12-2019

Deconvolute brain tumor genomic alterations based on DNA methylation

Jie Yang

DECONVOLUTE BRAIN TUMOR GENOMIC ALTERATIONS

BASED ON DNA METHYLATION

By

Jie Yang, M.S., B.M.

APPROVED:

_____
Erik P. Sulman, M.D., Ph.D.
Advisory Professor

_____
Jason T. Huse, M.D., Ph.D.
(Onsite advisor)

_____
Krishna P.L. Bhat, Ph.D.

_____
Nicholas Navin, Ph.D.

_____
Arvind Rao, Ph.D.

_____
Ann Klopp, M.D., Ph.D.

APPROVED:

_____
Dean, The University of Texas
MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

DECONVOLUTE BRAIN TUMOR GENOMIC ALTERATIONS BASED ON DNA METHYLATION

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Jie Yang, M.S., B.M.

Houston, Texas

December, 2019

In dedication to my father Wenqing Yang, my mother Juan Tu,

and my husband Yinsen Miao.

Acknowledgments

Time flies so fast and now it is my fifth-year PhD study. It would not have been possible without the support and guidance that I received from many people.

I would like to give my big thank you to my PhD advisor, Dr. Erik P. Sulman. It is a great honor to be his PhD student and I feel super lucky about it. I always remember what he told me about PhD training: it is not about publication, not about any significant findings, but learn how to be a scientist, how to think like a scientist. He is always kind, generous, and helpful.

I would like to thank my advisory committee, Dr. Jason Huse, Dr. Krishna Baht, Dr. Nick Navin, Dr. Arvind Rao, and Dr. Ann Klopp, for being part of my journey and witnessing my growth. Special acknowledgment to Dr. Huse, who is my onsite advisor after I moved to New York, for all the paper works. My sincere thanks also go to my lab mates, thank you guys for being so supportive.

I really appreciate the support received from my program and GSBS. I am especially grateful to Dr. Ram and Amy in Quantitative Science (QS) program, and Dr. Mattox, Brenda, Bunny, and Patricia in GSBS.

Last but not least, I would like to thank my family: my parents and my husband for everything you have done for me. I know I go crazy at some moments, but I know you will always be there for me, providing me love, comfort, and support.

DECONVOLUTE BRAIN TUMOR GENOMIC ALTERATIONS BASED ON DNA METHYLATION DATA

Jie Yang, M.S.

Advisory Professor: Erik Sulman, M.D. Ph.D.

Molecular classification based on mutations, expression subtypes, and copy number variants has improved diagnosis and treatment decision-making for patients with brain tumors, particularly malignant gliomas. However, the association between epigenetic signature and genetic alterations is poorly understood. For example, mutation of isocitrate dehydrogenase (*IDH*) is associated with genome-wide hypermethylation of CpG islands in gliomas. But other subtype-associated alterations, including telomerase reverse transcriptase (*TERT*) promoter mutation, alpha thalassemia/mental retardation syndrome X-linked (*ATRX*) mutation, chromosome 1p19q co-deletion (chr1p19q codel), and gene expression subtypes, have yet to be associated with any epigenetic signature. Therefore, we hypothesized that DNA methylation signatures can classify gliomas based on these alterations and give insight into subgroup characteristics. Machine learning models, including elastic net and random forest, were used to predict somatic mutations of *IDH*, *TERT*p, and *ATRX*, chr1p19q codel, and gene expression subtype of gliomas. Data from the NOA-04 randomized phase III trial were used for external validation. In total, 926 cases from The Cancer Genome Atlas were included in this study. Prediction accuracies for *IDH*, *TERT*p, and *ATRX* mutations, and chr1p19q codel were 100%, 98.3%, 90.48%, and 99.21%, respectively in test set. Accuracy for gene expression subtype prediction was 72.2%. The methylation-based prediction models for both *ATRX* and chr1p19q codel statuses proved superior to conventional assays for these biomarkers. Similarly, characteristic alterations associated with gene expression subtypes were better discriminated using methylation compared to transcriptome-based classification. DNA methylation signatures accurately predicted somatic alterations and improved over existing classifiers. The established Unified Diagnostic Pipeline (UniD) is a rapid and cost-effective diagnostic platform of genomic alterations and gene expression subtypes at initial clinical diagnosis and improves over individual assays currently in clinical use. The significant

relationship between genetic alterations and epigenetic signatures indicates the broad applicability of our approach to other malignancies.

Contents

List of Illustrations

List of Tables

Abbreviations

2-HG	2- hydroxyglutarate
3-mC	3-methylcytosine
5-caC	5-carboxylcytosine
5-fC	5-formylcytosine
5-hmC	5-hydroxymethylcytosine
5-mC	5-Methylcytosine
A	Adenine
ADD	ATRX-DNMT3-DNMT3L
AI	artificial intelligence
AID	activation-induced cytidine deaminase
a-KG	α–ketoglutarate
ALT	alternative lengthening of telomeres
AML	acute myeloid leukemia
APBs	ALT-associated PML bodies
ASMN	all sample mean normalization
ASXL	additional sex combs like 2, transcriptional regulator
ATL3	atlastin GTPase 3
ATR	ATR serine/threonine kinase
ATRX	Alpha Thalassemia/Mental Retardation Syndrome X-linked
AUC	area under the curve
BACR	BeadArray control reporter
BMI-1	polycomb group RING finger protein 4
BMIQ	beta-mixture quantile normalization
BMP	bone morphogenetic proteins
bp	base pair
BRAFV600E	v-raf murine sarcoma viral oncogene homolog B at V600
BRCA1	breast cancer 1, early onset
BRD4	bromodomian containing 4

C	Cytosine
CBS	circular binary segmentation
CBTRUS	Central Brain Tumor Registry of the United States
CDKN2A	cyclin dependent kinase inhibitor 2A
CFLAR	CASP8 and FADD like apoptosis regulator
CHI3L1	chitinase-3-like protein 1
chr1p19q codel	Chromosome 1p19q co-deletion
CIC	capicua transcriptional repressor
CL	classical
CML	chronic myelogenous leukemia
CNA	Copy Number Alterations
CNMF	consensus non-negative matrix factorization
CNS	central nervous system
CNV	copy number variation
CREBBP	CREB binding protein
CTSF	cathepsin F
CV	Cross-validation
D2HGDH	D-2-hydroxyglutarate dehydrogenase, mitochondrial
DAXX	Death-Domain Associated Protein
DEDD2	death effector domain containing 2
DNMT	DNA methyltransferase
DNMTs	DNA methyltransferase enzymes
DSBR	double-strand break repair
DSBs	Double-strand breaks
ECTR	abundant extrachromosomal telomeric repeat DNA
EGFR	epidermal growth factor receptor
EP300	histone acetyltransferase P300
ERBB2	erb-b2 receptor tyrosine kinase 2
EZH2	enhancer of zeste 2 polycomb repressive complex 2 subunit

EZH2	enhancer of Zeste 2 polycomb repressive complex 2 subunit
FAS	tumor necrosis factor receptor superfamily member 6
FBXO6	Gene F-box protein 6
FFPE	formalin-fixed paraffin-embedded
FGFR2	fibroblast growth factor receptor 2
FIS	Functional impact score
FISH	fluorescent in situ hybridization
FUBP1	far upstream element binding protein 1
G	Guanine
G-CIMP	glioblastoma CpG island methylator phenotype
G4	G-quadruplexes
G9a	enchromatic histone lysine methyltransferase 2
GABRA1	gamma-aminobutyric acid receptor alpha 1
GBM	glioblastoma
GDA	gaussian discriminant analysis
glmnet	Lasso and Elastic-Net Regularized Generalized Linear Model
GO	gene ontology
GR	Gain ratio
GSC	glioma sphere-forming cell
HCC	hepatocellular carcinoma
HDAC2	hisone deacetylase 2
HDAC5/7A	hisone deacetylase 5
HJs	Holliday junctions
HM	HumanMethylation
HM27k	Illumina HumanMethylation 27 BeadChip
HM450k	Illumina HumanMethylation 450 BeadChip
HR	homologous recombination
ICF syndrome	immunodeficiency-centromeric instability-facial anomalies syndrome
IDH	Isocitrate dehydrogenase

IG	Information gain
IHC	immunohistochemistry
ISM1	isthmin 1
JARID1B/C	lysine demethylase 5B
KCNIP3	potassium voltage-gated channel interacting protein 3
KNN	k-nearest neighbor
LASSO	least absolute shrinkage and selection operator
LDA	linear discriminant analysis
LGALS1	lectin, galactoside-binding, soluble, 1
LGG	Low-grade gliomas
LOH	loss of heterozygosity
LOS	ordinary least square
LSD1	lysine-specific demethylase 1
MAD	median absolute deviation
MB	medulloblastoma
MBD1/2	methyl-cpg binding domain protein 1 or 2
MDS	myelodysplastic syndromes
MES	mesenchymal
MGMT	O-6-methylguanine-DNA methyltransferase
mH2a	macroH2A
MI	mutual information
ML	machine learning
MLL1/2/3	lysine methyltransferase 2A/2D/2C
MLPA	multiplex ligation-dependent probe amplification
mRNA	messenger RNA
MS-PCR	methylation specific PCR
MS-qLNAPCR	methylation sensitive-quantitative locked nucleic acid PCR
NAD+	nicotinamide adenine dinucleotide
NADP+	nicotinamide adenine dinucleotide phosphate

NB	naïve Bayes
NEFL	neurofilament protein, light polypeptide
NF-kB	nuclear factor-kB
NF1	neurofibromin 1
NR4A1	nuclear receptor subfamily 4, group A, member 1
OLIG2	oligodendrocyte transcription factor
PAM	Prediction Analysis of Microarray
PBC	Peak-based correction
PCAF	lysine acetyltransferase 2B
PCNA	proliferating cell nuclear antigen
PCV	procarbazine/lomustine/vincristine
PDE7B	phosphodiesterase 7B
PDGFRA	platelet derived growth factor receptor alpha
PML	promyelocytic leukemia
PN	proneural
PNS	peripheral nervous system
PRMT1/5	protein arginine methyltransferase 1
PRODH	proline dehydrogenase 1
PTEN	phosphatase and tensin homolog
QC	quality control
R132H	arginine replaced with histidine
RARG	retinoic acid receptor, gamma
RB1	retinoblastoma protein 1
RBP1	retinol binding protein 1
RCCC	renal cell carcinoma
ROC	receiver operator characteristic
RRBS	sequencing-by-synthesis, representation bisulphite sequencing
RTOG	The Radiation therapy Oncology Group
SAM	S-adenosyl methionine

SDSA	synthesis-dependent strand annealing
SFK	SRC family kinase
SIRT1	NAD-dependent protein deacetylase sirtuin -1
SNF2	transcription regulatory protein SNF2
SNP	single nucleotide polymorphism
SNP6	Affymetrix 6.0 platform
SNVs	single nucleotide variations
SOCS2	suppressor of cytokine signaling 2
SQN	subset quantile normalization
SRC	v-Src Avian sarcoma viral oncogene homolog
ssDNA	single-stranded DNA
SU	Symmetrical Uncertainty
SVM	support vector machines
SWAN	subset quantile within array normalization
SWI/SNF	SWItch/Sucrose Non-Fermentable
SYT1	synaptotagmin 1
T	Thymine
t-SCES	telomere sister chromatid exchanges
TCGA	The Cancer Genome Atlas
TERC	telomerase RNA component
TERTp	Telomerase reverse transcriptase promoter
TET	ten-eleven translocation
TET1/2	TET methylcytosine dioxygenase 1 or 2
TMZ	temozolomide
TP53	tumor protein p53
TSS	transcriptional start site
U	uracil
UHRF1	ubiquitin-like, containing PHD and RING figure domains 1
UniD	Unified Diagnostic

UTX	lysine demethylase 6A
WHO	World Health Organization

Chapter 1

1. Introduction

A glioma is a type of brain tumor that originates in the glial cells and is the most commonly seen malignant brain tumor. Gliomas are classified into four different grades according to their malignancy, growth speed, cell infiltration, and aggressiveness. Grade IV gliomas are also called glioblastomas (GBMs) and are lethal in almost every patient. The standard care for glioma patients is surgery followed with or without radiation/chemotherapy. However, due to limited understandings of this devastating tumor, there is no efficient treatment for it and the prognosis is still very poor. In recent years scientists have been able to gain more insight about gliomas' molecular features because of the development of molecular biotechnology.

Glioma classification based on DNA methylation and gene expression profiles helps identify subgroups with characteristic molecular features. In addition to classification, specific genomic alterations play an important role in tumor initiation, progression, and prognosis. For example, somatic mutations of the gene isocitrate dehydrogenase (*IDH*), telomerase reverse transcriptase promoter (*TERT*p), and alpha thalassemia/mental retardation syndrome X-linked (*ATRX*) and chromosome 1p19q co-deletion (chr1p19q codel). It has been found that *IDH* mutation shows a strong relationship with a subgroup of GBM with genome-wide hypermethylation. Meanwhile, DNA methylation profile is believed to reflect the tumor cell of origins. Therefore, I decided to investigate the potential relationships between DNA methylation profiles and known somatic genomic alterations. My study will not only lead to insights about gliomas' genomic alterations but will also allow multiple genomic alteration status to be determined with only one simple assay.

The first chapter introduces the basic information about gliomas and the related molecular information, DNA methylation platform and microarray data, and machine learning models and techniques used in the study. The second chapter is focused on the methods and results of building predictive models for binary genomic alterations using DNA methylation data. The third chapter introduces the methods and results of building predictive models for gene expression subtypes using DNA methylation data. The last chapter includes this study's conclusions, discussions, limitations, and future directions. My studies of the relationship of DNA methylation and somatic genomic alterations show that DNA methylation can accurately predict genomic alterations in gliomas and reveal enrichment of characteristic genomic alterations. Furthermore, in addition to all the known advantages, methylation microarray holds the potential to expand DNA methylation biomarker predictive signatures in gliomas and other types of cancer.

1.1 Brain tumor

A brain tumor is a tumor grown in the brain and may be benign or malignant. Brain tumors can also be categorized as primary brain tumors and metastatic brain tumors based on their origins. A primary brain tumor is a tumor that originates from brain tissue, and a metastatic brain tumor usually begins in other parts of the body and later migrates to the brain through the blood. Most metastatic brain tumors originate from breast cancer, lung cancer, kidney cancer, or melanoma. The diagnosis of a brain tumor requires a combination of sophisticated imaging tools and determination of tumor biopsy histopathology. After diagnosis, patients with brain tumors are treated with surgery, and/or radio- or chemotherapy.

1.1.1 Glioma introduction

A glioma is a type of brain tumor that develops from glial cells. Glial cells are the most abundant type of cell in the brain and can provide support, nutrition, and energy to the neurons.

There are different types of glia cells including astrocyte, oligodendrocyte, microglia, and others in the peripheral nervous system.

Gliomas are the most commonly seen primary malignant brain tumor in adults. Based on World Health Organization (WHO) classification of tumors of the central nervous system (CNS (1)), there are four grades (grade I to IV) based on their malignancy, growth speed, tumor cell infiltration, and aggressiveness (**Table 1**). Grade I gliomas are called pilocytic astrocytoma and often occur in children. This type of tumor usually grows slowly and is relatively benign. Grade II gliomas include astrocytoma, oligodendroglioma, and mixed oligoastrocytoma. Grade III gliomas include anaplastic astrocytoma, anaplastic oligodendroglioma, and anaplastic mixed oligoastrocytoma. The highest-grade glioma, grade IV, is usually called GBM and is the most common and malignant type of brain tumor. Patients with GBM usually die within one year after diagnosis (2). Low-grade gliomas (LGGs, including grades I and II) usually are less malignant and aggressive and have better prognoses than high-grade gliomas (grade III and IV). However, LGGs have a high probability of evolving into high-grade gliomas.

Table 1: Glioma grading

Grade	benign/malignant	growth rate	major subtypes
I	benign	slow	pilocytic astrocytoma
II	malignant	medium	astrocytoma, oligodendroglioma, mixed oligoastrocytoma
III	malignant	medium	anaplastic astrocytoma, anaplastic oligodendroglioma, anaplastic mixed oligoastrocytoma
IV	malignant	fast	glioblastoma

Epidemiological research about brain and other CNS tumors reported 392982 incidences between 2011 and 2015 (3), 23.8% of which were gliomas. Among all gliomas, about 80% are malignant and 61.7% are GBMs. The median age of diagnosis with glioma is 57 years old. Gliomas are observed more in male (56.1%), white (88%) patients. Among all malignant brain and CNS tumors, GBM has the highest rate of 3.21 per 100,000 and the five-year survival rate

is about 5.6% for all ages. **Table 2** shows that the five-year survival rate decreases significantly as the grade of glioma increases.

Table 2: one-year and five-year survival rate (%) and 95% CI* (months) for glioma by grading

grade	subtype	one-year		five-year	
		survival rate	95% CI*	survival rate	95% CI*
I	pilocytic astrocytoma	97.9	97.4-98.3	94.1	93.2-94.8
II	astrocytoma	74.9	73.8-76.0	50.4	49.0-51.7
	oligodendroglioma	94.7	93.8-95.4	81.6	20.1-83.1
	oligoastrocytoma	88.7	87.3-90.0	63.7	61.4-65.8
III	anaplastic astrocytoma	65.3	63.9-66.8	30	28.4-31.5
	anaplastic oligodendroglioma	84.4	82.4-86.2	57.6	54.7-60.4
IV	glioblastoma	40.2	39.7-40.7	5.6	5.3-5.8

* CI: confidence interval

Note: Data summarized from Central Brain Tumor Registry of the United States (CBTRUS) 2018 version (3)

Current understandings of tumor initiation, evolution, and progression are poor. In recent years, many researchers have shown that high-throughput genomic data may help to understand this devastating tumor better. Several predictive biomarkers have been discovered, such as *IDH* mutation status, which is positively associated with younger age and longer survival time; chr1p19q codel, which is prognostic of improved survival and predictive of response to chemotherapy (1, 4); and O-6-methylguanine-DNA methyltransferase (*MGMT*) methylation status (5), which is beneficial for patients' treatment with alkylating chemotherapy such as carmustine or temozolomide. The first two biomarkers, *IDH* mutation and chr1p19q codel have been included in the recent revisions of WHO diagnostic criteria for gliomas. In addition, *TERT*p mutation (6) and *ATRX* mutation are mutually exclusive but both are functionally correlated with telomere length maintenance (7, 8).

