 154 GBM and 513 LGG samples, of which 37 in-frame fusions involved receptor tyrosine kinases (Table S2C). Collectively, these analyses recovered all known glioma driving events, including in *IDH1* (n = 457), *TP53* (n = 328), *ATRX* (n = 220), *EGFR* (n = 314), *PTEN* (n = 168), *CIC* (n = 80), and *FUBP1* (n = 45). Notable newly predicted glioma drivers relative to the earlier TCGA analyses were genes associated with chromatin organization such as *SETD2* (n = 24), *ARID2* (n = 20), *DNMT3A* (n = 11), and the *KRAS/NRAS* oncogenes (n = 25 and n = 5, respectively).

We overlapped copy number, mutation (n = 793), and fusion transcript (n = 649) profiles and confirmed the convergence of genetic drivers of glioma into pathways, including the Ras-Raf-MEK-ERK, p53/apoptosis, PI3K/AKT/mTOR, chromatin modification, and cell cycle pathways. The Ras-Raf-MEK-ERK signaling cascade showed alterations in 106 of 119 members

Table 1. Clinical Characteristics of the Sample Set Arranged by IDH and 1p/19q Co-deletion Status

Feature	IDH Wt (n = 520)	IDH mut - non-codel (n = 283)	IDH mut - codel (n = 171)	Unknown (n = 148)
Clinical				
Histology (n)				
Astrocytoma	52 (10.0%)	112 (39.6%)	4 (2.3%)	1 (0.7%)
Glioblastoma	419 (80.6%)	32 (11.3%)	2 (1.2%)	137 (92.6%)
Oligoastrocytoma	15 (2.9%)	69 (24.4%)	30 (17.5%)	0 (0%)
Oligodendroglioma	19 (3.7%)	37 (13.1%)	117 (68.4%)	1 (0.7%)
Unknown	15 (2.9%)	33 (11.7%)	18 (10.5%)	9 (6.1%)
Grade (n)				
G2	19 (3.7%)	114 (40.3%)	81 (47.4%)	2 (1.4%)
G3	67 (12.9%)	104 (36.7%)	70 (40.9%)	0 (0%)
G4	419 (80.6%)	32 (11.3%)	2 (1.2%)	137 (92.6%)
Unknown	15 (2.9%)	33 (11.7%)	18 (10.5%)	9 (6.1%)
Age				
Median (LQ-UQ)	59 (51–68)	38 (30–44)	46 (35–54)	55 (48–68)
Unknown (n)	16	33	18	9
Survival				
Median (CI)	14.0 (12.6–15.3)	75.1 (62.1–94.5)	115.8 (90.5–Inf)	12.6 (11.3–14.9)
Unknown (n)	14	32	18	12
KPS				
<70	85 (16.3%)	8 (2.8%)	5 (2.9%)	21 (14.2%)
70–80	196 (37.7%)	41 (14.5%)	18 (10.5%)	60 (40.5%)
90	29 (5.6%)	60 (21.2%)	32 (18.7%)	2 (1.4%)
100	51 (9.8%)	44 (15.9%)	30 (17.5%)	14 (9.5%)
Unknown	159 (30.6%)	129 (45.6%)	86 (50.3%)	51 (34.5%)
Molecular				
MGMT promoter				
Methylated	170 (32.7%)	242 (85.5%)	169 (98.8%)	32 (21.6%)
Unmethylated	248 (47.7%)	36 (12.7%)	1 (0.6%)	34 (23.0%)
Unknown	102 (19.6%)	5 (1.8%)	1 (0.6%)	82 (55.4%)
TERT promoter				
Mutant	67 (12.9%)	8 (2.8%)	86 (50.3%)	1 (0.7%)
Wild-type	19 (9.8%)	146 (51.6%)	2 (1.2%)	0 (0%)
Unknown	434 (83.5%)	129 (45.6%)	83 (48.5%)	135 (99.3%)
TERT expression				
Expressed	178 (34.2%)	14 (4.9%)	153 (89.5%)	6 (4.1%)
Not expressed	51 (9.8%)	242 (85.5%)	16 (9.4%)	7 (4.7%)
Unknown	291 (56.0%)	27 (9.5%)	2 (1.2%)	135 (91.2%)

detected across 578 cases (73%), mostly occurring in IDH-wild-type samples (n = 327 of 357, 92%). Conversely, we found that a set of 36 genes involved in chromatin modification was targeted by genetic alterations in 423 tumors (54%, n = 36 genes), most of which belonged to the IDH mutant-non-codel group (n = 230, 87%).

In order to identify new somatically altered glioma genes, we used MutComFocal to nominate candidates altered by mutation, as well as copy number alteration. Prominent among these genes was *NIPBL*, a crucial adherin subunit that is essential for loading cohesins on chromatin (Table S2D) (Peters and Nish-

iyama, 2012). The cohesin complex is responsible for the adhesion of sister chromatids following DNA replication and is essential to prevent premature chromatid separation and faithful chromosome segregation during mitosis (Peters and Nishiyama, 2012). Alterations in the cohesin pathway have been reported in 12% of acute myeloid leukemias (Kon et al., 2013). Mutations of the cohesin complex gene *STAG2* had been previously reported in GBM (Brennan et al., 2013). Taken together, 16% of the LGG/GBM showed mutations and/or CNAs in multiple genes involved in the cohesin complex, thus nominating this process as a prominent pathway involved in gliomagenesis.

Figure 1. Telomere Length Associations in Glioma

(A) Heatmap of relative tumor/normal telomere lengths of 119 gliomas, grouped by *TERTp* and *ATRX* mutation status.

(B) Telomere length decreases with increasing age (measured in years at diagnosis) in blood normal control samples (n = 137).

(C) Quantitative telomere length estimates of tumors and blood normal, grouped by *TERTp* mutant (n = 67, 56%), *ATRX* mutant (n = 40, 33%), and double negative (n = 13, 11%) status. *** = p < 0.0001; ** = p < 0.001.

To correlate *TERTp* mutations to telomere length, we used whole-genome sequencing and low pass whole-genome sequencing data to estimate telomere length in 141 pairs of matched tumor and normal samples. As expected, we observed an inverse correlation of telomere length with age at diagnosis in matching blood normal samples (Figure 1B) and tumor samples (Figure S1C). Glioma samples harboring *ATRX* mutations showed significantly longer telomeres compared to *TERTp* mutant samples (t test p value < 0.0001; Figure 1C).

Among *TERTp* mutation gliomas, there was no difference in telomere length between samples with and without additional *IDH1/IDH2* mutations, despite a difference in age. *ATRX* forms a complex with DAXX and H3.3, and the genes encoding these proteins are frequently mutated in pediatric gliomas (Sturm et al., 2012). Mutations in DAXX and H3F3A were identified in only two samples in our WGS dataset. The *ATRX*-DAXX-H3.3 complex is associated with the alternative lengthening of telomeres (ALT) and our observations confirm previously hypothesized fundamental differences between the telomere control exerted by telomerase and ALT (Sturm et al., 2014).

As demonstrated by the identification of *TERTp* mutations, somatic variants affecting regulatory regions may play a role in gliomagenesis. Using 67 matched whole-genome and RNA-seq expression pairs, we similarly sought to identify mutations located within 2 kb upstream of transcription start sites and associated with a gene expression change. Using strict filtering methods, we identified 12 promoter regions with mutations in at least 6 samples. Three of 12 regions related to a significant difference in the expression of the associated gene expression, suggesting possible functional consequences. Other than *TERT* (n = 37), promoter mutations of the ubiquitin ligase *TRIM28* (n = 8) and the calcium channel gamma subunit *CACNG6* (n = 7) correlated with respectively upregulation and downregulation of these genes, respectively (Table S2E). *TRIM28* has been reported to mediate the ubiquitin-dependent degradation of AMP-activated protein kinase (AMPK) leading to activation of mTOR signaling and hypersensitization to AMPK agonists, such as metformin (Pineda et al., 2015).

Telomere Length Is Positively Correlated with *ATRX*, but Not *TERT* Promoter Mutations

Mutations in the *TERT* promoter (*TERTp*) have been reported in 80% of GBM (Killela et al., 2013). We used *TERTp* mutation calls from targeted sequencing (n = 287) and complemented them with *TERTp* mutations inferred from whole-genome sequencing (WGS) data (n = 42). *TERTp* mutations are nearly mutually exclusive with mutations in *ATRX* (Eckel-Passow et al., 2015), which was confirmed in our cohort. Overall, 85% of diffuse gliomas harbored mutations of *TERTp* (n = 157, 48%) or *ATRX* (n = 120, 37%). *TERTp* mutations activate *TERT* mRNA expression through the creation of a de novo E26 transformation-specific (ETS) transcription factor-binding site (Horn et al., 2013), and we observed significant *TERT* upregulation in *TERTp* mutant cases (p value < 0.0001, Figure S1A). *TERT* expression measured by RNA-seq was a highly sensitive (91%) and specific (95%) surrogate for the presence of *TERTp* mutation (Figure S1B). We correlated *TERTp* status with glioma driving alterations and observed that nearly all *IDH*-wild-type cases with chromosome 7 gain and chromosome 10 loss harbored *TERTp* mutations or upregulated *TERT* expression (n = 52/53 and n = 134/147, respectively; Figure 1A). Conversely, only 45% of *IDH*-wild-type samples lacking chromosome 7/chromosome 10 events showed *TERTp* mutations or elevated *TERT* expression (n = 15/33 and n = 43/82, respectively). Thus, *TERTp* mutations may precede the chr 7/chr 10 alterations that have been implicated in glioma initiation (Ozawa et al., 2014).

Figure 2. Pan-glioma DNA Methylation and Transcriptome Subtypes

(A) Heatmap of DNA methylation data. Columns represent 932 TCGA glioma samples grouped according to unsupervised cluster analysis; rows represent DNA methylation probes sorted by hierarchical clustering. Non-neoplastic samples are represented on the left of the heatmap ($n = 77$) (Guintivano et al., 2013). (B) Heatmap of RNA sequencing data. Unsupervised clustering analysis for 667 TCGA glioma samples profiled using RNA sequencing are plotted in the heatmap using 2,275 most variant genes. Previously published subtypes were derived from Brennan et al. (2013) and Cancer Genome Atlas Research Network et al., 2015. (C) Tumor Map based on mRNA expression and DNA methylation data. Each data point is a TCGA sample colored coded according to their identified status. A live interactive version of this map is available at <http://tumormap.ucsc.edu/?p=ynewton.gliomas-paper>.

Unsupervised Clustering of Gliomas Identifies Six Methylation Groups and Four RNA Expression Groups Associated with IDH Status

To segregate the DNA methylation subtypes across the pan-glioma dataset, we analyzed 932 glioma samples profiled on the HumanMethylation450 platform (516 LGG and 129 GBM) and the HumanMethylation27 platform (287 GBM). In order to incorporate the maximum number of samples, we merged datasets from both methylation platforms yielding a core set of 25,978 CpG probes. To reduce computational requirements to cluster this large dataset, we eliminated sites that were methylated (mean β value ≥ 0.3) in non-tumor brain tissues and selected 1,300 tumor-specific methylated probes (1,300/25,978, 5%) to perform unsupervised k-means consensus clustering. This identified six distinct clusters, labeled LGM1–6 (Figure 2A and Tables S1 and S3A). Next, we sought to determine pan-glioma expression subtypes through unsupervised clustering analysis of 667 RNA-seq profiles (513 LGG and 154 GBM), which resulted in four main clusters labeled LGr1–4 (Figure 2B and Tables S1 and S3A). An additional 378 GBM samples with Affymetrix HT-HG-U133A profiles (but lacking RNA-seq data) were classified into the four clusters using a k-nearest neighbor classification procedure. IDH mutation status was the primary driver of methylation and transcriptome clustering and separated the cohort into two macro-groups. The LGM1/LGM2/LGM3 DNA methylation macro-group carried IDH1 or IDH2 mutations (449 of 450, 99%) and was enriched for LGG (421/454, 93%) while LGM4/

LGM5/LGM6 were IDH-wild-type (429/430, 99%) and enriched for GBM (383/478, 80%). LGM1–3 showed genome-wide hypermethylation compared to LGM4–6 clusters (Figure S2A), documenting the association between IDH mutation and increased DNA methylation (Noushmehr et al., 2010; Turcan et al., 2012). Principal component analysis using 19,520 probes yielded similar results, thus emphasizing that our probe selection method did not introduce unwanted bias (Figure S2B). The gene expression clusters LGr1–3 harbored IDH1 or IDH2 mutations (438 of 533, 82%) and were enriched for LGG (436/563, 77%), while the LGr4 was exclusively IDH-wild-type (376 of 387, 97%) and enriched for GBM (399/476, 84%).

We extended our analysis using Tumor Map (Supplemental Experimental Procedures) to perform integrated co-clustering analysis of the combined gene expression ($n = 1,196$) and DNA methylation ($n = 867$) profiles. An interactive Tumor Map is publicly available at <http://tumormap.ucsc.edu/?p=ynewton.gliomas-paper>. Tumor Map assigns samples to a hexagon in a grid so that nearby samples are likely to have similar genomic profiles and allows visualizing complex relationships between heterogeneous genomic data samples and their clinical or phenotypical associations. Thus, clusters in the map indicate groups of samples with high similarity of integrated gene expression and DNA methylation profiles (Figure 2C). The map confirms clustering by IDH status and additionally shows islands of samples that share previously reported GBM cluster memberships (Noushmehr et al., 2010; Verhaak et al., 2010). To assess

(legend on next page)

clustering sensitivity to pre-processing, we tried complementary methods and obtained similar results (Figure S2C).

To identify genes whose copy number changes are associated with concordant changes in gene expression, we combined expression and copy number profiles from 659 samples to define a signature of 57 genes with strong functional copy number (fCN) change (Table S3B). The fCN signature clustered gliomas into three macro-clusters, LGfc1–3, strongly associated with IDH and 1p/19q status (Figure S2D). The fCN analysis revealed the functional activation of a cluster of *HOXA* genes in the IDH-wild-type LGfc2 cluster, which were previously associated with glioma stem cell maintenance (Kurscheid et al., 2015).