GBM has also been classified into three subtypes base on characteristic gene expression signatures called classical (CL), proneural (PN), and mesenchymal (MES) (9). CL GBMs frequently have epidermal growth factor receptor (*EGFR*) amplification and higher expression

while MES GBMs have high rates of neurofibromin 1 (*NF1*) loss function mutation (9). GBMs have distinguished copy number alterations (CNAs) compared to other types of cancer. For example, GBM samples have characteristic chromosome 7 amplifications and chromosome 10 deletions. In the following subsections, those key genetic alterations are introduced in detail.

1.1.2 Molecular Classification

The Cancer Genome Atlas (TCGA) is a landmark cancer-related genomic program initiated in 2006. It has molecularly characterized over 20,000 primary cancer samples spanning 33 cancer types. GBM was the first systematically investigated cancer type in this program and then the program expanded to other LGGs. In this project, glioma samples were well-characterized in many genomic aspects, including messenger RNA (mRNA) gene expression level, DNA sequencing, copy number variations, DNA methylation profile, and others. This provided me the chance to study it as a whole set, therefore, different molecular classifications were proposed. I will give a brief introduction of some important classifications according to published papers.

1.1.2.1 Gene expression classification of GBM

In 2006, Phillips et al. (10) published the first high-impact paper about the gene expression subclasses of high-grade astrocytoma in Cancer Cell. They studied 76 newly diagnosed high-grade astrocytoma patients with DNA microarray gene expression data and clustered them into three subclasses: PN, MES and proliferation. These three subclasses of samples show differences in tumor grade, patient age, proliferation, angiogenesis, neurogenesis, copy number variations, and signaling pathway activation status. By comparing the paired primary and recurrent gliomas, recurrent samples were found to shift towards the mesenchymal phenotype and showed frequent loss of oligodendrocyte transcription factor (*OLIG2*) expression and chitinase-3-like protein 1 (*CHI3L1*) upregulation. More importantly, the authors tried to explain the progression process between these three subclasses in their paper.

In 2010, Verhaak et al. (11) published a paper about the gene expression subtypes of GBM using a larger data set from TCGA and integrating multiple genomic characterizations. In this paper, the authors classified GBM into four subtypes: PN, MES, neural, and CL. Each subtype was characterized in terms of gene expression level, clinical phenotypes, copy number variations, gene mutation rate, and relationship with cell types. Among them, some genes were well characterized for each subtype, including tumor protein p53 (*TP53*), *IDH1*, platelet-derived growth factor receptor alpha (*PDGFRA*), *EGFR*, *NF1*, and cyclin-dependent kinase inhibitor 2A (*CDKN2A*; **Table 3**). In short, PN subtype samples were characterized by high *IDH* mutation rates and high expression and amplification of *PDGFRA*. Neural subtype samples showed high expression of neuron marker genes, including neurofilament protein, light polypeptide (*NEFL*), gamma-aminobutyric acid receptor alpha 1 (*GABRA1*), synaptotagmin 1 (*SYT1*), and so on. MES subtype samples showed high mutation rates and low expression of *NF1*, focal hemizygous deletions at chr17q11.2, high nuclear factor-κB (NF-κB) pathway expression, and poor prognosis. CL GBM showed frequent chr10 loss and chr7 amplification, high *EGFR* expression and amplification, homozygous *CDKN2A* deletion, and high expression of Notch and Sonic hedgehog signaling pathway. Though these four subtypes did not show indications of a strong prognosis, they provided a systematic understanding in terms of mRNA expression and correlation of the subtypes to other genomic features.

Table 3: Characteristics of gene expression subtypes by 2010 Cancer Cell paper

Genomic feature	Proneural	Neural	Mesenchymal	Classical
median diagnostic age	51.8	63.8	57.7	55.7
IDH mutation rate	30.0% (11/37)	5% (1/19)	0% (0/38)	0% (0/22)
TP53 mutation rate	54% (20/37)	21% (4/19)	32% (12/38)	0% (0/22)
PDGFRA mutation rate	11% (4/37)	0% (0/19)	0% (0/38)	0% (0/22)
EGFR mutation rate	16% (6/37)	26% (5/19)	5% (2/38)	32% (7/22)
NF1 mutation rate	5% (2/37)	16% (3/19)	37% (14/38)	5% (1/22)

Note: Information summarized from Verhaak et al. (11)

In 2017, an improved gene expression-based classification system was proposed by Wang et al. (9) by using the single cell sequencing technology. With about 600 single glioma cells RNA sequencing data and glioma sphere-forming cell (GSC) RNA sequencing data, Wang et al. identified a set of genes that are uniquely expressed by glioma cells rather than tumor-associated host cells. With those gliomas' intrinsically expressed genes, the authors identified three gene expression subtypes within I*DH*-wild type GBMs: PN, MES, and CL. The subtype neural was excluded because its signals come from neurons. Moreover, a simplicity score was created to measure the intratumor heterogeneity by multi-subtype activation, which means one glioma sample with multiple gene expression subtypes is activated. Furthermore, this paper showed that these three different subtypes can differentially activate the immune environment, as mainly happens between the MES and non-MES samples. As reported before, MES samples have worse prognoses and lower tumor purity compared to non-MES samples.

1.1.2.2 DNA methylation classification

In 2010, TCGA project published a paper about GBM classification (12) using DNA microarray data. In this classification, a small subset of GBMs (about 5 to 10%) showed characteristics that distinguish them from other GBMs, such as overall hypermethylation, high *IDH1* and *IDH2* mutation rate, younger diagnostic age, and significantly longer survival time; these were named glioblastoma CpG island methylator phenotype (G-CIMP). CIMP, as a phenotype characterized as genome-wide gene promoter region hypermethylation, was first identified in colorectal cancer by Toyota et al. (13). They found methylation of cancer-specific clones exclusively existed in a subset of colorectal cancer. This kind of epigenetic aberration may be associated with tumor suppressor genes and can be used as biomarker. The relationship between G-CIMP and *IDH* mutation is further explained in the following *IDH* mutation section 1.1.3.1.

In 2018, a DNA methylation-based classification of all CNS tumors was published in Nature by Capper et al. (14). In this study, the researchers applied unsupervised analysis using

Illumina HumanMethylation450 (HM450k) DNA methylation data for over 100 CNS tumor types and then built a random forest classifier which could be used to classify new sample data (**Figure 1**). Results of this study are useful to help pathologists provide potential diagnoses in challenging cases and to help further avoid individual subjective effects of diagnoses. However, it is worth noting that Capper et al. did not classify samples using any specific biomarkers but using overall methylation profiles. Moreover, they covered all CNS tumor types, not just gliomas. These are the key differences between this project and my proposed project.

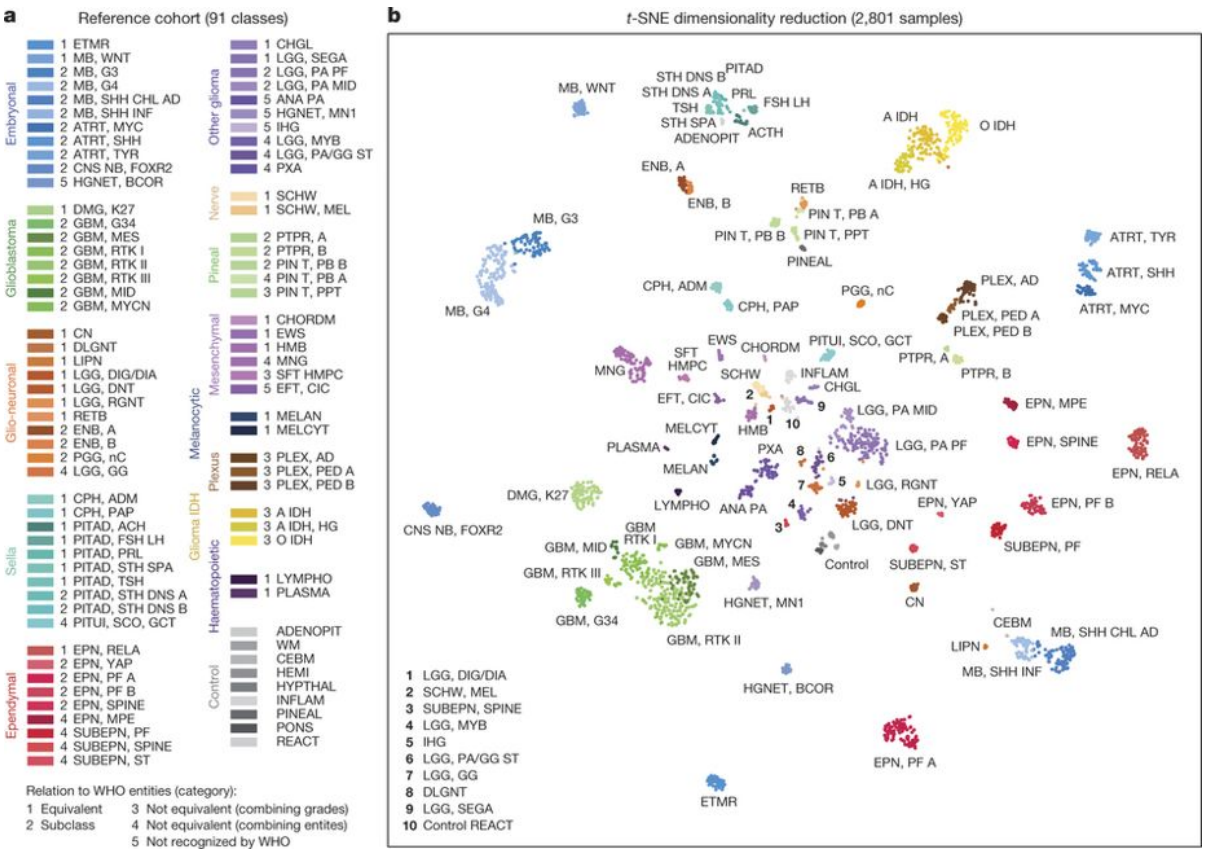


Figure 1: Establishing the DNA methylation-based CNS tumor reference cohort.

A Overview of the 82 CNS tumor methylation classes and nine control tissue methylation classes of the reference cohort. The methylation classes are grouped by histology and color-coded. Category 1 methylation classes are equivalent to a WHO entity, category 2 methylation classes are a subgroup of a WHO entity, category 3 methylation classes are not equivalent to a unique WHO entity with combining of WHO grades, category 4 methylation classes are not equivalent to a unique WHO entity with combining of WHO entities, and category 5 methylation classes are not recognized as a WHO entity. Full names and further details of the 91 classes are included in Supplementary Table 1. Embryonal tumors, shades of blue; glioblastomas, shades of green; other gliomas, shades of violet; ependymomas, shades of red; glio-neuronal tumors, shades of orange; IDH-mutated gliomas, shades of yellow; choroid plexus tumors, shades of brown; pineal region tumors, shades of mint green; melanocytic tumors, shades of dark blue; sellar region

tumors, shades of cyan; mesenchymal tumors, shades of pink; nerve tumors, shades of beige; haematopoietic tumors, shades of dark purple; control tissues: shades of grey. **B** Unsupervised clustering of reference cohort samples (n = 2,801) using t-SNE dimensionality reduction. Individual samples are color-coded in the respective class color (n = 91) and labelled with the class abbreviation. The color code and abbreviations are identical to **A**.

Note: Figure and legend from Capper, David, et al. "DNA methylation-based classification of central nervous system tumours." *Nature* 555.7697 (2018): 469. (14) with license number 4667250031818.

1.1.2.3 Diffuse glioma subtype classification

In 2016, TCGA consolidated all LGG and GBM samples (n = 1,122) and identified seven different subtypes with distinct biological and clinical characteristics (15). This classification no longer followed the traditional pathohistological grading rules but mixed LGG and GBM samples under the name diffuse glioma. All diffuse gliomas were first classified in *IDH* mutant or *IDH* wild type subgroups. Within the *IDH*-mutant subgroup, gliomas were further classified into G-CIMP-low, G-CIMP-high, and chr1p19q codel subgroup; within the *IDH* wild-type subgroup, gliomas were further classified into CL-like, MES-like, LGm6-GBM, and PA-like subgroups. This classification emphasized the effect of *IDH* mutation, which is believed to be an early event in tumor initiation and progress, and further showed the genomic similarity between *IDH* mutant LGG and GBM. In other words, according to the pathohistological grading, LGG and GBM are different grades of glioma. However, in terms of genetic similarity, they actually belong to the same subtype or may have the same origin if they harbor *IDH* mutation.

1.1.3 Somatic mutation

Generally, cancer is recognized as a genetic disease and genetic alterations and epigenetic alterations play important roles in cancer initiation, evolution, and prognosis. Genetic alterations usually refer to DNA sequence changes, or mutations. Mutations can be categorized into two different types: somatic and germline mutation. Somatic mutations usually occur in a single body cell and cannot be passed on to offspring, while germline mutations usually occur in gametes and can be passed on to offspring. A typical example of germline mutation is retinoblastoma, which mainly affects young children. Almost half of the retinoblastoma patients

have inherited mutated retinoblastoma protein 1 (*RB1*), which is a tumor suppressor gene, from their parents. In my dissertation, I will focus on the key somatic mutations for GBM, including *IDH*, *TERT*, and *ATRX* mutation.

1.1.3.1 Isocitrate dehydrogenase (*IDH*)

IDH is a family of genes which encode enzymes that can catalyze isocitrate to α-ketoglutarate (αKG). Three genes belong to this family: *IDH1* on chr2q32; *IDH2* on chr15q21; and *IDH3* on chr15q25. As shown in **Figure 2**, IDH1 located in the cytosol while IDH2 and IDH3 are located in mitochondria. To catalyze the conversion of isocitrate to αKG, IDH1 and IDH2 need nicotinamide adenine dinucleotide phosphate ($NADP^+$) and IDH3 needs nicotinamide adenine dinucleotide (NAD^+) as cofactors and therefore is a major pathway to generate NADPH and NADH.

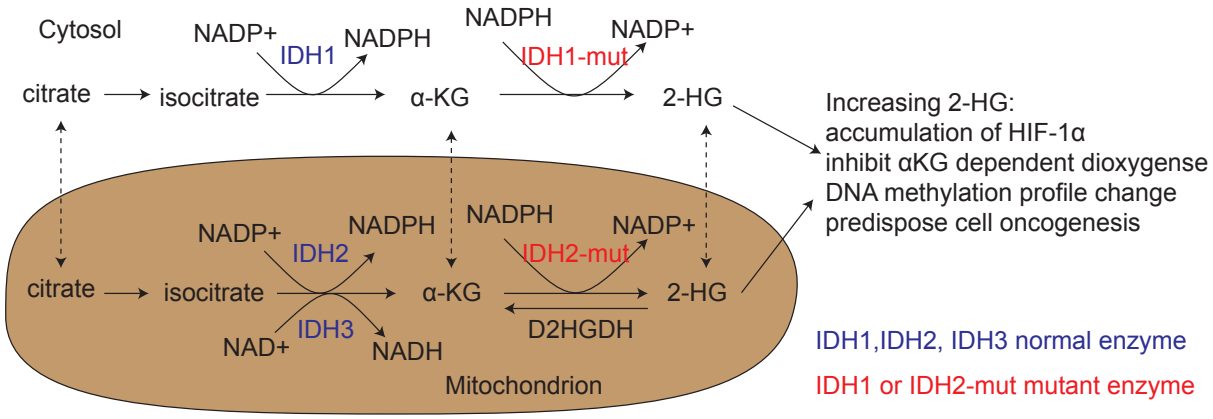


Figure 2: Wild-type *IDH* and mutant *IDH* functioning in cell.

For *IDH* loss function mutation, all mutations can be mapped to the key residue within the active site, which is critical for isocitrate binding, especially of codon R132 of *IDH1*, and codons R172 and R140 of *IDH2*. However, missense substitutions are quite flexible, which leads to the assumption that the location is more important than the substituted codons. The mutant IDH can lead to the decreased production of αKG and increased the production of 2- hydroxyglutarate (2-HG; **Figure 3**). Under homeostasis, 2-HG is usually maintained at a low level and can be

converted back to αKG through D-2-hydroxyglutarate dehydrogenase, mitochondrial (D2HGDH). However, the mutant IDH enzymes can convert αKG to 2-HG. It has been reported by Xu et al. (16) that 2-HG is a competitive inhibitor of αKG dependent dioxygenases, which include the histone demethylases and ten-eleven translocation (TET) family of 5-methlycytosine hydroxylases. Therefore, the accumulation of 2-HG can result in genome-wide DNA methylation and histone profile change (mainly increasing) and can block cellular differentiation, which is associated with tumor initiation and progression. Therefore, *IDH* mutation has been proposed as a potential target for cancer therapy. In fact, *IDH* mutation inhibitors such as AG-221 (17) (mutant *IDH2* inhibitor), AG-120 (18) (mutant *IDH1* inhibitor), and AG-881 (mutant *IDH1*/*IDH2* inhibitor) have been under evaluation and the preliminary data show promising results.