Finally, we clustered reverse phase protein array profiles, consisting of 196 antibodies on 473 samples. Two macro clusters were observed, and in contrast to the transcriptome/methylome/fCNV clustering, the primary discriminator was based on glioma grade (LGG versus GBM) rather than IDH status (Figure S2E). Compared to the LGG-like cluster, the GBM-like cluster had elevated expression of IGFBP2, fibronectin, PAI1, HSP70, EGFR, phosphoEGFR, phosphoAKT, Cyclin B1, Caveolin, Collagen VI, Annexin1, and ASNS, whereas the LGG class showed increased activity of PKC (alpha, beta, and delta), PTEN, BRAF, and phosphoP70S6K.

The above results confirm IDH status as the major determinant of the molecular footprints of diffuse glioma. To further elucidate the subtypes of diffuse glioma, we performed unsupervised clustering within each of the two IDH-driven macroclusters. We used 1,308 tumor-specific CpG probes defined among the IDH mutation cohort ($n = 450$) and identified three IDH mutant-specific DNA methylation clusters (Figure S3A). Using 914 tumor-specific CpG probes in the IDH-wild-type cohort ($n = 430$), we uncovered three IDH-wild-type-specific clusters (Figure S4A). The sets of CpG probes used to cluster each of the two IDH-driven datasets overlapped significantly with the 1,300 probes that defined the pan-glioma DNA methylation clustering (1162/1,300, 89% and 853/1,300, 66%, for IDH mutant and IDH-wild-type, respectively). The clusters identified by separating IDH mutant and IDH-wild-type gliomas showed strong overall concordance with pan-glioma DNA methylation subtypes (Table S3A). Similarly, unsupervised clustering of 426 IDH mutant RNA-seq profiles resulted in three subtypes (Figure S3A), and analysis of the 234 IDH-wild-type samples led to four mixed LGG/GBM clusters that showed enrichment for previously identified GBM expression subtypes (Figure S4C) (Verhaak et al., 2010).

An Epigenetic Signature Associated with Activation of Cell Cycle Genes Segregates a Subgroup of IDH Mutant LGG and GBM with Unfavorable Clinical Outcome

The three epigenetic subtypes defined by clustering IDH mutant glioma separated samples harboring the 1p/19q co-deletion into a single cluster and non-codel glioma into two clusters (Figure S3A). Conversely, non-codel glioma grouped nearly exclusively into a single expression cluster, and codels were split in two separated expression clusters (Figure S3A). A distinct subgroup of samples within the IDH mutant-non-codel DNA methylation clusters manifested relatively reduced DNA methylation (Figure S3B). The unsupervised clustering of IDH mutant glioma was unable to segregate the lower methylated non-codel subgroup as the 1,308 probes selected for unsupervised clustering included only 19 of the 131 differentially methylated probes characteristic for this subgroup ($FDR < 10^{-15}$, difference in mean methylation beta value > 0.27). The low-methylation subgroup consisted of both G-CIMP GBM (13/25) and LGGs (12/25) and was confirmed using a non-TCGA dataset (Figure S3C). The tumors with higher methylation in the split cluster were very similar to those grouped in the second non-codel cluster, and a supervised comparison identified only 12 probes as differentially DNA methylated (Figures 3A and 3B). We concluded that IDH mutant glioma is composed of three coherent subgroups: (1) the Codel group, consisting of IDH mutant-codel LGGs; (2) the G-CIMP-low group, including IDH mutant-non-codel glioma (LGG and GBM) manifesting relatively low genome-wide DNA methylation; and (3) the G-CIMP-high group, including IDH mutant-non-codel glioma (LGG and GBM) with higher global levels of DNA methylation. The newly identified G-CIMP-low group of glioma was associated with significantly worse survival as compared to the G-CIMP-high and Codel groups (Figure S3D). The clinical outcome of the tumors classified as G-CIMP-high was as favorable as that of Codel tumors, the subgroup generally thought to have the best prognosis among glioma patients (Figures 3C and S3D). We compared the frequencies of glioma driver gene alterations between the three types of IDH mutant glioma and found that 15 of 18 G-CIMP-low cases carried abnormalities in cell cycle pathway genes such as *CDK4* and *CDKN2A*, relative to 36/241 and 2/172 for G-CIMP-high and Codels, respectively (Figure 3D). Supervised analysis between gene expression of G-CIMP-low and G-CIMP-high resulted in 943 differentially expressed genes. We mapped the 943 deregulated genes to 767 nearest CpG probes (max distance 1 kb) and found the majority

Figure 3. Identification of a Distinct G-CIMP Subtype Defined by Epigenomics

- (A) Heatmap of probes differentially methylated between the two IDH mutant-non-codel DNA methylation clusters allowed the identification of a low-methylation subgroup named G-CIMP-low. Non-tumor brain samples ($n = 12$) are represented on the left of the heatmap.
- (B) Heatmap of genes differentially expressed between the two IDH mutant-non-codel DNA methylation clusters.
- (C) Kaplan-Meier survival curves of IDH mutant methylation subtypes. Ticks represent censored values.
- (D) Distribution of genomic alterations in genes frequently altered in IDH mutant glioma.
- (E) Genomic distribution of 633 CpG probes differentially demethylated between co-clustered G-CIMP-low and G-CIMP-high. CpG probes are grouped by UCSC genome browser-defined CpG Islands, shores flanking CpG island ± 2 kb and open seas (regions not in CpG islands or shores).
- (F) DNA methylation heatmap of TCGA glioma samples ordered per Figure 2A and the epigenetically regulated (EReg) gene signatures defined for G-CIMP-low, G-CIMP-high, and Codel subtypes. The mean RNA sequencing counts for each gene matched to the promoter of the identified cgID across each cluster are plotted to the right.
- (G) Heatmap of the validation set classified using the random forest method applying the 1,300 probes defined in Figure 2A.
- (H) Heatmap of probes differentially methylated between G-CIMP-low and G-CIMP-high in longitudinally matched tumor samples.

of the CpG probes (486/767, 63%) to show a significant methylation difference (FDR < 0.05, difference in mean methylation beta value > 0.01) between G-CIMP-low and G-CIMP-high, suggesting a mechanistic relation between loss of methylation and increased transcript levels.

Recent analysis of epigenetic profiles derived from colon cancers showed that transcription factors may bind to regions of demethylated DNA (Berman et al., 2012). Therefore, we asked whether transcription factors may be recruited to the DNA regions differentially methylated between G-CIMP-low samples and G-CIMP-high samples from the same methylation cluster, using 450K methylation profiles ($n = 39$). Globally, we detected 643 differentially methylated probes between 27 G-CIMP-low and 12 G-CIMP-high samples (absolute diff-mean difference ≥ 0.25 , FDR $\leq 5\%$). Most of these probes (69%) were located outside of any known CpG island but positioned within intergenic regions known as open seas (Figure 3E). This represents a 2.5-fold open sea enrichment compared to the expected genome-wide distribution of 450K CpG probes (chi-square p value < 2.2×10^{-16}). We also observed a 3.4-fold depletion within CpG islands (chi-square p value < 2.2×10^{-16}).

Using this set of intergenic CpG probes, we asked whether a DNA motif signature associated with distal regulatory elements. Such a pattern would point to candidate transcription factors involved in tumorigenesis of the G-CIMP-low group. A de novo motif scan and known motif scan identified a distinct motif signature TGTT (geometric test p value = 10^{-11} , fold enrichment = 1.8), known to be associated with the OLIG2 and SOX transcription factor families (Figure 3E) (Lodato et al., 2013). This observation was corroborated by the higher expression levels of SOX2, as well as 17 out of 20 other known SOX family members in G-CIMP-low compared to G-CIMP-high (fold difference > 2). The primary function of SOX2 in the nervous system is to promote self-renewal of neural stem cells and, within brain tumors, the glioma stem cell state (Graham et al., 2003). Interestingly, SOX2 and OLIG2 have been described as neurodevelopmental transcription factors being essential for GBM propagation (Suvà et al., 2014). Supervised gene expression pathway analysis of the genes activated in the G-CIMP-low group as opposed to G-CIMP-high group revealed activation of genes involved in cell cycle and cell division consistent with the role of SOX in promoting cell proliferation (Figure S3E). The enrichment in cell cycle gene expression provides additional support to the notion that development of the G-CIMP-low subtype is associated with activation of cell cycle progression and may be mediated by a loss of CpG methylation and binding of SOX factors to candidate genomic enhancer elements.

To validate the G-CIMP-low, G-CIMP-high, and Codel IDH mutant subtypes, we compiled a validation cohort from published studies, including 324 adult and pediatric gliomas (Lambert et al., 2013; Mur et al., 2013; Sturm et al., 2012; Turcan et al., 2012). The CpG probe methylation signatures used to classify the validation set are provided on the publication portal accompanying this publication (https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015/). Among them, 103 were identified as IDH mutant on the basis of their genome-wide DNA methylation profile. We classified samples in the validation set using the probes that defined the IDH mutant-specific DNA

methylation cluster analysis integrated in a supervised random forest method. The analysis recapitulated the clusters generated from the TCGA collection (Figure S3C). In order to determine epigenetically regulated (EReg) genes that may be characteristic of the biology of the IDH mutant diffuse glioma subtypes, we compared 450k methylation DNA methylation profiles and gene expression levels between 636 IDH mutant and IDH-wild-type gliomas and 110 non-tumor samples from 11 different tissue types. From the list of epigenetically regulated genes, we extracted 263 genes that were grouped into EReg gene signatures, which showed differential signals among the three IDH mutant subtypes (Figure 3F). These trends were confirmed in the validation set (Figure 3G).

We investigated the possibility that the G-CIMP-high group is a predecessor to the G-CIMP-low group by comparing the DNA methylation profiles from ten IDH mutant-non-codel LGG and GBM primary-recurrent cases with the TCGA cohort. We evaluated the DNA methylation status of probes identified as differentially methylated ($n = 90$) between G-CIMP-low and G-CIMP-high (FDR < 10^{-13} , difference in mean methylation beta-value > 0.3 and < -0.4). Four out of ten IDHmut-non-codel cases showed a demethylation pattern after disease recurrence, while partial demethylation was demonstrated in the remaining six recurrences, supporting the notion of a progression from G-CIMP-high to G-CIMP-low phenotype (Figure 3H).

An IDH-Wild-Type Subgroup of Histologically Defined Diffuse Glioma Is Associated with Favorable Survival and Shares Epigenomic and Genomic Features with Pilocytic Astrocytoma

IDH-wild-type gliomas segregated into three DNA methylation clusters (Figure S4A). The first is enriched with tumors belonging to the classical gene expression signature and was labeled Classic-like, whereas the second group, enriched with mesenchymal subtype tumors, was labeled Mesenchymal-like (Table S1) (Verhaak et al., 2010). The third cluster contained a larger fraction of LGG in comparison to the other IDH-wild-type clusters. We observed that the IDH-wild-type LGGs but not the IDH-wild-type GBM in this cluster displayed markedly longer survival (log-rank p value = 3.6×10^{-5} ; Figure 4A) and occurred in younger patients (mean 37.6 years versus 50.8 years, t test p value = 0.002). Supervised analysis of differential methylation between LGG and GBM in the third DNA methylation cluster did not reveal any significant probes despite significant differences in stromal content (p value < 0.005; Figure S4D), suggesting that this group cannot be further separated using CpG methylation markers.

Next, we sought to validate the methylation-based classification of IDH-wild-type glioma in an independent cohort of 221 predicted IDH-wild-type glioma samples, including 61 grade I pilocytic astrocytomas (PAs). Toward this aim, we used a supervised random forest model built with the probes that defined the IDH-wild-type clusters. Samples classified as Mesenchymal-like showed enrichment for the Sturm et al. (2012) Mesenchymal subtype (29/88), and gliomas predicted as Classic-like were all RTK II "Classic" (22/22), per the Sturm et al. (2012) classification (Figures 4B and S4B). We observed that PA tumors were unanimously classified as the third,

Figure 4. A Distinct Subgroup of IDH-Wild-Type Diffuse Glioma with Molecular Features of Pilocytic Astrocytoma

(A) Kaplan-Meier survival curves for the IDH-wild-type glioma subtypes. Ticks represent censorship.
 (B) Distribution of previous published DNA methylation subtypes in the validation set, across the TCGA IDH-wild-type-specific DNA methylation clusters.
 (C) Distribution of genomic alterations in genes frequently altered in IDH-wild-type glioma.
 (D) Heatmap of TCGA glioma samples ordered according to Figure 2A and two EReg gene signatures defined for the IDH-wild-type DNA methylation clusters. Mean RNA sequencing counts for each gene matched to the promoter of the identified cgID across each cluster are plotted to the right.
 (E) Heatmap of the validation set classified using the random forest method using the 1,300 probes defined in Figure 2A.

LGG-enriched group (Figure S4B). Based on the molecular similarity with PA, we labeled the LGGs in the third methylation cluster of IDH-wild-type tumors as PA-like. The GBMs in this group were best described as LGM6-GBM for their original pan-glioma methylation cluster assignment and tumor grade.