IDH gene mutations have been observed in many different cancer types, including acute myeloid leukemia (AML), GBM, intrahepatic cholangiocarcinoma, and others. Both *IDH1* and *IDH2* mutations are found related to glioma, and the most common mutation happens in *IDH1* at amino acid residue 132, which results in arginine being replaced with histidine (R132H). About 80% of LGGs and secondary GBMs are found to harbor *IDH* mutation, while only a small portion (about 5 to 10%) of primary GBMs have *IDH* mutation. This group of GBMs with *IDH* mutation show upregulated methylation profiles (**Figure 3**), younger age at diagnosis, longer survival time, and the majority of them are harboring *IDH* mutations. GBMs belong to this subgroup were named as G-CIMP samples. Also, this subgroup of primary GBM samples has been found showing PN subtype genetic characterization, which is more similar to LGG samples than typical GBM samples. The causal relation between *IDH* mutation and G-CIMP has been proven by Sevin et al. (19). By constructing isogenic astrocytes expressing either *IDH* R132H mutation, wild type *IDH*, or neither, the 2-HG production level and methylome profile were compared. It was found that *IDH* R132H mutation astrocytes show high expressed 2-HG and overall methylation level increased by cell passages.

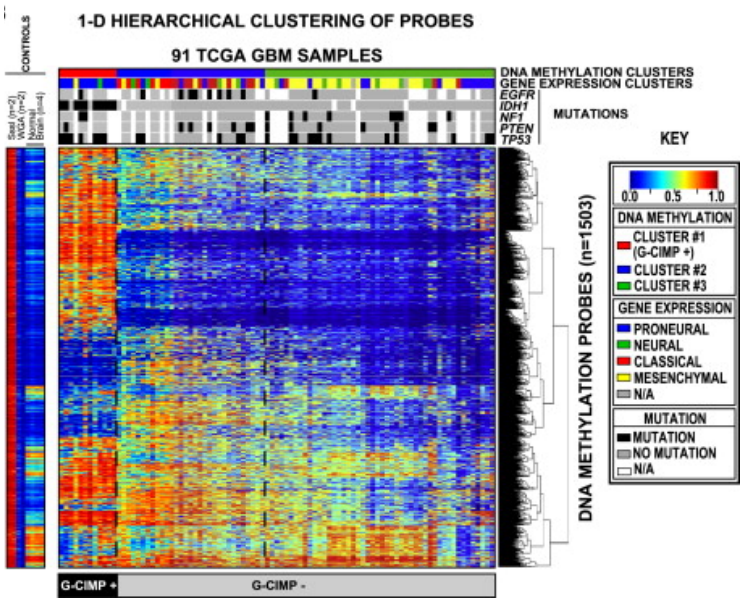


Figure 3: Unsupervised clustering analysis of GBM DNA methylation profile.

One-dimensional hierarchical clustering of the same 1503 most variant probes. Each row represents a probe; each column represents a sample. The level of DNA methylation (beta value) for each probe, in each sample, is represented with a color scale as shown in the legend; white indicates missing data. M.SssI-treated DNA (n = 2), WGA-DNA (n = 2), and normal brain (n = 4) samples are included in the heatmap but did not contribute to the unsupervised clustering. The probes in the eight control samples are listed in the same order as the y axis of the GBM sample heatmap.

Note: Figure and legend from Noushmehr, Houtan, et al. "Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma." *Cancer cell* 17.5 (2010): 510-522. (20) with license number 4667300694903.

IDH mutation has been used as a prognostic biomarker for GBMs that show better survival rates. Many hypotheses have been proposed to explain these better prognoses. For example, it has been suggested that cancer cells with reduced NADPH production are more vulnerable to irradiation and chemotherapy because of their decreased capacity to scavenge reactive oxygen species generated during irradiation and chemotherapy (21). Moreover, *IDH* mutation may lead to a suppressed immune response to glioma through down-regulated leukocyte chemotaxis, which in turn reduces the aggressiveness of tumors and leads to longer survival (22). Moreover, lower immune cell infiltration in *IDH* mutant glioma may also contribute.

1.1.3.2 Telomerase reverse transcriptase (*TERT*)

Telomere is a special heterochromatin structure located at the end of the linear chromosome. It consists of high GC DNA sequences and is enclosed by specific proteins. During the cell division process, the telomere loses some base pairs because the RNA primer is involved in the DNA replication process. The RNA primer will bind with DNA sequences, but this region cannot be replicated. As the cell division progresses, telomere gets shorter and shorter, and once it reaches a certain length, the cell will start apoptosis or stop further division. Telomere works like a clock to maintain the length and stability of chromosomes, self-renewal, and proliferation.

Two major mechanisms are involved in telomere length maintenance in cancer: telomerase activation, or upregulation (85%-90%) and the alternative lengthening of telomeres (ALT) pathway. These two mechanisms are mutually exclusive in most situations but may coexist in rare situations, and the switch between telomerase upregulation and ALT has also been observed. Two major components make up the telomerase enzyme: hTR, the RNA component, which provides the template for DNA synthesis, and hTERT, the protein that catalyzes components and adds the new DNA segments to the ends of chromosomes. hTR is encoded by the gene telomerase RNA component (*TERC*), located on chr3q26, and hTERT is encoded by the gene *TERT*, located on chr5p13. Both components are essential for telomere maintenance. *TERT* is usually silenced in normal cells but is activated in cancer cells. Different mechanisms can activate or upregulate the hTERT expression including promoter mutation, alteration in alternative splicing of pre-mRNA, gene amplification, epigenetic changes, and others (23). I will focus on *TERT*p mutation in this section.

*TERT*p mutation has been observed in more than 50 cancer types and is a major mechanism to activate the telomerase in cancer cells. For example, the *TERT*p mutation rate is 43% in CNS tumors (~80% for GBM), 59% in bladder cancer, and 29% in skin cancer (~70% in melanoma; (24). Two *TERT*p mutation hotspots have been reported: C228T and C250T, which are located -124bp and -146bp upstream from the *TERT* transcriptional start site (TSS). Studies suggested that promoter region mutation can lead to the change of binding preferences and

adding additional binding sites (**Figure 4**) (23). The telomerase expressions are higher in cancer samples with *TERT*p mutation compared to *TERT*p wild type.

In general, *TERT*p mutation has been observed to be associated with decreased overall survival rates. However, the coexistence with other prognostic biomarkers, such as *IDH*, *MGMT*, and chr1p19q codel, should be considered for GBMs when evaluating the prognostic effects. The molecular classification of gliomas according to *TERT*p mutation is useful and powerful. Anti-telomerase therapeutics focus on inducing apoptosis and cell death in cancer cells while minimizing the risk of telomere shortening in normal cells. Different approaches have been adopted such as hTERT- or hTR-targeted antisense oligonucleotides and small molecule inhibitors (23).

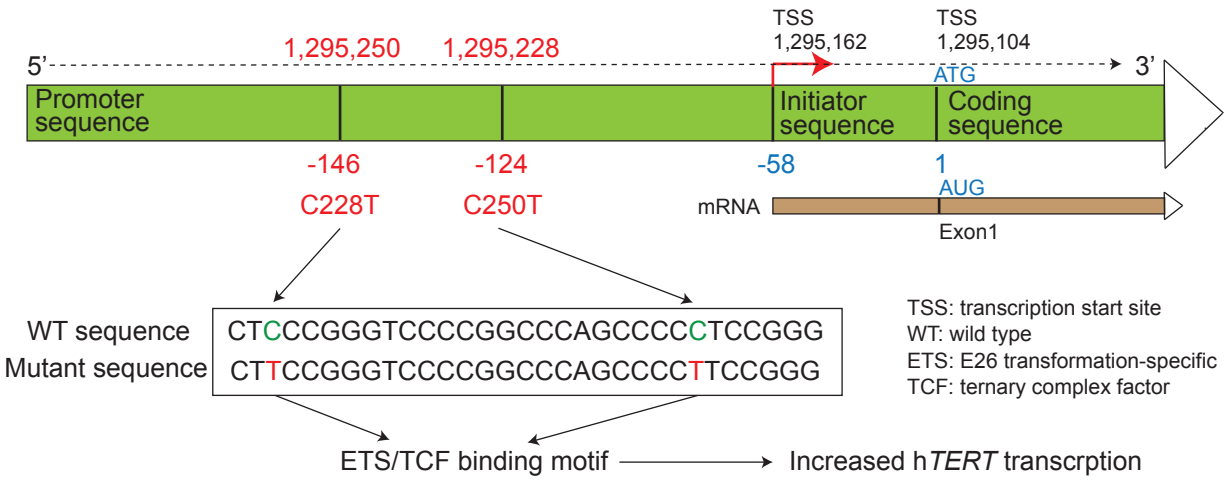


Figure 4: *TERT* transcription and promoter mutations.

Two *TERT*p mutation hotspots are showed in the figure: C250T and C228T. The transcription is regulated by transcription factors. *TERT*p mutations create E26 transformation-specific/ternary complex factor (ETS/TCF) binding motifs. Then more transcription factors can bind to their respective sites and promote *TERT* transcription.

1.1.3.3 Alpha thalassemia/mental retardation syndrome X-linked (*ATRX*)

Other than telomerase, ALT is another major mechanism to maintain telomere length in cancer cells. It is hypothesized that the ALT process involves homologous recombination (HR)-mediated DNA replication; however, the precise mechanism is still less understood (**Figure 5**). ALT cells are believed to be highly heterogeneous and usually have fluctuating telomere lengths,

abundant extrachromosomal telomeric repeat DNA (ECTR), high levels of telomere sister chromatid exchanges (t-SCES), and ALT-associated promyelocytic leukemia (PML) bodies (APBs) that have a specialized telomeric DNA nuclear structure (25). Though *TP53* inactivation has been widely observed in many cancer types, it is believed to be necessary for ALT cancer cells (26). About 10% to 15% of human cancers utilize the ALT rather than the telomerase-dependent mechanism to maintain their telomere length. Moreover, studies have reported that mesenchymal origin cancers are more likely to utilize the ALT mechanism and are less likely to express telomerase (27, 28).

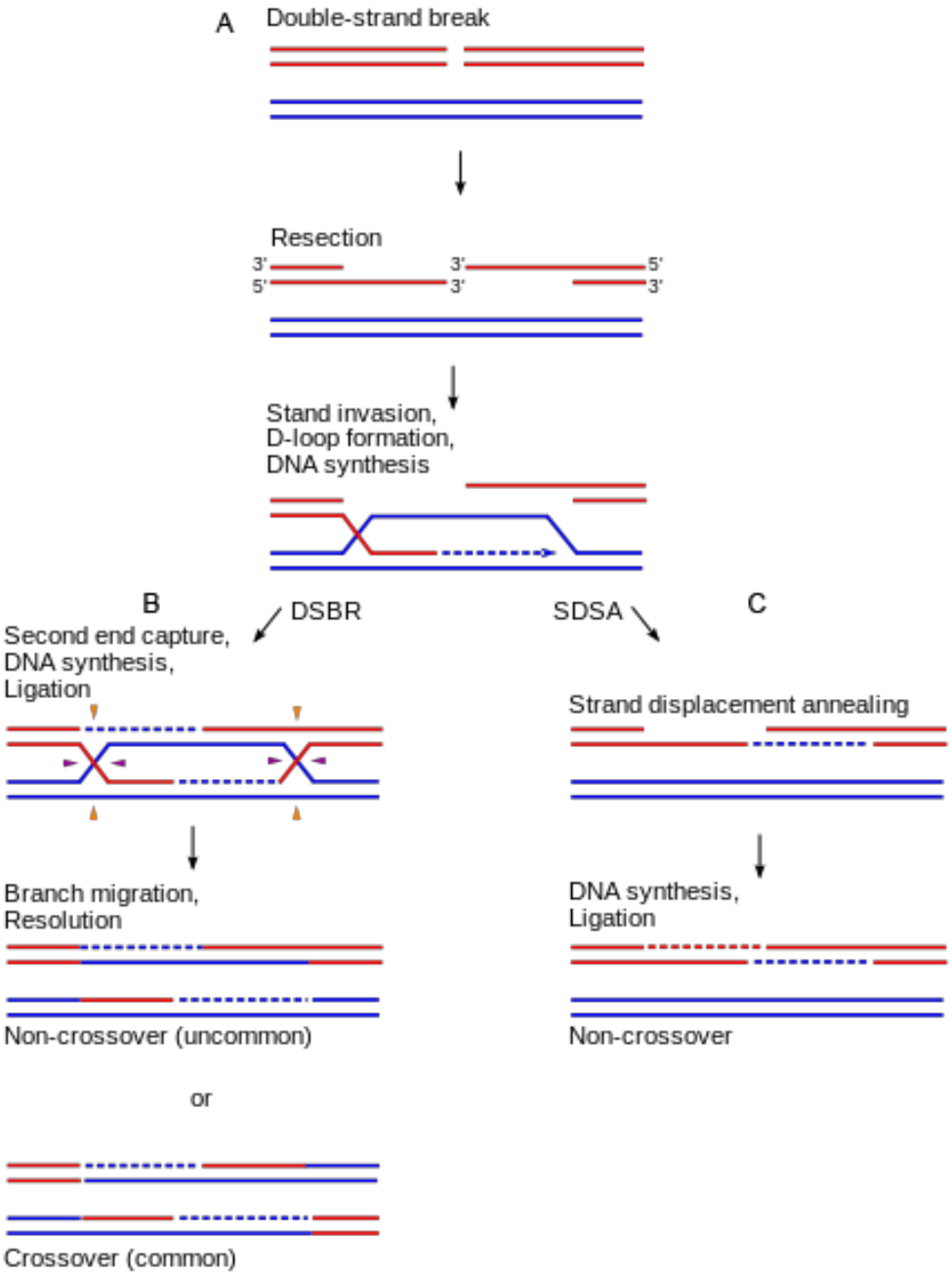


Figure 5 Repair of DNA double-strand breaks by DSBR and SDSA.

Double-strand breaks (DSBs) can be repaired by several HR-mediated pathways, including double-strand break repair (DSBR) and synthesis-dependent strand annealing (SDSA). **A** In both pathways, repair is initiated by resection of a DSB to provide 3′ single-stranded DNA (ssDNA) overhangs. Strand invasion by these 3′ ssDNA overhangs into a homologous sequence is followed by DNA synthesis at the invading end. **B** After strand invasion and synthesis, the second DSB end can be captured to form an intermediate with two Holliday junctions (HJs). After gap-repair DNA synthesis and ligation, the structure is resolved at the HJs in a non-crossover (black arrow heads at both HJs) or crossover mode (green arrow heads at one HJ and black arrow heads at the other HJ). **C** Alternatively, the reaction can proceed to SDSA by strand displacement,

annealing of the extended single-strand end to the ssDNA on the other break end, followed by gap-filling DNA synthesis and ligation. The repair product from SDSA is always non-crossover.

Note: Figure and legend are from Sung, Patrick, and Hannah Klein. "Mechanism of homologous recombination: mediators and helicases take on regulatory functions." *Nature reviews Molecular cell biology* 7.10 (2006): 739. (29) with license number 4667681409176.

Mutations of *ATRX*, located on chrXq21, or death-domain associated protein (*DAXX*), located on chr6p21, have been frequently observed in cancer cells with the ALT phenotype. The gene *ATRX* encodes SWItch/Sucrose Non-Fermentable (SWI/SNF) like chromatin remodeling protein and therefore have the chromatin remodeling function (30): with the ATRX-DNMT3-DNMT3L (ADD) domain and HP1α, *ATRX* can bind and regulate genomic regions enriched for H3K9me3 (31). It can also regulate transcription by recruiting enhancer of zeste 2 polycomb repressive complex 2 subunit (EZH2) to deposit H3K27me3. By binding DNA through its transcription regulatory protein SNF2 N-terminal helicase domain, *ATRX* can resolve tandem repeats and further promote transcription at the α-globin locus. *ATRX* can deposit H3.3 with *DAXX* to repetitive regions and negatively regulate macroH2A (mH2a) deposition at the α-globin locus and telomeres. *DAXX* is a highly conserved protein associated with nuclear and cytoplasmic events during apoptosis. *ATRX*/*DAXX* complexes function as histone chaperone to deposit the histone variant H3.3 to repetitive heterochromatin. It has been proposed that the loss function mutations of either *ATRX* or *DAXX* coexist with *IDH* and *TP53* mutation in gliomas and are mutually exclusive with far upstream element binding protein 1 (*FUBP1*), capicua transcriptional repressor (*CIC*) and chr1p19q codel.

Many theories have been proposed to explain how the loss function of *ATRX* mutation affects the ALT. Clynes et al. (32) proposed that with loss function mutation of *ATRX* or *DAXX*, the H3.3 cannot be deposited at the G-rich repeats in the telomere region, which leads to the formation of a G4 structure, replication fork stalling, and the HR of telomeres and finally the ALT. O'Sullivan et al. (33) proposed that ALT depends on the replicative stress by ATR serine/threonine kinase (ATR) and that *ATRX* can recognize G-quadruplexes (G4) structure and prevent replicative stress. Therefore, *ATRX* loss function is required for cancer cells to maintain

the ALT phenotype. G4 structure is a secondary DNA structure that can lead to replicative stress or block transcriptional processes, which is formed by the G-rich tandem repeats upstream of the alpha-globin genes. In summary, the loss function of *ATRX* mutation is a prerequisite of the ALT mechanism.