Pilocytic astrocytomas are characterized by frequent alterations in the MAPK pathway, such as *FGFR1* mutations, *KIAA1549-BRAF*, and *NTRK2* fusions (Jones et al., 2013). The frequency of mutations, fusions, and amplifications in eight PA-associated genes (*BRAF*, *NF1*, *NTRK1*, *NTRK2*, *FGFR1*, and *FGFR2*) rated from 11% (n = 12/113) of Classic-like, 13% (n = 21/158) of Mesenchymal-like IDH-wild-type tumors to 32% (n = 7/22) of LGM6-GBM and 52% (n = 13/25) of PA-like LGG (Fisher's exact test [FET] p value < 0.0001; Figure 4C). Conversely, only 2 of 25 (8%) PA-like LGG tumors showed

TERT expression, compared to 5 of 12 LGM6-GBM (43%), 60 of 65 Classic-like (92%), and 82 of 98 Mesenchymal-like (84%, FET p value < 0.0001). The PA-like group was characterized by relatively low frequency of typical GBM alterations in genes such as *EGFR*, *CDKN2A/B*, and *PTEN* and displayed euploid DNA copy number profiles (Figure S4E). To ascertain that the histologies of the PA-like subgroup had been appropriately classified, we conducted an independent re-review. This analysis confirmed the presence of the histologic features of diffuse glioma (grade II or grade III) in 23 of the 26 cases in the cluster. The remaining three cases were re-named as PA (grade I). An independent review of the magnetic resonance diagnostic images from 13 cases showed a similar pattern, with the majority of tumors showing behavior consistent with grade II or grade III glioma. Taken together, the epigenetic analysis of the

Table 2. DNA Methylation Subtypes Are Prognostically Relevant in Multivariable Analysis and in External Validation Data

		Discovery (n = 809)			Validation (n = 183)		
		C-Index: 0.835 ± 0.019			C-Index: 0.745 ± 0.032		
Predictor	Levels	n	HR (95% CI)	Signif.	n	HR (95% CI)	Signif.
Age at diagnosis	per year	809	1.05 (1.03–1.06)	***	183	1.02 (1–1.04)	*
WHO Grade	II	214	1.0 (ref)		41	1.0 (ref)	
	III	241	1.96 (1.15–3.33)	*	51	1.24 (0.55–2.76)	
	IV	354	2.38 (1.3–4.34)	*	91	2.6 (1.08–6.3)	*
Subgroup	IDHmut-codel	156	1.0 (ref)		57	1.0 (ref)	
	G-CIMP-low	22	5.6 (2.49–12.62)	***	2	0 (0–Inf)	
	G-CIMP-high	219	1.92 (1.05–3.51)	*	15	1.25 (0.43–3.66)	
	classic-like	143	5.4 (2.79–10.44)	***	22	4.55 (1.8–11.49)	*
	mesenchymal-like I	204	8.71 (4.59–16.53)	***	61	5.55 (2.52–12.21)	***
	LGM6-GBM	39	5.79 (2.78–12.1)	***	22	6.8 (2.58–17.91)	**
	PA-like	26	2.02 (0.71–5.71)		4	3.64 (0.79–16.78)	.

Survival regression analysis indicates that an optimal model of prognosis includes age, grade, and methylation subtype. These predictors are statistically significant in both our discovery dataset and an external validation dataset. Significance codes: 0 “***”; 0.001 “**”; 0.05 “*”; 0.1 “.”

IDH-wild-type group of adult glioma revealed the existence of a novel subgroup sharing genetic and DNA methylation features with pediatric PA and favorable clinical outcome compared to diffuse IDH-wild-type glioma. This group may include but extends beyond *BRAF*-mutated grade II oligodendroglioma that were previously recognized as a unique clinical entity (Chi et al., 2013).

Through comparison of the methylation profiles of 636 glioma and 110 non-neoplastic normal samples from different tissue types, we defined EReg signatures consisting of 27 genes that showed differential signals among IDH-wild-type subtypes in the TCGA (Figure 4D) and the validation set (Figure 4E). EReg4 comprised a group of 15 genes hypermethylated and downregulated in particularly Classic-like. EReg5 was defined as a group of 12 genes associated with hypomethylation in LGM6/PA-like compared to all other LGM clusters. These ERegs aided in characterizing the biological importance of IDH-wild-type subtypes and were subsequently used to evaluate the prognostic importance of the IDH-wild-type clusters.

The Epigenetic Classification of Glioma Provides Prognostic Value Independent of Age and Grade

In order to assess whether the DNA methylation-based subtypes we identified carry prognostically relevant information independent of known overall survival predictors, we constructed a series of survival regression models. To find the optimal model for survival prediction, we studied covariates individually and in combination with other covariates. Age at diagnosis, histology, IDH/codel subtype, *TERT* expression, and epigenetic subtype all contribute to survival in single-predictor analysis (log-rank *p* value < 0.05, Table S4). As expected, age was a highly significant predictor (*p* < 0.0001, C-Index 0.78) and was included in all subsequent multi-predictor models. We found that histology and grade are highly correlated. Histology provided only marginal improvement to a model that includes grade (likelihood ratio test [LRT] *p* value = 0.08) and was therefore not included in further analyses. Conversely, grade markedly impacted a histol-

ogy-based predictor model (LRT *p* value = 0.0005, Table S4) and was retained in the subsequent models. In contrast to previous reports (Eckel-Passow et al., 2015), we failed to observe a statistically significant and independent survival association with *TERT* expression (LRT *p* value = 0.82, Table S4) or *TERT* mutations after accounting for age and grade (LRT *p* value = 0.85, data not shown). Thus, the optimal survival prediction model includes age, grade, and epigenetic subtype (LRT *p* value < 0.0001, C-Index 0.836; Table 2).

To confirm that the epigenetic subtypes provide independent prognostic information, we tested the survival model on the validation dataset. Epigenetic subtypes in these samples were determined as described above. The distinction between LGM6-GBM and PA-like gliomas was made on the basis of tumor grade and not by DNA methylation signature. Using a subset of 183 samples in the validation cohort with known survival, age, and grade, we found that epigenetic subtypes are significant independent predictors of survival in the multivariate analysis (LRT *p* value < 0.0001, C-Index 0.746, Table 2). This generalization of our model supports the epigenetic subtypes as a means to improve the prognostication of glioma.

Activation of Cell Cycle/Proliferation and Invasion/ Microenvironmental Changes Marks Progression of LGG to GBM

We observed that, in spite of morphological differences between LGG and GBM, such as high cell density and microvascular proliferation, clustering of gene expression profiles frequently grouped LGG and GBM together within the same subtype. Gene Set Enrichment Analysis of the genes activated in G-CIMP GBM as opposed to the IDH mutant-non-codel within LGR3 (Figure 2B) revealed four major groups, including cell cycle and hyperproliferation, DNA metabolic processes, response to stress, and angiogenesis (Figure S5A and Table S5). These biological functions are consistent with the criteria based on mitotic index used by pathologists to discriminate lower and high-grade glioma and the significance of activated microglia for tumor

aggressiveness (Roggendorf et al., 1996). Conversely, compared with the G-CIMP GBM, IDH mutant-non-codel LGG in LGr3 were characterized by enrichment of genes associated with neuro-glial functions such as ion transport and synaptic transmission, possibly suggesting a more differentiated nature. The comparison of co-clustered GBM and LGG in LGr3 by the PARADIGM algorithm that integrates DNA copy number and gene expression to infer pathway activity confirmed that GBMs express genes associated with cell cycle, proliferation, and aggressive phenotype through activation of a number of cell cycle, cell replication, and NOTCH signaling pathways whereas LGGs exhibit an enrichment of neuronal-differentiation-specific categories, including synaptic pathways (Figure S5C and Table S5).