1.1.4 Copy number variation (CNV)

Copy number variation is another key genomic feature of cancer cells and involves gain or loss of genomic DNA. Currently, the most commonly used high-throughput technology to measure the copy number variation is the single nucleotide polymorphism (SNP)-array, for example, the Affymetrix 6.0 platform (SNP6) has been used in TCGA project. On this microarray, probes are designed to test the genotype of the SNP and the probe signal intensities are used to calculate the copy number. It is assumed that the majority of the genome is diploid and it is used as the baseline for signal intensities normalization. For example, if a certain genome region has amplification, then all probes located within this region are expected to have higher signal intensities no matter which genotype they are and vice versa. After data cleaning and noise reduction, circular binary segmentation (CBS) (34) is used to call the copy number. This method evolved from binary segmentation, which only considers one change-point at a time, while the CBS considers two change-points. Assuming the arc from *i+1* to *j* and its complement have different means of log-ratio normalized intensities gives the following equations:

$$Z_{i,j} = \{\frac{1}{j-i} + 1/(n-j+1)\}^{-1/2}\{\frac{S_j - S_i}{j-i} - (S_n - S_j + S_i)/(n-j+i)\}$$

$$Z_c = max_{1 \le i < j \le n}|Z_{i,j}|$$

The null hypothesis is rejected if the statistic exceeds the critical value based on the null hypothesis and the critical value can be computed using Monte Carlo simulation or the approximation of tail probability when X_i follows normal distribution. This procedure is applied recursively to search for all copy number breakpoints.

The role of CNV as a risk factor for cancer is underestimated. The changes of copy number are commonly associated with the deletion of tumor suppressor genes or amplification

of oncogenes. For example, CL GBM has its special CNAs, which are characterized as chr7 amplifications and chr10 deletions involving the change of *EGFR* and phosphatase and tensin homolog (*PTEN*).

1.1.4.1 Chromosome 1p19q co-deletion (chr1p19 codel)

A subgroup of *IDH* mutant gliomas show chr1p19q codel: the complete deletion of both the short arm of chr1 and the long arm of chr19. This genomic feature has nearly always been observed only in oligodendroglioma. The loss of one chromosome arm each in chr1 and chr19 leads to the loss of heterozygosity (LOH). Chr1p19q codel is caused by an unbalanced whole-arm translocation between chr1 and chr19, which happens in the early stage of cancer initiation. It usually coexists with *IDH*, *TERT*p, *CIC* and *FUBP1* mutation and is mutually exclusive to *TP53* and *ATRX* mutation, except few rare cases. Though the reason why this specific deletion happens is unknown, it is an independent prognostic biomarker associated with better survival durations. Studies have shown that patients with *IDH* mutation and chr1p19q codel have a median overall length of survival of 8.0 years, while patients with *IDH* mutation and chr1p19q intact have a median length of survival of 6.3 years. In addition to its prognostic value, chr1p19q codel can also be used to predict the anaplastic oligodendrogliomas chemotherapy response. Randomized controlled clinical trials comparing the procarbazine/lomustine/vincristine (PCV) chemotherapy plus radiotherapy to radiotherapy alone have shown survival benefits for first-line chemotherapy in oligodendroglioma with chr1p19q codel (35, 36). With all these clinical values, chr1p19q codel has been included in the new WHO classification criteria. There are many different methods to detect the chr1p19q codel and fluorescent in situ hybridization (FISH) is one of the most widely-used methods. With the current high-throughput copy number evaluation methods available, chr1p19q codel can also be identified by checking the chromosome arm-level copy number alterations.

1.1.4.2 Other genomic features

Genomic alterations, including mutation and copy number variation (CNV), are essential components in cancer etiology and progression and have been used for cancer diagnosis, classification, and as therapeutic targets. Epigenetic disruptions can interact with genetic alterations and contribute to abnormal cancer genomes. For example, tumor suppressor gene inactivation by epigenetic silencing is typically mutually exclusive of gene inactivation by mutation or deletion as exemplified by cyclin dependent kinase inhibitor 2A (*CDKN2A*) in lung squamous cell carcinoma (37) and breast cancer 1, early onset (*BRCA1*) in breast and ovarian cancer (38). In addition to single gene suppression, CIMP is another known association between somatic mutation and epigenetic signature that usually presents as hypermethylation within CpG site-enriched regions, such as gene promoters. The first identified CIMP was in colorectal cancer (13) and the CIMP-high subgroup has been associated with mutation of v-raf murine sarcoma viral oncogene homolog B at V600 ($BRAF^{V600E}$; (39), though the molecular mechanism is still unclear. The CIMP subgroup has also been identified in many other tumors (40, 41) (42), including GBM (12). G-CIMP patients show a strong association with *IDH1* and *IDH2* mutation and demonstrate longer survival compared to patients with non-G-CIMP tumors.

1.1.4.3 MGMT promoter methylation

The *MGMT* gene is located on chromosome 10q26.3 and encodes the DNA repair protein, which is involved in cellular defense against mutagenesis and toxicity from alkylating agents. The protein can transfer the methyl group from O^6-alkylguanine to its own molecule, thus enhancing the fatal effects of alkylating agents. In contrast, defective *MGMT* will cause the base mis-repairing and mismatch repair failure during DNA replication and finally lead to cell cycle arrest and apoptosis. Therefore, *MGMT* promoter methylation will lead to decreased level of functional MGMT protein and inadequate repair of DNA alkylation during chemotherapy with an alkylating agent.

About 30 to 50% GBM shows *MGMT* promoter methylation and it can be used as an efficient biomarker for alkylating chemotherapy response prediction, including temozolomide (TMZ). Patients who are treated with TMZ and with *MGMT* promoter methylation have longer survival times compared to patients who are treated with TMZ without *MGMT* promoter methylation (5). In addition to this predictive value, *MGMT* promoter methylation also has prognostic value: patients with *MGMT* promoter methylation have better survival rates regardless of whether they are treated with TMZ and radiation or radiation alone (43, 44). The most widely used methods to measure *MGMT* promoter methylation are the methylation-specific PCR (MS-PCR; (45) and the methylation sensitive-quantitative locked nucleic acid PCR (MS-qLNAPCR). With DNA methylation microarray data, Bady et al. (46) built a logistic regression model (MGMT-STP27) to predict *MGMT* promoter methylation status using two probes available on the BeadChip.

1.2 Epigenetics and DNA methylation

Epigenetics play a crucial role in cancer (47, 48) and lead to extensive reprogramming of cancer mechanisms through DNA methylation, histone variation (49), and noncoding RNA (50). As one of the most well-known epigenetic mechanisms, DNA methylation is a stable feature and can reflect both inter- and intratumor heterogeneity. It has been used to classify different types of tumors (13, 39, 51-53). For example, the recently published DNA methylation-based histopathological classification of CNS tumors (14) has challenged the conventional histologic classification and tumor grading. DNA methylation also provides clarity in unknown primary tumor classification (54). Another important application of DNA methylation is developing predictive and prognostic biomarkers (55), such as for tumor recurrence (56), treatment response and patient survival (57).

1.2.1 DNA methylation

1.2.1.1 DNA methylation definition and types

DNA methylation is the process of adding a methyl group to DNA molecules. Usually, it happens to the cytosine of a CpG site that represents cytosine-phosphate-guanine (**Figure 6**). This transfer process is catalyzed by DNA methyltransferase enzymes (DNMTs). Therefore, if enzyme function is affected, the DNA methylation profile will change accordingly. This mechanism can happen in multiple cytosine positions, for example, 5-Methylcytosine (5-mC), 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC), 5-carboxylcytosine (5-caC), and 3-methylcytosine (3-mC) (**Figure 7**). These are differentiated by the position where the methyl group is added to the cytosine and the added methyl group. The most commonly seen is 5-mC, which is also what I targeted in the Illumina DNA methylation microarray. A CpG island indicates the region where enriched for CpG sites. It is common to see a CpG island in the gene promoter region.

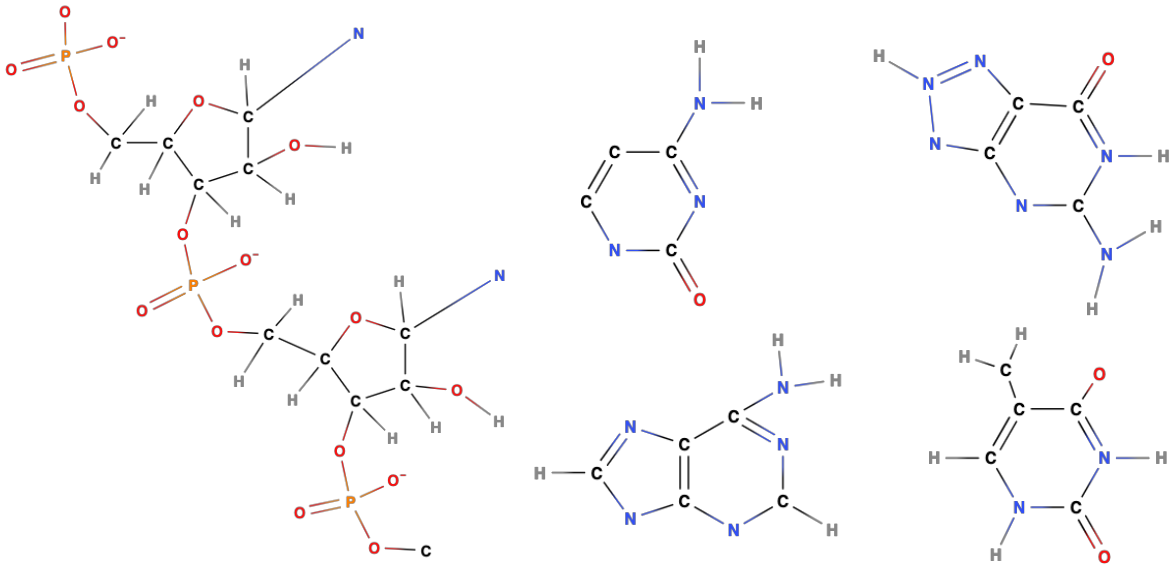


Figure 6: DNA single strand structure and nucleotide structure.

Left panel shows the bone of single-strand DNA molecules structures. The blue N represents the nucleobase and the difference of base which determine the nucleotide A, T, G or C. The right panel show the structure of four different nucleotides in DNA: top left: Cytosine (C); top right: Guanine (G); bottom left: Adenine (A); bottom right Thymine (T).

1.2.1.2 DNA methylation functions

DNA methylation is the major epigenetic mechanism and plays an important role in regulating cell stage and differentiation. The most well-known function of DNA methylation is silencing the gene expression by inhibiting the binding of a transcription factor to gene promoter regions.

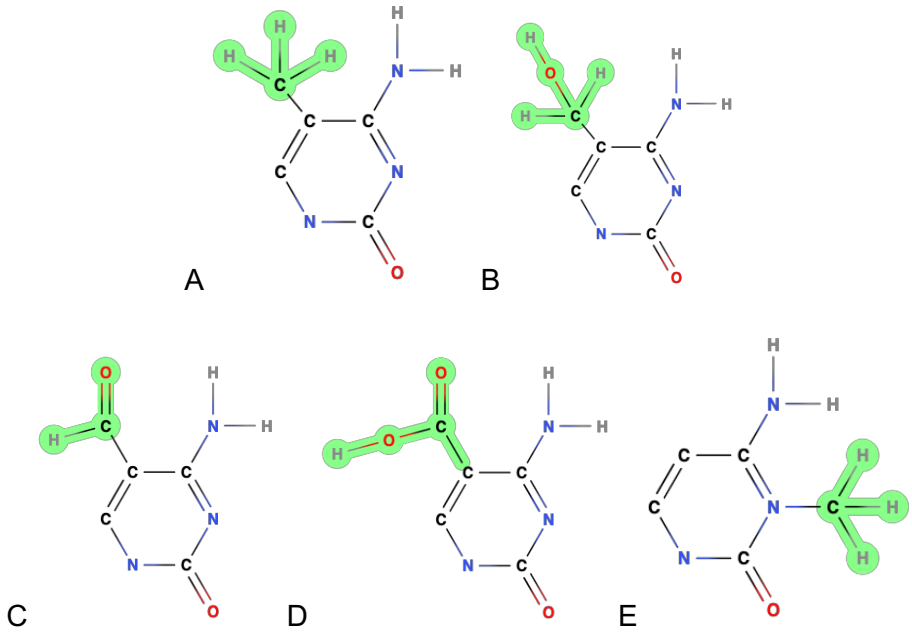


Figure 7: Different type of DNA methylation structure.
The green highlighted part is the adding group to cytosine. A: 5-mC; B: 5-hmC; C: 5-fC; D: 5-caC; E: 3-mC.

1.2.1.3 DNA methylation testing methods

DNA methylation can be measured in many different ways, and how to choose the appropriate method depends on research interests (**Figure 8**). For example, MS-PCR, or pyrosequencing, is helpful if the study is concerned with only a few positions while sequencing-by-synthesis, representation bisulfite sequencing (RRBS), or microarray is more suitable for global analysis, genome-wide methylation profile. Bisulfite conversion is a key step in measuring

the level of DNA methylation. Bisulfite conversion is a process whereby genomic DNA is treated with sodium bisulfite, leading to the deamination of unmethylated cytosines into uracil (U), while the methylated cytosines are protected by the methyl group and stay unchanged during the treatment (**Figure 9**).

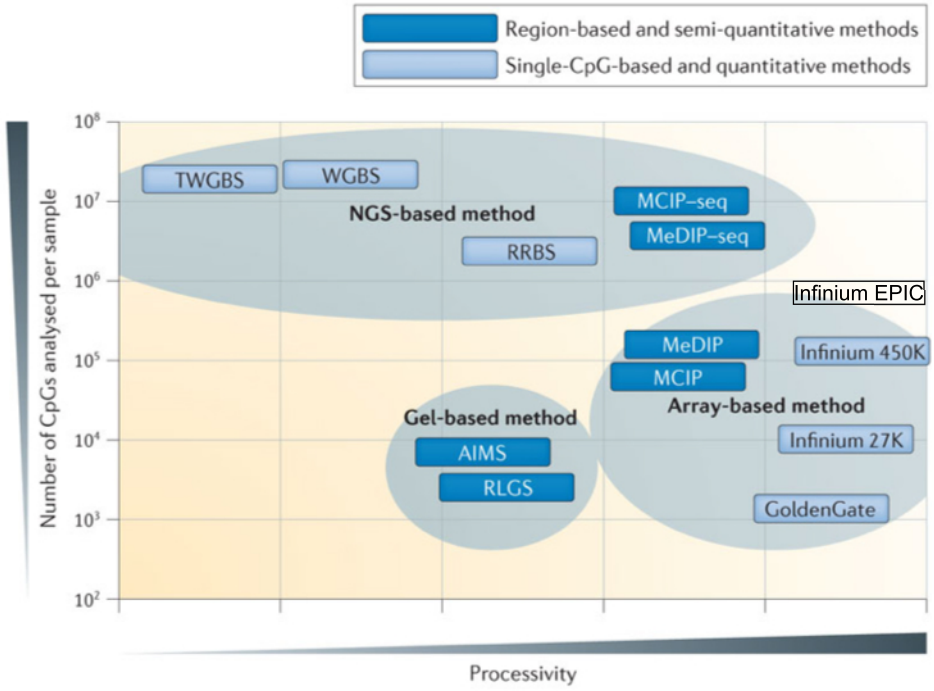


Figure 8: DNA methylation profiling technologies.

DNA methylation profiling technologies ordered based on the 'processivity' of a technique (*x* axis), as measured by an estimate of the total number of samples analysed, and the number of CpGs that can be analysed per sample (*y* axis). Processivity was measured based on published data, but it also reflects the cost per assay, the time for post-processing of data and the ease of handling.

Note: Figure and legend adapted from Plass, Christoph, et al. "Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer." *Nature reviews genetics* 14.11 (2013): 765. (58) with license number 4671380789944.

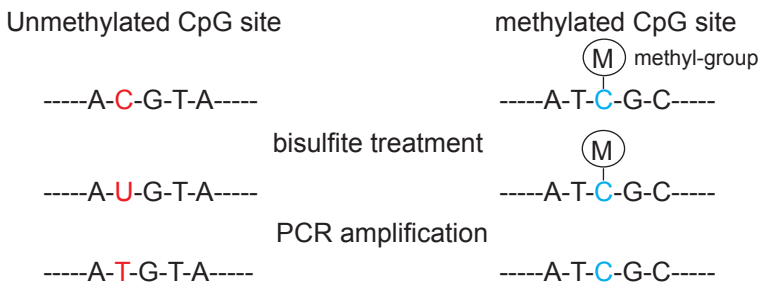


Figure 9: Bisulfite conversion process.

Left panel shows unmethylated CpG site (red) and the right panel shows methylated CpG site (blue). After bisulfite conversion, the left cytosine has been changed to uracil and then replaced by T during amplification; the right cytosine is protected by the methyl group and keep as cytosine.

1.2.2 DNA methylation microarray.

Illumina HumanMethylation (HM) microarray is the most commonly used microarray platform for human genome-wide DNA methylation profiles. For each targeted CpG position, there are corresponding beads with specific-designed oligos attached to them. From the first platform (GoldenGate), HumanMethylation 27 (HM27k), to the current platform (EPIC), the number of probes available on the array and covered genes has exponentially increased (**Table 4**). The platform available now is called Illumina HM EPIC array, and it has more than 850,000 probes available. It is worth noting that more than 90% of probes available with the previous platform, HM450k, are also available in the current EPIC array. This continuity of available probes gives researchers the chance to merge data from different platforms and increase the sample size efficiently. More importantly, this platform can be used to formalin-fixed paraffin-embedded (FFPE) samples, which is the most common way for clinical sample storage. Depending on storage methods, a sample's DNA may degrade at different levels; therefore, a restoration process is needed before bisulfite conversion treatment.

Table 4: DNA methylation microarray evolution

Platform	availability	number of probes	number of covered genes	Infinium probe type
GoldenGate	discarded	1536	371	-
HM27k	discarded	27578	14495	Infinium I
HM450k	discarded	485764	21231	Infinium I and II
EPIC	available	867531	277365	Infinium I and II

1.2.2.1 Microarray probe

The Infinium probe design has been used since the HM27k platform. There are two different types of Infinium probes: Infinium I and II (**Figure 10**). For the Infinium I type, two probes are needed for each targeting position: one for methylated and one for unmethylated; for the Infinium II type probe, one probe is enough. This helps to put more probes on the limited space on the chip. Probes are single strand oligos (length = 50; base pair, bp) with designed DNA sequence which can hybridize with input DNA fragments. Probes are expected to hybridize with targeted DNA sequences, but mis-hybridization can happen. This leads to the so-called multi-hit probes, which refers to probes that hybridize with multiple DNA sequences from different chromosome regions (59). Besides the multi-hit problem, the SNP may also affect the probe hybridization (60). Those probes may cause errors during the hybridization process and further affect the methylation value. Therefore, it is suggested to exclude them in downstream analysis.

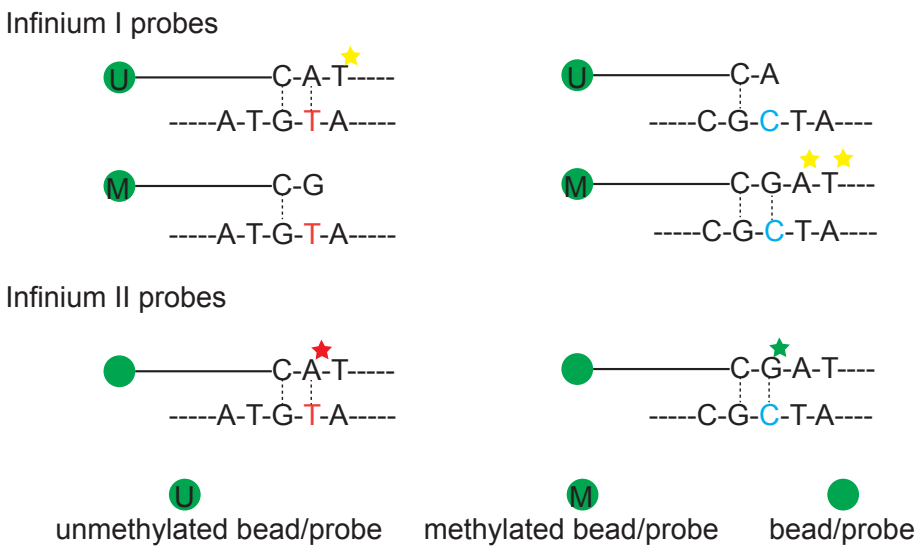


Figure 10: Infinium type I and II probes.

The top panel shows the Infinium I probe and the bottom panel shows the Infinium II probes. There are two probes for each targeted position in the Infinium I probe: the unmethylated bead/probe and the methylated bead/probe. For the unmethylated bead/probe, single strand oligos end with CA, which can hybridize with unmethylated CpG sites and prolong the hybridization with the signal (the yellow stars) while stopping the hybridization with methylated CpG site without a detected signal. For the methylated bead/probe, single strand oligos end with CG, which can hybridize with the methylated CpG sites and release the signal (the yellow stars) while stopping the hybridization with the unmethylated CpG sites. In contrast, Infinium type II only has one type of probe for both methylated and unmethylated CpG sites. The single strand oligos end with C, which enable hybridization with both unmethylated and unmethylated CpG sites. Then either A (unmethylated CpG sites) or G (methylated CpG sites) are attached to prolong the hybridization and release different colored signals. Then the signal intensities are used to calculate the methylation level.

1.2.2.2 Infinium microarray protocol and controls

Illumina Infinium array usually takes three days to run. Some steps involve internal controls: those probes were designed to support quality control (QC) of the assay's stringent performance criteria and to demonstrate its robustness. Built-in control probes can help to identify samples for which data characteristics are significantly different than expected and may need to be excluded for further analysis. They can also provide some information about the experiment's steps. There are two types of internal probes: sample-dependent and sample-independent (**Figure 11**, **Table 5**).

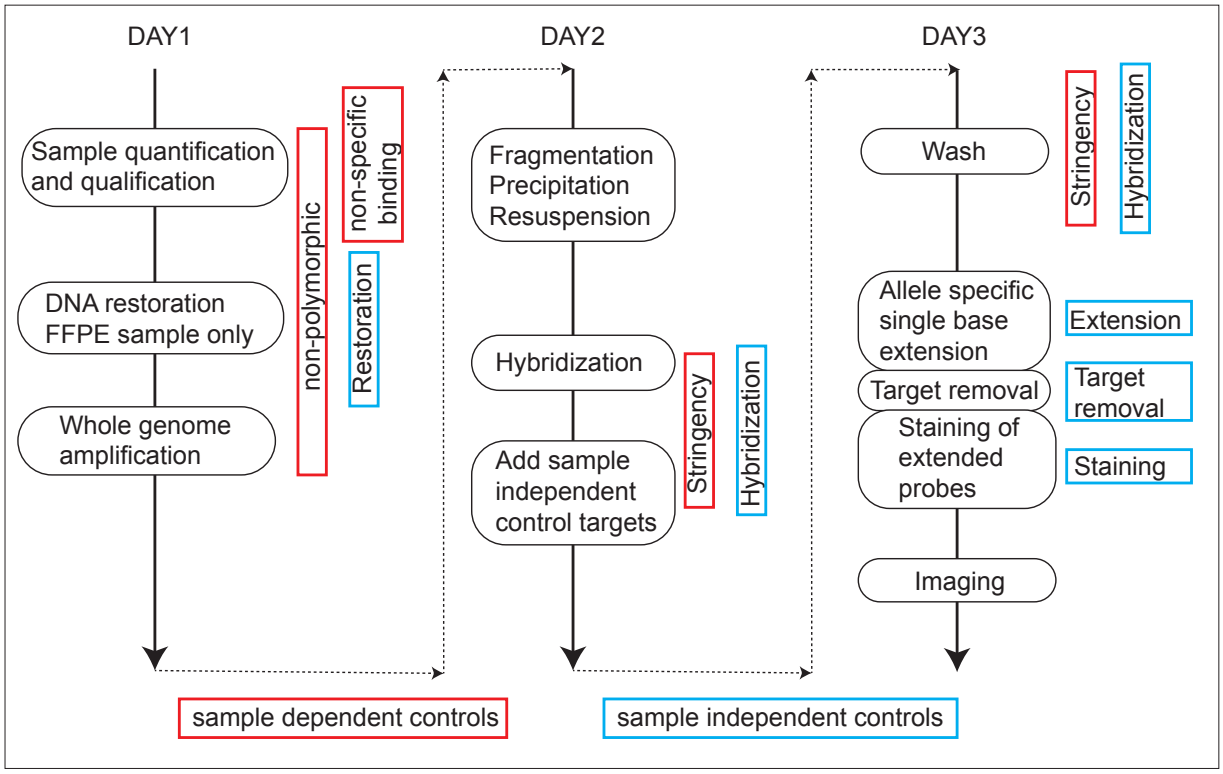


Figure 11: Overview of Infinium microarray protocol and build-in controls

All built-in controls are shown within the box; red boxes indicate sample-dependent controls and blue boxes indicate sample-independent controls. Sample-dependent control probes are used to evaluate sample quality and performance (red boxes) including nonpolymorphic controls, nonspecific binding controls, and stringency controls. Sample-dependent stringency and nonpolymorphic control probes are designed for human DNA and are not informative when working with nonhuman samples. Sample-independent control probes are used to evaluate bead chip and reagent performance and efficiency of hybridization and staining processes, including restoration, hybridization, extension, target removal, and staining controls. The restoration step is applied and evaluated for FFPE samples only.

Table 5: Number of internal control probes in Infinium bead chip

Type	controls	#probes in HM450k	#probes in EPIC
sample-independent controls	staining	4	6
	extension	4	4
	hybridization	3	3
	target removal	2	2
sample-dependent controls	bisulfite conversion I	12	10
	bisulfite conversion II	4	4
	specificity I	12	12
	specificity II	3	3
	negative	613	411
	nonpolymorphic	4	9
sample-dependent	norm_A	32	27
	norm_T	61	58
	norm_C	61	58
	norm_G	32	27

Note: Table adapted from Illumina technology sheet.

For all those internal control probes, researchers use probe methylation and unmethylation intensity as metrics. For all probes, intensity is evaluated using both the red and green channels. Based on the Illumina BeadArray control reporter (BACR), specific criteria are applied (**Table 6**).

Table 6: Internal control criteria

Control	criteria	note
staining green biotin high > biotin Bkg	(high/biotin Bkg) > 5	Threshold can be increased green channel-use lowest C or G and highest A or T intensity red channel-use lowest A or T and highest C or G intensity
staining red DNP high > DNP Bkg	(high/DNP Bkg) > 4	
Extension green lowest CG/Highest AT	(C or G/A or T) >5	
Extension Red Lowest AT/Highest CG	(C or G/A or T) >5	
Hybridization Green high > Medium > Low	(High/Med)>1 and (Med/Low)>1	bkg=extension green highest A or T intensity
Target removal Green ctrl1 <= Bkg	((bkg+x)/ctrl) >1	
Target removal Green ctrl2 <= Bkg	((bkg+x)/ctrl) >1	
Bisulfite conversion I green C1,2,3 > U1,2,3	(C/U)>1	Threshold can be increased. use lowest C intensity use highest U intensity green channel- bkg=extension green highest AT red channel-bkg=extenstion red highest CG
Bisulfite conversion I green U <= bkg	((bkg+x)/U)>1	
bisulfite conversion I red C4,5,6>U4,5,6	(C/U)>1	
bisulfite conversion I red U<= bkg	((bkg+x)/U)>1	
Bisulfite conversion II C red > C green	(C red/C green)>1	
bisulfite conversion II C green <= bkg	((bkg+x)/C green)>1	
Specificity I green PM>MM	(PM/MM)>1	use lowest PM intensity use highest MM intensity bkg=extension Green highest A or T intensity
Specificity I red PM>MM	(PM/MM)>1	
Specificity II S red > S green	(S red/S green)>1	
specificity II S green <= bkg	((bkg+x)/S green)>1	

1.2.3 DNA methylation data

For Illumina DNA microarrays, the raw data is saved as IDAT format which is human-unreadable. After data processing, researchers can obtain both the methylated and unmethylated signal intensities for each probe available on the bead chip.

1.2.3.1 Methylation metrics

The methylation level can be measured with two values: the β-value and the M-value, and these can be converted into each other. The β-value has a range of 0 to 1; β-value closer to 0 indicate less methylation, and β-value closer to 1 indicate more methylation. Because this microarray is used with bulk tissue, it is rare to observe exactly 0 or 1 for methylation value. In

contrast, β-values usually show two peaks in terms of their distribution: one around 0.1 and the other around 0.9 (**Figure 12**). The M-value is the log transformation of methylation intensities and its range is usually between -6 and 6 (**Figure 12**). A comparison (61) between these two metrics leads to the conclusion that it is better to use the M-value for data processing and calculation and to use the β-value for data visualization and demonstration.

$$\beta - \text{value} = \frac{\text{methylated signal}}{\text{methylated} + \text{unmethylated signal} + \alpha}$$

$$\text{M} - \text{value} = \log2(\frac{\text{methylated signal} + \text{c}}{\text{unmethylated signal} + \text{c}})$$

$$\beta - \text{value} = \frac{2^{M-value}}{2^{M-value} + 1}$$

$$M - value = log2(\frac{\beta - \text{value}}{1 - \beta - \text{value}})$$

Note: α is usually set as 0 or 100 and c is a constant usually set as 1.

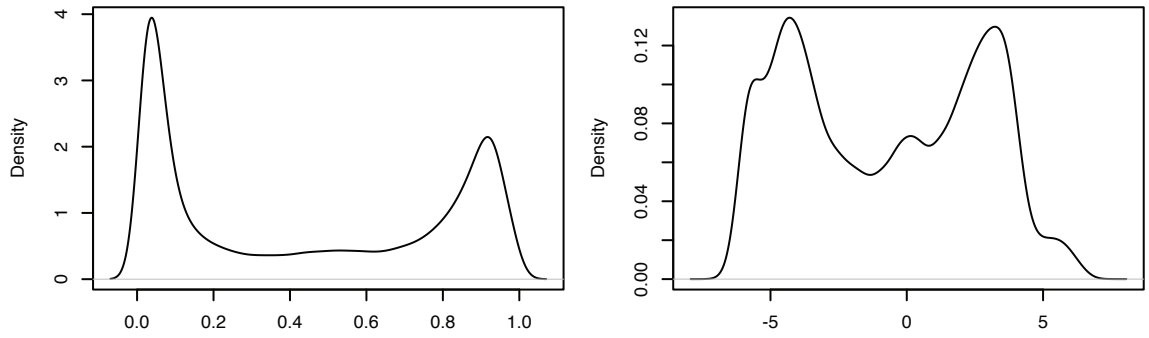


Figure 12: methylation β-value and M-value distribution.

A TCGA sample with HM450k data available is used for demonstration. Left panel shows the β-value distribution of all probes, we can clearly see two peaks which located around 0.1 and 0.9. These two peaks represent the unmethylated and methylated locus, respectively. Right panel shows the corresponding M-value distribution. We also see two peaks, which are the log-transformed peaks of β-value.

1.2.3.2 Probe annotations

Illumina has provided comprehensive annotation for probes on the DNA methylation microarray. Probes are annotated by CpG island relationship, functional genomic groups,

Infinium type, and so on. The probes' relationship to CpG island can be classified as CpG island, shore, shelf, and open sea. Shore indicates probes are 0-2k bp away from CpG islands, shelf indicates they are 2-4k bp away from CpG islands, and open sea indicates they are are more than 4k bp away from CpG islands. Probes can also be classified as TSS200 (within TSS 200 bp), TSS1500 (within TSS 1,500 bp), 5'UTR or 3' UTR (untranslated regions), 1st exon, body, and intergenic. Other than body, 3'UTR and intergenic, all categories belong to the promoter region. Using the probes located on HM450k platform as an example, **Figure 13** has provided the rough distribution.

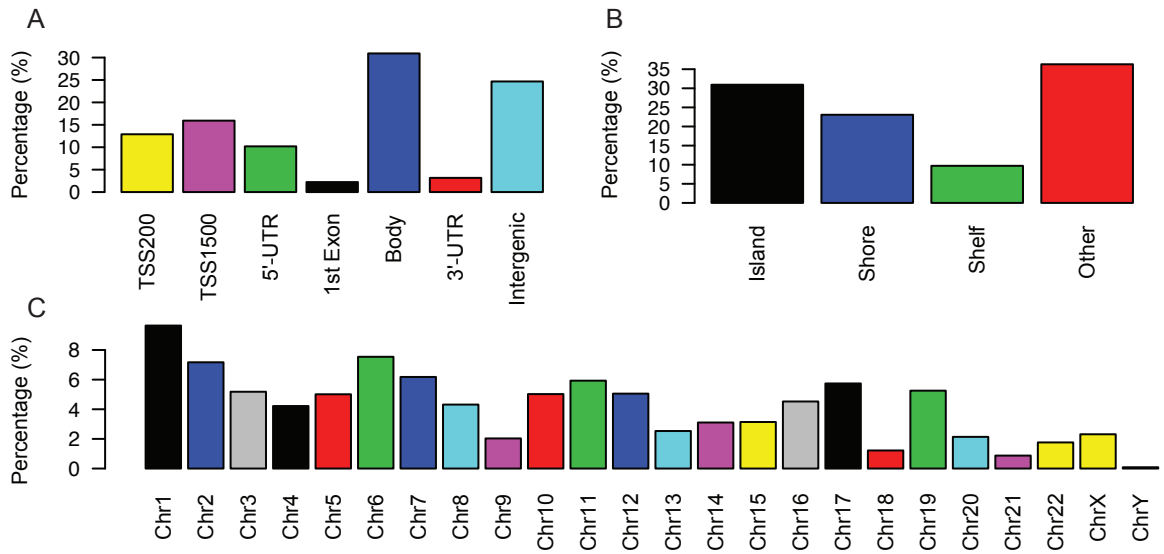


Figure 13: HM450k probes percentage of annotation

Y-axis shows the percentage of probes category over all probes and X-axis shows each category. **A**: functional genomic distribution among four different subgroups: promoter (TSS200, TSS1500, 5'-UTR, and 1st Exon), body, 3'UTR, and intergenic. **B**: probe to relationship of CpG islands: CpG islands, shore, shelf, and open sea. **C** Probes distribution of chromosome location.

1.2.3.3 Platform normalization

As introduced previously, two different probe types exist for both the HM450k and the current HM EPIC array. Due to the different chemical technologies and different sensitivities and specificities, they are not directly comparable: the data distribution is different (**Figure 14**). Many different methods have been proposed to merge the data generated using Infinium type I and type II probes, including peak-based correction (PBC; (62), subset quantile normalization (SQN;

(63), subset quantile within array normalization (SWAN; (64), all sample mean normalization (ASMN; (65) and beta-mixture quantile normalization (BMIQ; (66). These normalization methods are useful when trying to merge the data from the HM27k and HM450k platforms because the HM27k only has Infinium I probes while HM450k has both Infinium I and II probes. Some of the targeted CpG sites are measured with Infinium I probes in HM27k but with Infinium II probes in HM450k. Therefore, normalization is usually applied to the HM450k data to adjust the data distribution. This will be further explained in Chapter 2 in the methods section.