The analysis of the genes activated in GBM versus the LGG component of LGr4, which grouped IDH-wild-type tumors, identified an inflammation and immunologic response signature characterized by the activation of several chemokines (*CCL18*, *CXCL13*, *CXCL2*, and *CXCL3*) and interleukins (*IL8* and *CXCR2*) enriching sets involved in inflammatory and immune response, negative regulation of apoptosis, cell cycle and proliferation, and the IKB/NFKB kinase cascade Map (Figure S5B and Table S5). These characteristics suggest differences in the relative amount of microglia. We used the ESTIMATE method to estimate the relative presence of stromal cells, which revealed significantly lower (p value 10^{-6}) stromal scores of LGG IDH-wild-type versus GBM IDH-wild-type (Figure S5F) (Yoshihara et al., 2013). Resembling the functional enrichment for LGG within LGr3, functional enrichment of LGG IDH-wild-type in comparison to GBM within LGr4 showed activation in LGG of special glial-neuronal functions involved in ion transport, synaptic transmission, and nervous system development.

Finally, we aimed to identify transcription factors that may exert control over prominent gene expression programs, known as master regulators. Master regulator analysis comparing the IDH-wild-type group to the IDH mutant group revealed transcription factors that were upregulated in IDH-wild-type gliomas and showed an increase in expression of target genes, including *NKX2-5*, *FOSL1*, *ETV4*, *ETV7*, *RUNX1*, *CEBPD*, *NFE2L3*, *ELF4*, *RUNX3*, *NR2F2*, *PAX8*, and *IRF1* (Table S5). No transcription factors (TFs) were found to be upregulated in IDH mutant gliomas relative to IDH-wild-type gliomas (at a log fold change > 1).

DISCUSSION

This study represents the largest multi-platform genomic analysis performed to date of adult diffuse glioma (WHO grades II, III, and IV). A simplified graphical summary of the identified groups and their main clinical and biological characteristics is reported in Figure 5. The clustering of all diffuse glioma classes and grades within similarly shaped methylation-based and expression-based groups has allowed us to pinpoint specific molecular signatures with clinical relevance. The DNA methylation classification proposed should be considered as a basis and it is likely that future studies involving significantly larger cohorts and more refined profiling methods will be able to further reduce intra-subtype heterogeneity. The dissection of the IDH mutant non-codel G-CIMP LGG and GBM into two separate subgroups (G-CIMP-

low and G-CIMP-high) based on the extent of genome-wide DNA methylation has crucial biological and clinical relevance. In particular, the identification of the G-CIMP-low subset, characterized by activation of cell cycle genes mediated by SOX binding at hypomethylated functional genomic elements and unfavorable clinical outcome, is an important finding that will guide more accurate segregation and therapeutic assessment in a group of patients in which correlations of conventional grading with outcome are modest (Olar et al., 2015; Reuss et al., 2015). The finding that G-CIMP-high tumors can emerge as G-CIMP-low glioma at recurrence identifies variations in DNA methylation as crucial determinants for glioma progression and provides a clue to the mechanisms driving evolution of glioma. Our results unify previous observations that linked the cell cycle pathway to malignant progression of low-grade glioma (Mazor et al., 2015). Future updates of the TCGA glioma clinical annotation and independent validation of our findings may be able to consider additionally important clinical confounders such as extent of resection and performance status to further optimize the weights of the currently known prognostic variables and their association to the molecular subtypes we identified.

Analysis of IDH-wild-type glioma revealed the PA-like LGG subset that harbors a silent genomic landscape, confers favorable prognosis relative to other IDH-wild-type diffuse glioma, and displays a molecular profile with high similarity to PA. Re-review by neuropathologists and neuroradiologists confirmed that the majority were correctly diagnosed as diffuse glioma, emphasizing the need for integration of molecular signatures into clinical classification (Chi et al., 2013) for this subgroup of patients that may be spared potentially unnecessary intensive treatments.

The large number of exomes in our dataset allowed identification of novel glioma-associated somatic alterations, including in the *KRAS* and *NRAS* genes, which were frequently used in genetically engineered glioma mouse models (Holland et al., 2000). Our analysis further nominates glial tumors to join an increasing number of tumor types characterized by a deactivated cohesin pathway (Kon et al., 2013; Solomon et al., 2011). Cohesin mutant tumors may infer increased sensitivity to DNA damage agents and PARP inhibitors (Bailey et al., 2014), suggesting that gliomas with genetic alterations of key cohesin regulatory factors may represent biomarkers and therapeutic opportunities.

Overexpression of *TERT* mRNA was found to be associated with increased telomere length in urothelial cancer (Borah et al., 2015). Our results revealed that, in gliomas, increased telomere length is associated with *ATRX* mutations, suggesting an alternative lengthening of telomeres (ALT) mechanism. ALT has been associated with sensitivity to inhibition of the protein kinase ATR (Flynn et al., 2015).

In summary, our pan-glioma analysis has expanded our knowledge of the glioma somatic alteration landscape, emphasized the relevance of DNA methylation profiles as a modality for clinical classification, and quantitatively linked somatic *TERT* pathway alterations to telomere maintenance. Combined, these findings are an important step forward in our understanding of glioma as discrete disease subsets and the mechanisms driving gliomagenesis.

(legend on next page)

EXPERIMENTAL PROCEDURES

Patient and Sample Characteristics

Specimens were obtained from patients with appropriate consent from institutional review boards. Details of sample preparation are described in the [Supplemental Experimental Procedures](#).

Data Generation

In total, tumors from 1,132 patients were assayed on at least one molecular profiling platform, which platforms included: (1) whole-genome sequencing, including high coverage and low pass whole-genome sequencing; (2) exome sequencing; (3) RNA sequencing; (4) DNA copy-number and single-nucleotide polymorphism arrays, including Agilent CGH 244K, Affymetrix SNP6.0, and Illumina 550K Infinium HumanHap550 SNP Chip microarrays; (5) gene expression arrays, including Agilent 244K Custom Gene Expression, Affymetrix HT-HGU133A and Affymetrix Human Exon 1.0 ST arrays; (6) DNA methylation arrays, including Illumina GoldenGate Methylation, Illumina Infinium HumanMethylation27, and Illumina Infinium HumanMethylation450 Bead-Chips; (7) reverse phase protein arrays; (8) miRNA sequencing; and (9) miRNA Agilent 8 × 15K Human miRNA-specific microarrays. Details of data generation have been previously reported ([Brennan et al., 2013](#); [Cancer Genome Atlas Research Network et al., 2015](#)). To ensure cross-platform comparability, features from all array platforms were compared to a reference genome.

Data Analysis

The data and analysis results can be explored through the Broad Institute FireBrowse portal (<http://firebrowse.org/?cohort=GBMLGG>), the cBioPortal for Cancer Genomics (http://www.cbioportal.org/study.do?cancer_study_id=lgggbm_tcga_pub), in a Tumor Map (<http://tumormap.ucsc.edu/?p=ynewton.gliomas-paper>), the TCGA transcript fusion portal (<http://www.tumorfusions.org>), TCGA Batch Effects (<http://bioinformatics.mdanderson.org/tcgambatch/>), Regulome Explorer (<http://explorer.cancerregulome.org/>), Next-Generation Clustered Heat Maps (<http://bioinformatics.mdanderson.org/TCGA/NGCHMPortal/>). See also [Supplemental Information](#) and the TCGA publication page (https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015/).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and five tables and can be found with this article online at <http://dx.doi.org/10.1016/j.cell.2015.12.028>.