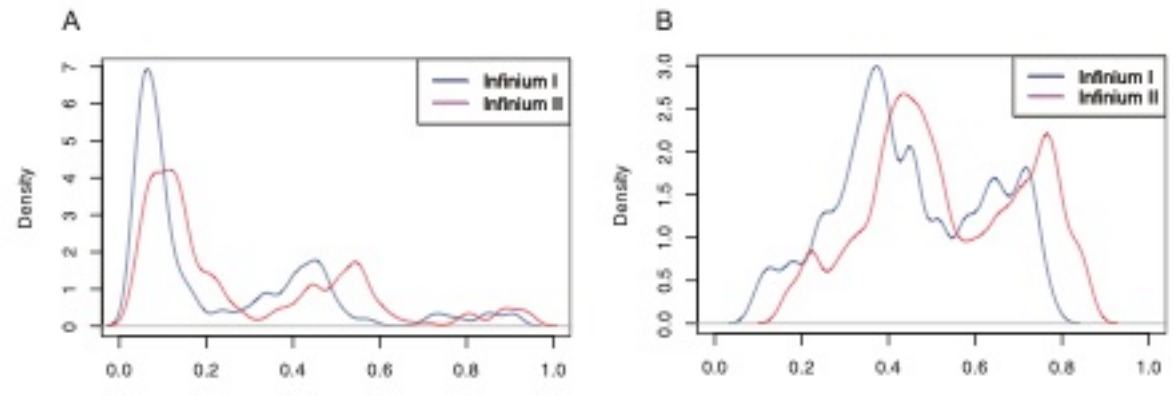


Figure 14: Beta value distribution for Infinium type I and type II probes.

A set of AML samples from TCGA with both HM27k and HM450k available are used for visualization. **A**: Probe cg00005847 is Infinium I type in HM27k and Infinium II type in HM450k. The density plot shows the different distribution of β-value. **B**: Probe cg00025991 is Infinium I type in HM27k and Infinium II type in HM450k. The density plot also shows different distribution of β-value.

1.3 Machine learning models

As an important subsection of artificial intelligence (AI), machine learning (ML) algorithms have benefited our lives in many ways, such as through facial recognition, recommendation engines, self-driving cars and other technological advancements. ML algorithms differ from traditional algorithms in that computers learn the data by themselves. ML is an interdisciplinary field that requires knowledge of statistics, computer science, and programming, and a background in the domain where it is applied. General speaking, machine learning questions can be categorized into two fields: supervised learning and unsupervised learning. Supervised learning methods train algorithms based on input and output data that is labeled by humans, and

unsupervised learning methods train algorithms without labeled output data and find the internal data structure by themselves. In short, if your data has provided labels, it is a question of supervised learning and vice versa. Of course, semi-supervised learning in another option when labeled and unlabeled data are mixed.

By applying theory to practical problems, researchers use ML models to explore the data by exploring the data's hidden patterns and structure or by building predictive models and applying the models to future prediction. For example, researchers can use ML algorithms to explore potential cancer subtypes by clustering analysis or to predict tomorrow's weather by building predictive models using past weather information.

The most commonly used unsupervised clustering algorithms are clustering methods, including k-means clustering, hierarchical clustering, Gaussian mixture models, and others. To apply clustering analysis, the most important step is to calculate the similarity or dissimilarity between samples. With the similarity or dissimilarity metric, samples can be grouped together using different algorithms.

For supervised learning questions, there is usually a set of example data with both independent variables and dependent variables. In this case, researchers train models to predict the dependent variable using the independent variables. Many algorithms are available for supervised learning questions, including support vector machines (SVM), linear regression, logistic regression, naïve Bayes (NB), linear discriminant analysis (LDA), Gaussian discriminant analysis (GDA), and decision trees. According to the data type of the dependent variable, algorithms can also be categorized into regression or classification algorithms; if the dependent variable is numeric variable, regression, and if the dependent variable is categorical variable, classification. Regression problems sometimes can be transformed into classification problems, such as when using thresholds to categorize numeric value into multiple categories. But this transformation can lead to loss of information.

1.3.1 Fundamentals of ML

To train an algorithm for a set of independent ($x_1, x_2, x_3, ...$) and dependent variables (y), the main goal is to minimize the difference between predicted $\hat{y}$ and real observed y ($difference = y - \hat{y}$). For example, the most commonly used solution for linear regressions is ordinary least square (OLS), which involves minimizing the sum of the difference between predicted and observed values ($\sum_{i=1}^{n}(y_i - \hat{y_i})$). Model fitting involves many interesting topics, such as variance bias tradeoff, variable selection, overfitting, and so on.

1.3.1.1 variance and bias tradeoff

Variance and bias tradeoff is also called a bias-variance dilemma or bias-variance problem, and it refers to how predictive models with a lower bias in parameter estimation will have higher variance, and vice versa. The difference between what is observed (y) and what is predicted ($\hat{y}$) is called error, and the error can be decomposed to variance, bias, and irreducible error. With fixed error, variance increase leads to lower bias, and bias increase leads to lower variance. Generally speaking, with increasing model complexity, the bias decreases and variance increases, which indicates potential overfitting risk. In contrast, with decreasing model complexity, the bias increases and variance decreases, which indicates underfitting risk (**Figure 15**). For model building purposes, it is important to find a good balance with certain model complexity but without overfitting to the current available data.

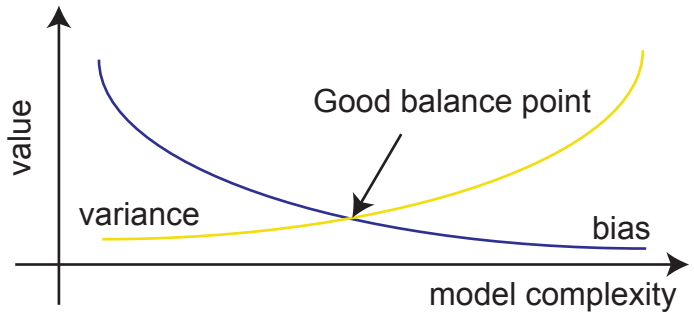


Figure 15: Bias and variance tradeoff plot

Y-axis represents the model complexity and X-axis represents the value of variance or bias. We can see, bias decreases as model complexity increases and variance increases as model

complexity increases. There exists a crossing point between the variance and bias line which is the good balance point we are looking for.

Assuming a set of data with input data points: $x_1, x_2, ..$, and targeted data points y_i with function: $y = f(x) + \epsilon$, $\epsilon \in N(0, \sigma^2)$ and a function $\hat{f}(x)$ as a good approximation of function $f(x)$, the squared loss or squared error would be:

$$Err(x) = E\left[(y - \hat{f}(x))^2\right]$$

Then the squared error can be decomposed as the sum of variance and bias as below:

$$\hat{f}(x) = \hat{y}$$

$$E[(y - \hat{y})^2] = E[(y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2]$$

$$= E[(y - E[\hat{y}])^2 + (E[\hat{y}] - \hat{y})^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})]$$

$$bias = \ y - E[\hat{y}] \ and \ variance = (E[\hat{y}] - \hat{y})^2 = (\hat{y} - E[\hat{y}])^2$$

This yields the following:

$$E[(y - \hat{y})^2] = bias^2 + Variance + 2 * E[2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})]$$

$$= bias^2 + Variance + 2(y - E[\hat{y}])(E[E[\hat{y}] - \hat{y}])$$

$$= bias^2 + Variance + 2(y - E[\hat{y}])(E[\hat{y}] - E[\hat{y}])$$

$$= bias^2 + Variance$$

1.3.1.2 Variable selection

In linear regression models, it is suggested to have 5-10 times the number of samples more than the number of parameters to get a robust model. However, in biological fields, it is more common to see more parameter than samples. For example, the number of genes available in gene expression data usually reach the thousands, and the number of samples is limited to a few hundred or less. With this type of high-dimension data, the data becomes very sparse, which hurts the robustness of model. This problem is called the curse of dimensionality. It is reasonable to expect smaller errors when increasing the model complexity or including more variables, but this may lead to overfitting and is not applicable for future data.

To avoid this problem, variable selection is often applied in data cleaning and preprocessing steps; this is an important step in building statistical models. It can help to select the most relevant features and exclude redundant features. It can increase productivity and accuracy and improve model interpretability. There are three main categories of feature selection:

(1) Filters: Filters investigate only the intrinsic characteristics of a given data set. They are fast and independent of any learning methods. Two types of filters exist: feature weighting-based and feature searching-based filters. According to certain evaluation criterion, filtering by feature weighting independently measures the relevance of each feature to the target problem. Usually the output is weight or ranking for each feature, and features are selected based on a given threshold. Filtering by feature searching takes inter-feature interaction into account and generates a subset of features that tends to be more relevant to the target problem.

(2) Wrappers: Wrappers differ from filters by integrating with a specific learning method. They search for a subset of features with optimal model performance. Usually wrappers can provide better results than filters, but are more computationally intensive. The variable selection models used in my dissertation belong to this category: variable selection within linear regression and random forests.

(3) Embedded: To overcome the disadvantages of both filters and wrappers, the embedded method combines them and looks for a tradeoff between performance and computational costs by using the internal information of learning methods.

The most commonly seen variable selection methods in linear regression are called forward and backward selection. Forward selection starts with a null model without any variables or parameters. Then it fits simple linear regression models by adding one parameter at a time. If the newly added parameters can statistically improve the model, then it is kept in the model, otherwise, it is excluded from the model. This evaluation step is repeated many times until all parameters have been evaluated. Backward selection is the opposite; it starts with a full model with all variables and parameters. Then the parameters are evaluated for their influence on the model. Starting with the parameter which has the least influence, parameters are excluded from

the model until all remaining parameters are statistically significant for the model. It is more common to apply forward and backward selection together; parameters are repeatedly included or excluded from the model to find the best combination of parameters.

1.3.1.3 Overfitting

As I mentioned above, with increasing model complexity, the bias decreases and variance increases, which indicates potential risk of overfitting. Overfitting means the model is overfitted to the current data set. For example, **Figure 16** shows an attempt to fit a linear model in such a way to separate the circles from the stars. Most of the stars are located at the left top region, and all circles are located at the right bottom region. A simple linear regression model can separate them well, except in a few cases (red straight line). It is also possible to fit a more complicated polynomial model to make sure all circles and stars are well separated (yellow curve). However, this yellow curve is highly possibly overfitting to the current data and will not achieve good performance with future data.

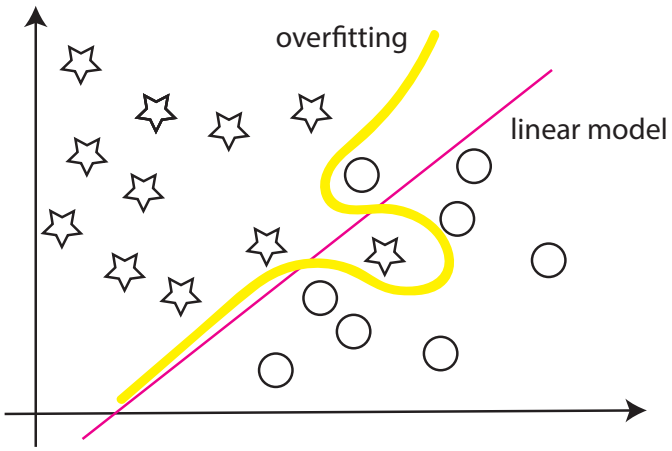


Figure 16: Overfitting example.

Overfitting is an important issue to consider when fitting models, especially when the models are used for future predictions. Overfitting causes the predictive model to be useless and even provide wrong conclusions about future data. The most commonly used methods to avoid overfitting include: (1) using sample or feature bagging or bootstrapping for model fitting process; (2) splitting data into training, development, and test sets and fitting models with training and

development sets then validated in test sets; (3) stopping early when the model is not significantly improved; (4) variable selection.

1.3.1.4 Linear regression assumptions

To apply linear regression models, four assumptions need to be satisfied (**Table 7**).

Table 7: Four assumptions of linear regression model

assumption	it means…	be detect by…	be adjust by…
linearity	Y show linearity with X	residual versus X or residual versus predicted Y, or X versus Y	nonlinear transformation of X or Y, such as logarithm, exponential, polynomial
independence	error independent over time, often seen in time-series data	residual versus time	adding lag of X or Y; adding variable to control the seasonal pattern
homoscedasticity	error independent of X or Y	residual versus predicted Y, residual versus X	nonlinear transformation of Y
Normality	error follow normal distribution and expect to be zero	quantile-quantile plot (qq-plot)	use non-linearity function; split data when it follows multiple distribution

1.3.1.5 Cross-validation

Cross-validation (CV), also called out-of-sample testing is a widely applied model validation technique, often used for parameter tuning. It splits all training data into n-folds, then uses n-1 folds to build the model and apply to the left fold. Then it changes the fold which is used as the test set and keeps rotating until all folds have been used as test sets (**Figure 17**). Cross-validation is usually used for parameter tuning.

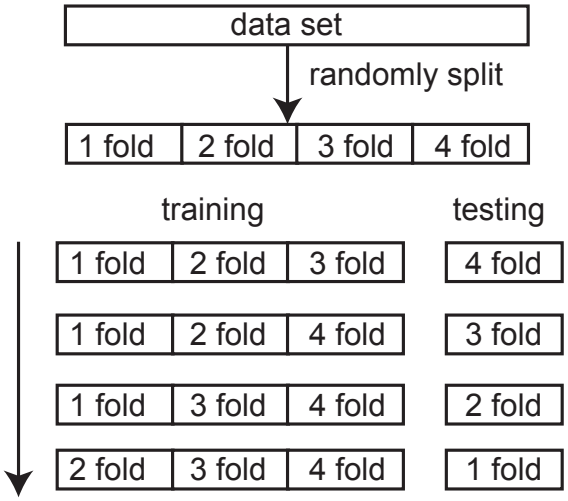


Figure 17: Cross-validation example

A whole data set has been randomly split into four folds. During the cross validation (CV) process, the 1st, 2nd, and 3rd folds are used to train the model and then the model is applied to the 4th fold in the first round. For the second round, the 1st, 2nd, and 4th folds are used to train the model and then model is applied to the 3rd fold. For the third round, the 1st, 3rd, and 4th folds are used to train the model and then model is applied to the 2nd fold. For the fourth round, the 2nd, 3rd, and 4th folds are used to train the model and then model is applied to the 1st fold. The order of fold usage is arbitrary and should not affect the results.

1.3.2 Elastic Net

In my dissertation, I used two ML methods to build predictive models: elastic net and random forest. A linear model fits a linear relationship among one or more independent variables and a dependent variable. This is the most simple and widely used type of model. The dependent variable can be either categorical or numerical data. Elastic net is a type of regularized linear regression model with L1 and L2 penalty (L1 penalty: $\|\beta\|_1 = \sum_{i=1}^{n}|\beta_i|$ and L2 penalty: $\|\beta\|_2 = \sum_{i=1}^{n}\beta_i^2$). It combines the least absolute shrinkage and selection operator (LASSO) and ridge regularization to overcome the disadvantages of each method alone.

1.3.2.1 LASSO regression

LASSO was discovered and popularized by Robert Tibshirani in 1996. It improves the prediction accuracy and interpretability of regression models by variable selection. By adjusting the tuning parameter, λ, the LASSO function can penalize the absolute size of the regression coefficients (**Figure 18**). Therefore, it can drive the coefficients of irrelevant variables to 0 and

lead to variable selection. However, the LASSO method has problems with consistency: it generates biased parameter estimates when applied to a large set of variables. In other words, if researchers run LASSO multiple times with different seeds or different sample sets, it may select a different set of variables.

Consider the common Gaussian linear regression model

$$\mathbf{y} = X\beta + \epsilon,$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ are the responses, $\beta = (\beta_1, \ldots, \beta_d)^T$ are the regression coefficients, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_d)$ is the covariate matrix, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \sim N(0, \sigma^2 I_n)$ are the normal noises.

The Lasso estimate is the solution to

$$\min_{\beta} (\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta), \quad \text{s.t.} \sum_{j=1}^{d} |\beta_j| \leq t. \tag{2.1}$$

Here $t \geq 0$ is a tuning parameter. Let $\hat{\beta}^0$ be the ordinary least square (OLS) estimate and $t_0 = \sum |\hat{\beta}_j^0|$. Values of $t < t_0$ will shrink the solutions toward 0. As shown in Tibshirani (1996), the Lasso gives sparse interpretable models and has excellent prediction accuracy. An alternative formulation of the Lasso is to solve the penalized likelihood problem

$$\min_{\beta} \frac{1}{n}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta) + \lambda \sum_{j=1}^{d} |\beta_j|. \tag{2.2}$$

The formulations (2.1) and (2.2) are equivalent in the sense that, for any given $\lambda \in [0, \infty)$, there exists a $t \geq 0$ such that the two problems have the same solution, and vice versa.

Figure 18: Lasso formula

Note: cited from the textbook "*Elements of Machine Learning*"

1.3.2.2 Ridge regression

Ridge regression is also called Tikhonov regularization. It is often used to deal with the problem of multicollinearity. Multicollinearity refers to explanatory variables showing collinearity, which means that as one variable increase, the other variable will also increase (positive relationship) or decrease (negative relationship). This can be evaluated by drawing a scatter plot

between two explanatory variables or by calculating correlation coefficients between variables. When data show multicollinearity, their least squares estimates are unbiased, but their variances are large so they may be inaccurate. Therefore, ridge regression adds a degree of bias to the regression estimates to reduce the standard errors. As shown in the formula below, the function adds a L2-norm in the model. This shrinks the coefficients towards zero as λ increases, but they will never reach zero. Moreover, ridge regression assigns the same weights to variables to prevent multicollinearity.

$$min_{\beta} \frac{1}{n} ||y - x\beta||_2^2 + \lambda ||\beta||_2^2$$

1.3.2.3 Elastic net

Elastic net is a combination of the ridge and LASSO methods and involves adjusting the composition between them. In general, an elastic net model's estimation is defined as:

$$\hat{\beta} = argmin(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|)$$

In this way, for fixed λ_2, elastic net first finds the ridge regression coefficients, then does a LASSO type shrinkage. Many software packages can apply elastic net while the most commonly used is an R package named glmnet: Lasso and elastic-net regularized generalized linear model (67). Instead of adding two parameters to each of the L1 and L2 norms, glmnet uses the following function to control the ratio between LASSO and ridge. When $\alpha = 0$, it applies ridge regression and when $\alpha = 1$, it applies LASSO regression. Then it uses the parameter λ to control the penalty level. The response variable can be multiple types, including quantitative families such as Gaussian and Poisson, categorical data type as binary or multi-classes, and survival data.

$$\lambda\alpha||\beta||_1 + \lambda(1-\alpha)||\beta||_2^2 \qquad \text{with } 0 \leq \alpha \leq 1;\ \lambda \geq 0$$

1.3.3 Random forest

Random forest is a type of ensemble learning that constructs multiple decision trees. Decision tree is a decision support tool with a tree-like model of decisions and their possible consequences. It can be used for both regression and classification problems. It can also be

used to calculate the variable importance using the normalized out of bag error. Out of bag error technique evaluates how important the variable is by calculating the error between models with and without the variable.

Random forest has the following advantages: robust to overfitting issues and robust to irrelevant features. As described earlier, overfitting is a critical issue that may significantly affect a model's application. To avoid overfitting, researchers use the sample-level or feature-level bagging techniques, which involve fitting the model with randomly selected samples or features with a replacement. Since the irrelevant variables are rarely selected and used in trees, it does not affect the model performance that much. Usually, researchers can ensemble multiple shallow trees to avoid overfitting and reduce the biases.

1.3.3.1 Entropy-based feature selection

I used entropy-based algorithms in evaluation of variable importance for gene expression subtype prediction. **Entropy** is the measure of impurity, disorder, or uncertainty in a group of examples. It is the expectation of information from samples (X).

$$\mathrm{H}(X) = E\big(I(X)\big)$$

If the information of X is defined as:

$$I(X) = -\log 2\ (P(X))$$

then the **entropy** is defined as:

$$H(X) = E\left(-log2\big(P(X)\big)\right) = -\sum_{i=1}^{n} P(x_i)\log P(x_i)$$

$$P(x_i)\ is\ the\ proportion\ of\ sample\ x_i\ belonging\ to\ each\ class\ i\ \in (1, 2, \dots n\ )$$

The base of the logarithm can be e, 2, or 10 and the most commonly used is 2. X is a discrete random variable with potential values: $\{x_1, x_2, \dots, x_n\}$. In my case, I assume that X is the classes I have. For example, if there are three subtypes for gene expression, then X can be $\{x_1 = CL, x_2 = MES, x_3 = PN\}$. The maximum entropy is when $p(x_i) = \frac{1}{n}$. In this case, the probability of

each subtype was 1/3. The minimum entropy is when the probability belonging to one subtype is 1 and the probability belonging to the other two subtypes is 0.

$$H(x)_{max} = -\sum_{i=1}^{n} p(x_i) log2\, p(x_i) = -\sum_{i=1}^{n} \frac{1}{n} log2\left(\frac{1}{n}\right) = -n * \left(\frac{1}{n} log2\left(\frac{1}{n}\right)\right) = -log2\left(\frac{1}{n}\right) = log2n$$

$$H(x)_{min} = -\sum_{i=1}^{n} p(x_i) log2\, p(x_i) = -\big(1 * log2(1) + 0 * log2(0)\big) = 0$$

The **conditional entropy** of two events X and Y with values of x_i, y_j is:

$$H(X|Y) = -\sum_{i,j} p\big(x_i, y_j\big) log2(\frac{p\big(x_i, y_j\big)}{p\big(y_j\big)})$$

The idea of using entropy in feature selection is that a good feature can make a better partition of samples. Entropy is smaller when samples partitioned into one class are more "pure" or "homogenous." In contrast, entropy is higher when samples within one subclass are impure. In my study, I used three metrics for evaluation of variable importance that are suitable for categorical variables.

1.3.3.1.1 Information gain

Information gain (IG) measures how much "information" a feature can give about the class. Features that perfectly partition samples should give maximal information. Unrelated features should give no information. IG is often used in decision trees and random forest.

$$IG = H - \frac{m_l}{m} H_l - \frac{m_r}{m} H_r$$

m is the total number of samples.

m_l *is the total number of samples belonging to* l *class.*

H_l *is the entropy of samples belonging to* l *class.*

This formula supposes there exist only two classes: l *and r.*

Using three subtypes as an example, suppose there are two features under consideration, feature 1 (**Table 8**) and feature 2 (**Table 9**), and three subtypes (CI, MES, PN) targeted for

prediction. It is worth noting that IG can only be used for categorical independent variables. If the independent variables are numerical values, they are categorized into classes.

Table 8: Confusion matrix between feature 1 annotation and observed subtypes

	subtype: CL	subtype: MES	subtype: PN	Subtotal
feature 1: CL	10	10	0	20
feature 1: MES	10	15	10	35
feature 1: PN	5	5	20	30
subtotal	25	20	30	85

Note: columns are showing the real subtype membership and the rows are showing the predicted subtype membership using feature 1.

Table 9: Confusion matrix between feature 2 annotation and observed subtypes

	subtype: CL	subtype: MES	subtype: PN	Subtotal
feature 2: CL	15	5	5	20
feature 2: MES	5	20	5	35
feature 2: PN	5	5	20	30
subtotal	25	20	30	85

Note: columns are showing the real subtype membership and the rows are showing the predicted subtype membership using feature 1.

In **Table 8** and **Table 9**, a confusion matrix is built between subtypes predicted by feature (feature 1 and feature 2) and the real observed subtype. Intuitively, it is clear that feature 2 provides better prediction or annotation than feature 1 because it has higher prediction accuracy (feature 1 accuracy = (10 + 15 + 20) / 85 = 0.5294 and feature 2 accuracy = (15 + 20 + 20) / 85 = 0.6470). Then IG is calculated to compare these two features' performance in predicting subtypes.

Entropy of observed subtypes:

$$H = -\big(p(CL) * log2p(CL) + p(MES) * log2p(MES) + p(PN) * log2p(PN)\big)$$

$$= -\left(\frac{25}{85} * log2\left(\frac{25}{85}\right) + \frac{30}{85} * log2\left(\frac{30}{85}\right) + \frac{30}{85} * log2\left(\frac{30}{85}\right)\right) = 1.579863$$

IG of feature 1 annotation:

$$IG(feature1) = H - \frac{m_{CL}}{m}H(CL) - \frac{m_{MES}}{m}H(MES) - \frac{m_{PN}}{m}H(PN)$$

$$H(CL) = -\left(\frac{10}{20} * log2\left(\frac{10}{20}\right) + \frac{10}{20} * log2\left(\frac{10}{20}\right) + \frac{0}{20} * log2\left(\frac{0}{20}\right)\right) = 1$$

$$H(MES) = -\left(\frac{10}{35} * log2\left(\frac{10}{35}\right) + \frac{15}{35} * log2\left(\frac{15}{35}\right) + \frac{10}{35} * log2\left(\frac{10}{35}\right)\right) = 1.556657$$

$$H(PN) = -\left(\frac{5}{30} * log2\left(\frac{5}{30}\right) + \frac{5}{30} * log2\left(\frac{5}{30}\right) + \frac{20}{30} * log2\left(\frac{20}{30}\right)\right) = 1.251629$$

$$IG(feature1) = 1.579863 - \frac{20}{85} * 1 - \frac{35}{85} * 1.556657 - \frac{30}{85} * 1.251629 = 0.2618411$$

IG of feature 2 annotation:

$$IG(feature2) = H - \frac{m_{CL}}{m} H(CL) - \frac{m_{MES}}{m} H(MES) - \frac{m_{PN}}{m} H(PN)$$

$$H(CL) = -\left(\frac{15}{25} * log2\left(\frac{15}{25}\right) + \frac{5}{25} * log2\left(\frac{5}{25}\right) + \frac{5}{25} * log2\left(\frac{5}{25}\right)\right) = 1.370951$$

$$H(MES) = -\left(\frac{5}{30} * log2\left(\frac{5}{30}\right) + \frac{20}{30} * log2\left(\frac{20}{30}\right) + \frac{5}{30} * log2\left(\frac{5}{30}\right)\right) = 1.251629$$

$$H(PN) = -\left(\frac{5}{30} * log2\left(\frac{5}{30}\right) + \frac{5}{30} * log2\left(\frac{5}{30}\right) + \frac{20}{30} * log2\left(\frac{20}{30}\right)\right) = 1.251629$$

$$IG(feature2) = 1.579863 - \frac{25}{85} * 1.370951 - \frac{30}{85} * 1.251629 - \frac{30}{85} * 1.251629 = 0.2931393$$

Based on this calculation, it is clear that feature 2 provides more information than feature1. This matches the observations from the confusion table. The drawback of IG is that IG tends to select the features with a large number of distinct values. Applying this method, R package Fselector uses following formula:

$IG = H(C) - H(C|F)$

$where\ H(C) = -1 * sigma_i(P(c_i) log2\ P(c_i))$

$$H(C|F) = sigma_j\left(P(f_j) * H(C\ |f_j)\right)$$

$$H(C|f_j) = -1 * sigma_k(P(c_k|f_j) log2 P(c_k|f_j))$$

1.3.3.1.2 Gain Ratio

Gain ratio (GR) is the modified IG that reduces its bias on highly branching features (a drawback of IG). GR considers the number and size of branches when choosing a feature and achieves this by normalizing IG by the "intrinsic information" of a split. Intrinsic information is defined as the potential information generated by splitting the data set into v partitions. It can be calculated as:

$$intrinsic\ Information(D) = -\sum_{j=1}^{v} \frac{|D_j|}{D} * log2(\frac{|D_j|}{D})$$

D_j *is the probability that sample D belongs to class j.*

High intrinsic information occurs when partitions have similar sample sizes and low intrinsic information occurs when few partitions have majority samples, meaning data are purer. The GR is defined as follows:

$$GainRatio(F) = \frac{Gain(F)}{Intrinsic\ ifnor(F)} \quad (F\ is\ the\ feature\ under\ evaluation)$$

With GR calculated for each feature, selection of variables with high GR is easy. The application of GR can be explained with the same example shown in **Table 8** and **Table 9**.

GR of feature 1 annotation:

$$GainRatio(f1) = \frac{0.2618411}{intrinsic\ Infor(f1)}$$

$$intrinsic\ infor(f1) = -\left(\frac{20}{85} * log2\left(\frac{20}{85}\right) + \frac{35}{85} * log2\left(\frac{35}{85}\right) + \frac{30}{85} * log2\left(\frac{30}{85}\right)\right) = 1.548565$$

$$GainRatio(f1) = \frac{0.2618411}{1.548565} = 0.1690863$$

GR of feature 2 annotation:

$$GainRatio(f2) = \frac{0.2931393}{intrinsic\ Infor(f1)}$$

$$intrinsic\ infor(f2) = -\left(\frac{25}{85} * log2\left(\frac{25}{85}\right) + \frac{30}{85} * log2\left(\frac{30}{85}\right) + \frac{30}{85} * log2\left(\frac{30}{85}\right)\right) = 1.579863$$

$$GainRatio(f2) = \frac{0.2931393}{1.579863} = 0.1855473$$

By comparing the GR calculated for both features, it is possible to conclude that feature 2 has a higher GR than feature 1, meaning feature 2 can better characterize the subtypes. In summary, compared with IG, which is biased toward multivalued features, GR tends to prefer unbalanced splits in which one partition is much smaller than the other. Because the unbalanced split will generate smaller intrinsic information value, which leads to higher gain ratio. It normalizes the IG by the number and size of branches. However, the problem with GR is that it may overcompensate and choose a subtype just because its intrinsic information is low. This can be fixed only by considering subtypes with greater IG than average IG.

1.3.3.1.3 Symmetrical uncertainty

Symmetrical uncertainty (SU) is one of the variants of mutual information (MI). MI of two random variables measures the mutual dependence between them. It quantifies the amount of information possible to obtain from one random variable by observing the other random variable. The MI between discrete random variables X and Y can be calculated as:

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)}\right)$$

p(x,y) is the joint probability function of X and Y.

p(x) and p(y) have marginal probability distribution.

The MI between continuous random variable X and Y can be calculated as:

$$MI(X;Y) = \int \int p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$$

p(x,y) is the joint probability density function of X and Y.

p(x) and p(y) are marginal probability density functions of X and Y respectively.

MI has two features: nonnegative ($I(X;Y) \geq 0$) and symmetric ($I(X;Y) = Y(Y;X)$). Minimum MI happens when X and Y are independent variables. In other words, information cannot be inferred about one variable from the other variable. This yields $p(x,y) = p(x) * p(y)$.

Maximum MI happens when X and Y provide the same entropy. In other words, all information about one variable can be inferred from the other variable.

SU is defined as:

$$SU(X,Y) = 2R = 2\frac{I(X;Y)}{H(X)+H(Y)}$$

$$R = \frac{I(X;Y)}{H(X)+H(Y)}$$

R is another measure called redundancy.

SU can also be calculated using the harmonic mean of the two uncertainty coefficients $C_{XY}\ and\ C_{YX}$. An uncertainty coefficient is a measure of nominal association based on the information entropy.

$$SU(X|Y) = C_{XY} = \frac{H(X)-H(X|Y)}{H(X)} = \frac{I(X;Y)}{H(X)}$$

$$SU(Y|X) = C_{YX} = \frac{H(Y)-H(Y|X)}{H(Y)} = \frac{I(Y;X)}{H(Y)}$$

FSelector (R package) achieves SU as follows:

$$SU = 2 * \frac{IG(C|F)}{H(C)+H(F)} = \frac{H(C)-H(C|F)}{H(C)+H(F)}$$

$$where\ H(C) = -1 * sigma_i\left(P(c_i)log2\ P(c_i)\right)$$

$$H(C|F) = sigma_j\left(P(f_j) * H\left(C|f_j\right)\right)$$

$$H(C|f_j) = -1 * sigma_k(P(c_k|f_j)log2P(c_k|f_j))$$

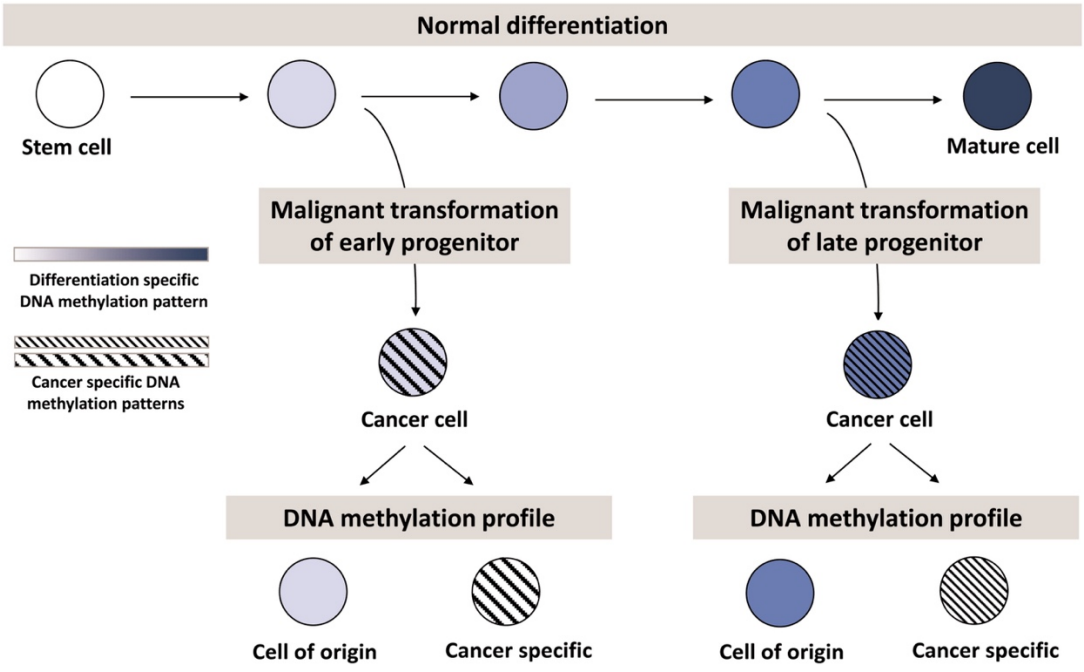
$$H(F) = -1 * sigma_i\left(P(f_i)log2\ P(f_i)\right)$$

1.4 Significance and importance

In my study, I focused on developing DNA methylation predictive signatures and applicable models for gliomas' characteristic genomic alterations. Why is this important and how can it affect current study in the field? It is already well-known that *IDH* mutation shows a strong association with the GCIMP signature in gliomas, but no one has ever looked into whether there are DNA

methylation signatures for other genomic alterations, such as somatic mutations, copy number variations, and gene expression subtypes. Moreover, DNA methylation is a relatively stable biomarker during tumor progression. For example, when comparing primary and recurrent gliomas, the DNA methylation profile does not show significant changes, although the gene expression profile can show subtypes switch. Some researchers have proposed that DNA methylation, as an epigenetic mechanism, can reflect the cell of origin (68). Researchers proposed that during the cell differentiation process, cells will acquire lineage-specific DNA methylation pattern which can be easily validated by comparing the methylation profile between different levels of differentiated cells. At the same time, new DNA methylation pattern are added when cells evolved to cancer cells (**Figure 19**) (69). Therefore, each cancer cell should harbor the stacked DNA methylation pattern from its lineage-specific and cancer-specific pattern. Even though researchers have already done comprehensive molecular profiling of tumors, information about tumor initiation is scarce. If DNA methylation is a stable marker of cell origin, DNA methylation signatures can help identify the cells with high potential to develop cancer.