AUTHOR CONTRIBUTIONS

Conceptualization and project administration: R.G.W.V., A.I., and H.N.; supervision: S.R.S., K.D.A., P.W.L., M.G., D.H., D.J.B., D.H.G., R.R., C.C.L., J.S.B.-S., C.G.C., D.P.C.T., W.K.A.Y., J.H., L.C., M.M., and T.M.; formal analysis: R.G.W.V., A.I., H.N., M.C., F.P.B., T.M.M., T.S.S., O.M., Y.N., S.M.P., P.Z., L.P., A. Radenbaugh, G.R., R.A., J.W., G.M., S.L., S.A., A. Rao, B.A.M., A.D.C., and H.Z.; investigation: D.J.B., L.C., and L.P.; data curation: D.J.B., L.P., and F.P.B.; writing - original draft: R.G.W.V., A.I., H.N., M.C., F.P.B., T.M.M., and T.S.S.; manuscript review: D.J.B., K.A.D., S.R.S., M.W., N.L., and D.H.G.

ACKNOWLEDGMENTS

This study was supported by NIH grants U24CA143883, U24CA143858, U24CA143840, U24CA143799, U24CA143835, U24CA143845,

U24CA143882, U24CA143867, U24CA143866, U24CA143848, U24CA144025, U54HG003067, U54HG003079, U54HG003273, U24CA126543, U24CA126544, U24CA126546, U24CA126551, U24CA126554, U24CA126561, U24CA126563, U24CA143731, U24CA143843, P30CA016672, P50 CA127001, U54CA193313, R01CA179044, R01CA185486, R01 CA190121, and P01 CA085878; Cancer Prevention & Research Institute of Texas (CPRIT) R140606; and São Paulo Research Foundation (FAPESP) 2014/02245-3, 2015/07925-5, 2015/02844-7, and 2015/08321-3. D.J.W. is a consultant for Zymo Research Corporation. R.B. is a consultant for and received grant funding from Novartis. A.D.C. and M.M. received grant support from Bayer.

Received: July 17, 2015

Revised: October 20, 2015

Accepted: December 11, 2015

Published: January 28, 2016

REFERENCES

- Bailey, M.L., O'Neil, N.J., van Pel, D.M., Solomon, D.A., Waldman, T., and Hietter, P. (2014). Glioblastoma cells containing mutations in the cohesin component STAG2 are sensitive to PARP inhibition. *Mol. Cancer Ther.* 13, 724–732.
- Berman, B.P., Weisenberger, D.J., Aman, J.F., Hinoue, T., Ramjan, Z., Liu, Y., Noshmeh, H., Lange, C.P., van Dijk, C.M., Tollenaar, R.A., et al. (2012). Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* 44, 40–46.
- Borah, S., Xi, L., Zaig, A.J., Powell, N.M., Dancik, G.M., Cohen, S.B., Costello, J.C., Theodorescu, D., and Cech, T.R. (2015). Cancer. TERT promoter mutations and telomerase reactivation in urothelial cancer. *Science* 347, 1006–1010.
- Brennan, C.W., Verhaak, R.G., McKenna, A., Campos, B., Noshmeh, H., Salama, S.R., Zheng, S., Chakravarty, D., Sanborn, J.Z., Berman, S.H., et al.; TCGA Research Network (2013). The somatic genomic landscape of glioblastoma. *Cell* 155, 462–477.
- Cancer Genome Atlas Research Network, Brat, D.J., Verhaak, R.G., Aldape, K.D., Yung, W.K., Salama, S.R., Cooper, L.A., Rheinbay, E., Miller, C.R., Vitucci, M., et al. (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* 372, 2481–2498.
- Chi, A.S., Batchelor, T.T., Yang, D., Dias-Santagata, D., Borger, D.R., Ellisen, L.W., Iafrate, A.J., and Louis, D.N. (2013). BRAF V600E mutation identifies a subset of low-grade diffusely infiltrating gliomas in adults. *J. Clin. Oncol.* 31, e233–e236.
- Eckel-Passow, J.E., Lachance, D.H., Molinaro, A.M., Walsh, K.M., Decker, P.A., Sicotte, H., Pekmezci, M., Rice, T., Kosel, M.L., Smirnov, I.V., et al. (2015). Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors. *N. Engl. J. Med.* 372, 2499–2508.
- Flynn, R.L., Cox, K.E., Jeitany, M., Wakimoto, H., Bryll, A.R., Ganem, N.J., Bersani, F., Pineda, J.R., Suvà, M.L., Benes, C.H., et al. (2015). Alternative lengthening of telomeres renders cancer cells hypersensitive to ATR inhibitors. *Science* 347, 273–277.
- Frattini, V., Trifonov, V., Chan, J.M., Castano, A., Lia, M., Abate, F., Keir, S.T., Ji, A.X., Zoppoli, P., Niola, F., et al. (2013). The integrated landscape of driver genomic alterations in glioblastoma. *Nat. Genet.* 45, 1141–1149.
- Graham, V., Khudyakov, J., Ellis, P., and Pevny, L. (2003). SOX2 functions to maintain neural progenitor identity. *Neuron* 39, 749–765.

Figure 5. Overview of Major Subtypes of Adult Diffuse Glioma

Integrative analysis of 1,122 adult gliomas resulted in 7 different subtypes with distinct biological and clinical characteristics. The groups extend across six DNA methylation subtypes of which the LGM6 cluster was further separated by tumor grade into PA-like and LGM6-GBM. The size of the circles is proportional to the percentages of samples within each group. DNA methylation plot is a cartoon representation of overall genome-wide epigenetic pattern within glioma subtypes. Survival information is represented as a set of Kaplan-Meier curves, counts of grade, histology and LGG/GBM subtypes within the groups are represented as bar-plots, whereas age is represented as density. Labeling of telomere length and maintenance status is based on the enrichment of samples within each column, similarly for the biomarkers and the validation datasets.

- Guintivano, J., Aryee, M.J., and Kaminsky, Z.A. (2013). A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* 8, 290–302.
- Holland, E.C., Celestino, J., Dai, C., Schaefer, L., Sawaya, R.E., and Fuller, G.N. (2000). Combined activation of Ras and Akt in neural progenitors induces glioblastoma formation in mice. *Nat. Genet.* 25, 55–57.
- Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. *Science* 339, 959–961.
- Jones, D.T., Hutter, B., Jäger, N., Korshunov, A., Kool, M., Warnatz, H.J., Zichner, T., Lambert, S.R., Ryzhova, M., Quang, D.A., et al.; International Cancer Genome Consortium PedBrain Tumor Project (2013). Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. *Nat. Genet.* 45, 927–932.
- Killela, P.J., Reitman, Z.J., Jiao, Y., Bettegowda, C., Agrawal, N., Diaz, L.A., Jr., Friedman, A.H., Friedman, H., Gallia, G.L., Giovannella, B.C., et al. (2013). TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc. Natl. Acad. Sci. USA* 110, 6021–6026.
- Kim, H., Zheng, S., Amini, S.S., Virk, S.M., Mikkelsen, T., Brat, D.J., Grimsby, J., Sougnez, C., Muller, F., Hu, J., et al. (2015). Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Res.* 25, 316–327.
- Kon, A., Shih, L.Y., Minamino, M., Sanada, M., Shiraishi, Y., Nagata, Y., Yoshida, K., Okuno, Y., Bando, M., Nakato, R., et al. (2013). Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nat. Genet.* 45, 1232–1237.
- Kurscheid, S., Bady, P., Sciuscio, D., Samarzija, I., Shay, T., Vassallo, I., Griesinger, W.V., Daniel, R.T., van den Bent, M.J., Marosi, C., et al. (2015). Chromosome 7 gain and DNA hypermethylation at the HOXA10 locus are associated with expression of a stem cell related HOX-signature in glioblastoma. *Genome Biol.* 16, 16.
- Lambert, S.R., Witt, H., Hovestadt, V., Zucknick, M., Kool, M., Pearson, D.M., Korshunov, A., Ryzhova, M., Ichimura, K., Jabado, N., et al. (2013). Differential expression and methylation of brain developmental genes define location-specific subsets of pilocytic astrocytoma. *Acta Neuropathol.* 126, 291–301.
- Lodato, M.A., Ng, C.W., Wamstad, J.A., Cheng, A.W., Thai, K.K., Fraenkel, E., Jaenisch, R., and Boyer, L.A. (2013). SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs and NPCs to specify cell state. *PLoS Genet.* 9, e1003288.
- Louis, D.N., Ohgaki, H., Wiestler, O.D., Cavenee, W.K., Burger, P.C., Jouvett, A., Scheithauer, B.W., and Kleihues, P. (2007). The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* 114, 97–109.
- Mazor, T., Pankov, A., Johnson, B.E., Hong, C., Hamilton, E.G., Bell, R.J., Smirnov, I.V., Reis, G.F., Phillips, J.J., Barnes, M.J., et al. (2015). DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors. *Cancer Cell* 28, 307–317.
- Mur, P., Mollejo, M., Ruano, Y., de Lope, A.R., Fiaño, C., García, J.F., Castresana, J.S., Hernández-Lain, A., Rey, J.A., and Meléndez, B. (2013). Codeletion of 1p and 19q determines distinct gene methylation and expression profiles in IDH-mutated oligodendroglial tumors. *Acta Neuropathol.* 126, 277–289.
- Noushmehr, H., Weisenberger, D.J., Diefes, K., Phillips, H.S., Pujara, K., Berman, B.P., Pan, F., Pelloski, C.E., Sulman, E.P., Bhat, K.P., et al.; Cancer Genome Atlas Research Network (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 17, 510–522.
- Olar, A., Wani, K.M., Alfaro-Munoz, K.D., Heathcock, L.E., van Thuijl, H.F., Gilbert, M.R., Armstrong, T.S., Sulman, E.P., Cahill, D.P., Vera-Bolanos, E., et al. (2015). IDH mutation status and role of WHO grade and mitotic index in overall survival in grade II–III diffuse gliomas. *Acta Neuropathol.* 129, 585–596.
- Ozawa, T., Riester, M., Cheng, Y.K., Huse, J.T., Squatrito, M., Helmy, K., Charles, N., Michor, F., and Holland, E.C. (2014). Most human non-GCIMP glioblastoma subtypes evolve from a common proneural-like precursor glioma. *Cancer Cell* 26, 288–300.
- Peters, J.M., and Nishiyama, T. (2012). Sister chromatid cohesion. *Cold Spring Harb. Perspect. Biol.* 4, a011130.
- Pineda, C.T., Ramanathan, S., Fon Tacer, K., Weon, J.L., Potts, M.B., Ou, Y.H., White, M.A., and Potts, P.R. (2015). Degradation of AMPK by a cancer-specific ubiquitin ligase. *Cell* 160, 715–728.
- Reuss, D.E., Mamatjan, Y., Schrimpf, D., Capper, D., Hovestadt, V., Kratz, A., Sahm, F., Koelsche, C., Korshunov, A., Olar, A., et al. (2015). IDH mutant diffuse and anaplastic astrocytomas have similar age at presentation and little difference in survival: a grading problem for WHO. *Acta Neuropathol.* 129, 867–873.
- Roggendorf, W., Strupp, S., and Paulus, W. (1996). Distribution and characterization of microglia/macrophages in human brain tumors. *Acta Neuropathol.* 92, 288–293.
- Schwartzbaum, J.A., Fisher, J.L., Aldape, K.D., and Wrensch, M. (2006). Epidemiology and molecular pathology of glioma. *Nat. Clin. Pract. Neurol.* 2, 494–503.
- Solomon, D.A., Kim, T., Diaz-Martinez, L.A., Fair, J., Elkahoul, A.G., Harris, B.T., Toretsky, J.A., Rosenberg, S.A., Shukla, N., Ladanyi, M., et al. (2011). Mutational inactivation of STAG2 causes aneuploidy in human cancer. *Science* 333, 1039–1043.
- Sturm, D., Witt, H., Hovestadt, V., Khuong-Quang, D.A., Jones, D.T., Konermann, C., Pfaff, E., Tönjes, M., Sill, M., Bender, S., et al. (2012). Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* 22, 425–437.
- Sturm, D., Bender, S., Jones, D.T., Lichter, P., Grill, J., Becher, O., Hawkins, C., Majewski, J., Jones, C., Costello, J.F., et al. (2014). Paediatric and adult glioblastoma: multifactorial (epi)genomic culprits emerge. *Nat. Rev. Cancer* 14, 92–107.
- Suvà, M.L., Rheinbay, E., Gillespie, S.M., Patel, A.P., Wakimoto, H., Rabkin, S.D., Riggi, N., Chi, A.S., Cahill, D.P., Nahed, B.V., et al. (2014). Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell* 157, 580–594.
- Suzuki, H., Aoki, K., Chiba, K., Sato, Y., Shiozawa, Y., Shiraishi, Y., Shimamura, T., Niida, A., Motomura, K., Ohka, F., et al. (2015). Mutational landscape and clonal architecture in grade II and III gliomas. *Nat. Genet.* 47, 458–468.
- Turcan, S., Rohle, D., Goenka, A., Walsh, L.A., Fang, F., Yilmaz, E., Campos, C., Fabius, A.W., Lu, C., Ward, P.S., et al. (2012). IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* 483, 479–483.
- van den Bent, M.J. (2010). Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathol.* 120, 297–304.
- Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., et al.; Cancer Genome Atlas Research Network (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110.
- Yan, H., Parsons, D.W., Jin, G., McLendon, R., Rasheed, B.A., Yuan, W., Kos, I., Batinic-Haberle, I., Jones, S., Riggins, G.J., et al. (2009). IDH1 and IDH2 mutations in gliomas. *N. Engl. J. Med.* 360, 765–773.
- Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-García, W., Treviño, V., Shen, H., Laird, P.W., Levine, D.A., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4, 2612.