Figure 19: DNA methylation patterns in normal differentiation and cancer.



Normal differentiation results in the acquisition of a specific DNA methylation pattern (indicated in blue). Additional to the DNA methylation pattern of the cell of origin, the cancer cell acquires cancer specific DNA methylation changes (depicted by black stripes).

Note: Figure adapted from Kulis, Marta, et al. "Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer." *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1829.11 (2013): 1161-1174 with order license ID: 1001061-1.

Therefore, I hypothesize that DNA methylation signature can be identified for somatic genomic alterations and those signatures can provide insights about genomic alterations. This can help develop understandings of the biology of those important genomic alterations. Furthermore, those genomic alterations are important features that can be used to describe and define tumors and their subtypes. However, there is no easy way to obtain them in clinical application. In clinical practice, all procedures are more complicated and must be validated and standardized by multiple organizations. The many rules make it hard to run tests on each patient, which may not seem like a big deal in research labs. For example, there hasn't been any survival application in clinical trials that utilize the gene expression subtypes to categorize patients. This is because there is no way to obtain that information in clinical practice. With this dissertation done, we can easily obtain all important genomic alteration information only based on DNA methylation microarray, which is fast, efficient, affordable, and suitable for FFPE samples. Last but not least, the expansion ability of the microarray bead chip can even provide more biomarkers in the future.

In summary, with my project I have developed a single-platform based DNA methylation biomarker prediction package, called the Unified Diagnostic (UniD) platform (**Figure 20**), using programming language R. This package accepts raw IDAT data and can run comprehensive data cleaning and processing procedures to generate clean, ready-to-use DNA methylation. More importantly, this platform can provide the binary status (mutant or wild type, codel or intact) of the genes *IDH*, *ATRX*, *TERT*p, and chr1p19q and quantitative multi-class value for gene expression subtypes. This package can be easily accessed through Github.

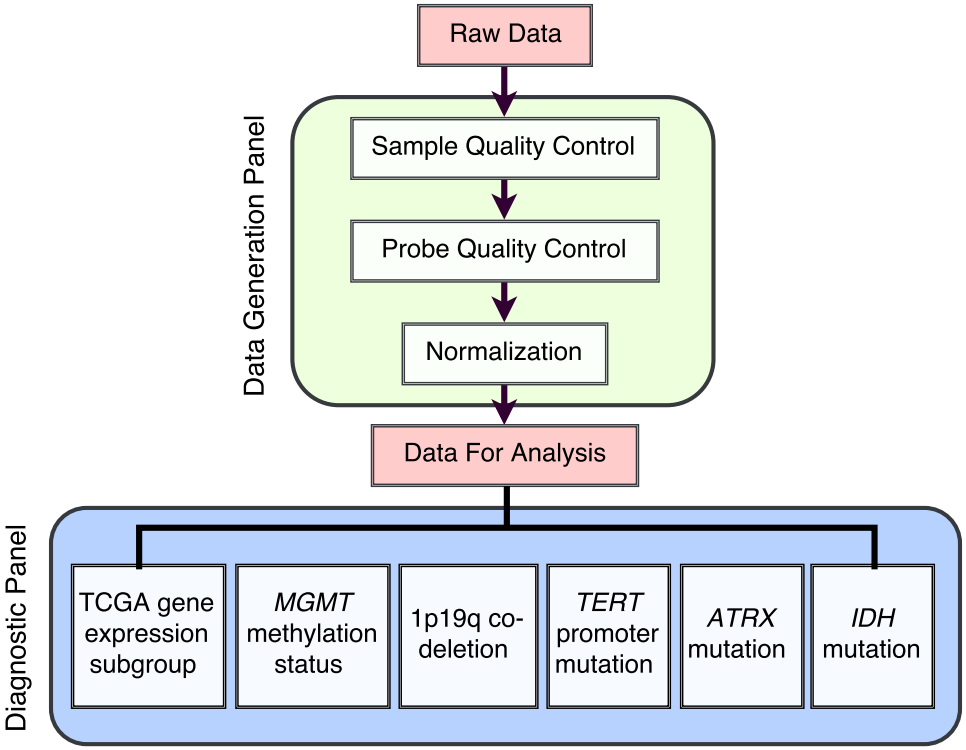


Figure 20: workflow of UniD package

There are two major parts of this pipeline: data generation and diagnostic. In the data generation panel, it applies sample-level QC, probe-level QC, and normalization. With all filters and normalization applied, it can generate ready-to-use DNA methylation data for the next step. In the diagnostic panel, multiple genomic alterations can be predicted with DNA methylation data, including TCGA gene expression subtypes, *MGMT* methylation status, chr1p19q codel status, and somatic mutation status of *TERT*p, *ATRX*, and *IDH*. *MGMT* methylation status is applying the MGMT-STP27 model.

Chapter 2

2 Binary genomic alterations prediction

In this chapter, I will introduce the methods and results of predictive model building for binary genomic alterations, including mutation status of *IDH*, *ATRX*, *TERT*p and chr1p19q codel. Because the DNA methylation processing itself is complex, I will describe it in its own section. Next, in the methods section, I will briefly introduce DNA methylation data processing, genomic alterations annotation, model building, and downstream analysis. In the results section, I will introduce model performance, model signatures, model validation with external data sets, and comparison of predicted status with existing DNA methylation-based CNS classification results.

Many different assays can be used to measure genome-wide DNA methylation profiles. DNA methylation microarray is one of them and is the most popular assay due to many advantages. For this purpose, I used the glioma samples with HM450k data available from TCGA project. I used the R package UniD for data processing and details are described below.

2.1 DNA methylation data processing

2.1.1 Data cleaning

In the UniD package, the data generation panel has three major categories: probe-level QC, sample-level QC, and normalization. The probe-level and sample-level QC were applied to all samples and normalization was applied to HM450k data only for data merging purposes (merge data from HM450k with HM27k).

2.1.1.1 Probe-level QC

There are two types of probe-level QC: (1) sample-dependent probe exclusion, in which the probe is set as a missing value and it is applied sample-wisely; (2) sample-independent probe exclusion, in which probes are usually predefined as potential “bad probes” and will be excluded for all samples. It is worth noting that these two types of probe-level QC (sample-dependent or

sample-independent) are not the same concepts discussed as related to the internal control probes.

sample-dependent probe exclusion: The missing value is assigned for a specific probe if it does not pass the criterion (**Table 10**). This is sample-dependent which means the missing value may be assigned to one probe in sample A but not necessarily in sample B. Three criteria are applied here. First of all, probes which do not pass the internal control threshold are set as missing. Second, probes which do not pass a background check, which means probes are not significantly detected when compared to background probe sets, are set as missing. Third, probes with less than three beads available are considered as failures and are set as missing. Those criteria are sample-dependently applied, which means one sample failing with a specific probe does not affect the same probe in other samples. Probes with high percentages of missing values across all samples may indicate poor quality which should be excluded from the data.

Table 10: probe level quality control details

probe type	details of probes	action for those probes
sample dependent	probes did not pass the internal control threshold	assign as missing value
	probes with detection p-value >0.05	
	probes with less than three beads	
sample independent	probes located on chromosome X and Y	recommend
	probes hybridization may affect by SNP (SNPhit)	recommend
	probes may hybridize to multiple location (multihit)	recommend
	probes not targeted to CpG methylation sites	optional
	probes available on EPIC array but not on HM450k array	optional

sample-independent probe exclusion: Probes that have been predefined as “bad probes” or as having “potential inaccuracy” will be excluded (**Table 10**). For example, probes located on chromosome X and Y are usually excluded to avoid the sex bias in data analysis. Probes can

also be excluded based on research questions. For example, probes located on HM450k but not on EPIC will be excluded if researchers want to apply HM450k-based models to EPIC array.

2.1.1.2 sample-level quality control

If a sample has a high rate of missing values after the sample-dependent QC, it is highly possible that this sample has low quality DNA or something was wrong with the experimental procedures. Usually, at least 90% of probes pass QC (except sample-independent probe filters). If a sample has missing values for more than 10% of probes, this sample should be used carefully; though other probes have been successfully detected from the bead chip but their methylation level may be biased.

2.1.2 normalization

A normalization step is utilized when it is necessary to merge DNA methylation data from HM27k and HM450k. As described in the introduction, HM27k and HM450k use different probe designs and different chemical techniques: HM27k only uses Infinium I probes and HM450k uses both Infinium I and II probes. Infinium I and II probes generate different data distributions and their data are not directly comparable. Many different methods are proposed to adjust the difference and normalize the data to the same distribution.

Below I present part of my master's thesis (70) in which I discuss a data-driven method to compare HM450k normalization methods (*material from my master thesis in italics).*

Comparison of Normalization Methods

Probes in the HM450k array include both Infinium I and Infinium II types, while HM27k probes include only the Infinium I type. The β-value derived from Infinium II probes has a smaller dynamic range and lower sensitivity compared to data from Infinium I probes (71). Approximately 21,000 probes overlapped between the HM450k and HM27k platforms based on probe ID. However, because the majority of overlapping probes from HM450k are Infinium II type probes and those from HM27k are mostly Infinium I probes, it is likely batch effect results when combining data sets. Batch adjustment is required to ensure the data are comparable between platforms. Several

*adjustment methods have been published, such as: BMIQ (66)R package: ChAMP), SWAN (64) R package: lumi), and PBC (72)R package: wateRmelon). AML DNA methylation data from TCGA were used to evaluate adjustment methods. All 194 AML samples are available in both HM27k and HM450k platforms. Three steps are used to evaluate these methods, briefly described below (**Figure 21**).*

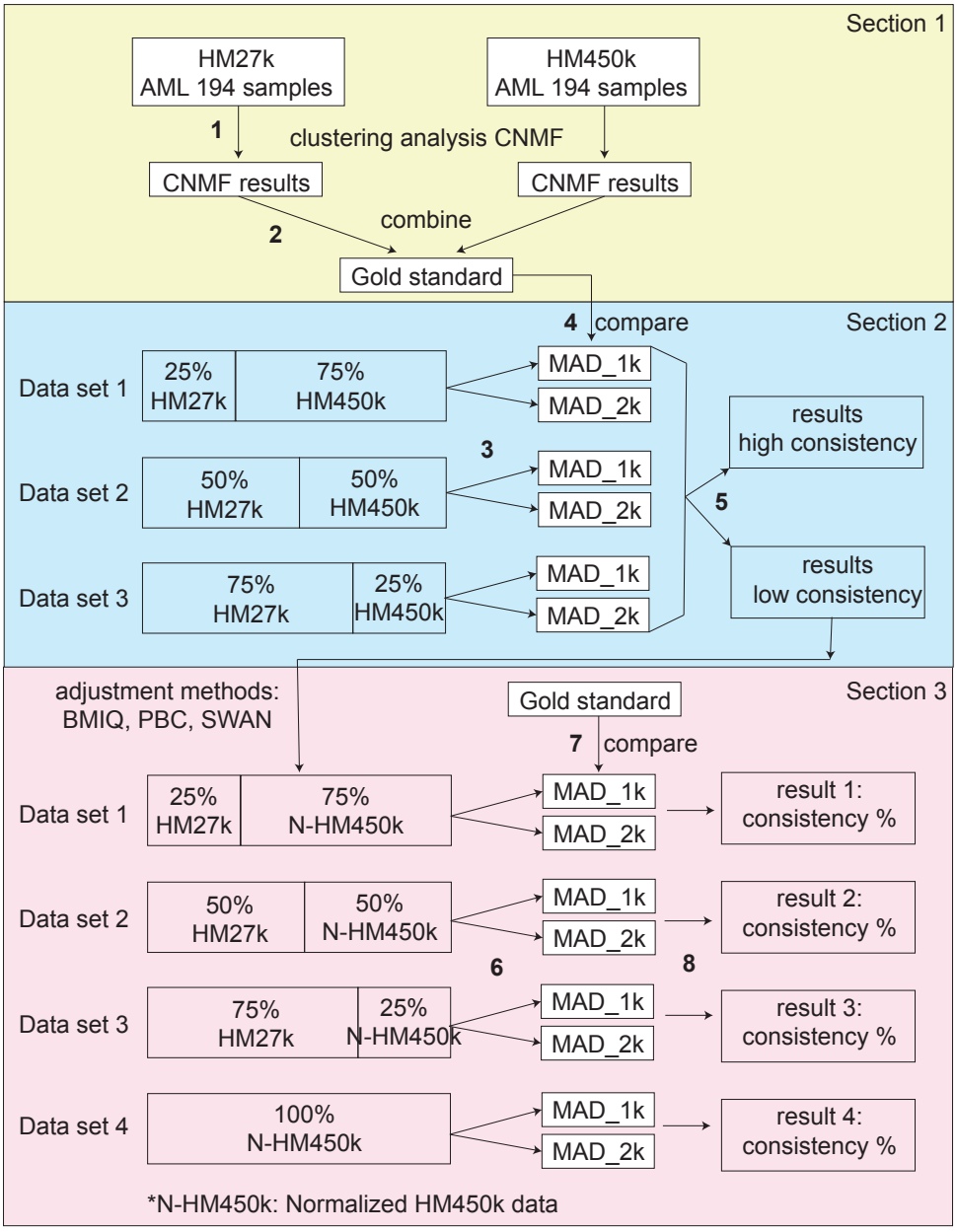


Figure 21: Data simulation procedures for evaluating HM450k data normalization methods.

Section 1: Generate the gold standard. Section 2: Evaluate simulated data sets with unadjusted HM450k data. Section 3: Evaluate simulated data sets with adjusted HM450k data. Adjustment methods include BMIQ, PBC, and SWAN. Each step is described in details. 1: Use data from

HM27k or HM450k to do the clustering analysis with CNMF and obtain the membership for each platform. 2: Combine the membership results from each data set and generate the final results as the gold standard. 3: Simulate three data sets with HM27k and unadjusted HM450k data, create two subsets for each data set, and obtain the clustering results for each subset. 4: Compare the six results with the gold standard. 5: Obtain the concordance percentage. If the concordance percentage is high, the data from HM450k are used without any adjustment. If the concordance percentage is low, three published adjustment methods are evaluated (BMIQ, PBC, and SWAN). 6: For each adjustment method, four data sets are simulated. Two subsets are generated by MAD value. CNMF is applied for each subset. 7: Compare the membership results with the gold standard. 8: Generate a consistent percentage for each simulated data set.

Section 1: 194 AML samples were clustered independently using the data from HM27k and HM450k. For each assay, probes were sorted by median absolute deviation (MAD) and the top 1000 and top 2000 probes were clustered using the consensus non-negative matrix factorization (CNMF; (73) method. These two cluster results were compared and the concordance between these two platforms was high. This initial classification was used as the gold standard.

Section 2: Three data sets consisting of admixtures of data from the HM27k and HM450k data sets were simulated. Data set 1 consisted of 25% samples coming from the HM27k data set and 75% samples from the HM450k data set. Data set 2 consisted of 50% of samples from HM27k and 50% from HM450k. Data set 3 consisted of 75% samples from HM27k and 25% from HM450k. CNMF clustering was applied to each of the admixture data sets. The clustering results from each admixture data sets were compared with the gold standard.

Section 3: For each adjustment method, the admixture process was repeated and a fourth data set containing only HM450k data was created. CNMF was applied to each admixture data set and the membership for each sample was obtained. Membership indicated the subgroup each sample belonged to, for example, the first AML sample may belong to subgroup 1 while the second AML sample belongs to subgroup 2. The membership of the classification was compared with the gold standard as was done for the unadjusted data sets.

Normalization method comparison results

*There are 194 AML samples available from TCGA and they were used for evaluation of adjustment methods. After deleting the probes with missing values, the HM27k data set contained 22,288 probes and the HM450k data set contained 393152 probes. There were 20,794 probes overlapping between these two platforms among the 194 AML samples and the MAD value was calculated for each probe. The distribution of the MAD value between the HM27k and HM450k was highly positive-skewed (**Table 11** and **Figure 22**). The MAD value is used to evaluate the variability of quantitative data. A high MAD value indicates high variability and a low MAD value indicates low variability. Most of the probes had a very low MAD value indicating a low variability among the samples for both HM27k and HM450k. Moreover, the β-value density distribution was different between the two data sets.*

Table 11: MAD distribution of data sets from HM27k and HM450k

MAD	*min*	*1st Q*	*median*	*mean*	*3rd Q*	*max*
HM27k	*0.0006097*	*0.0034990*	*0.0124800*	*0.0506500*	*0.0565700*	*0.5397000*
HM450k	*0.0009585*	*0.0058600*	*0.0132800*	*0.0499000*	*0.0549500*	*0.485300*

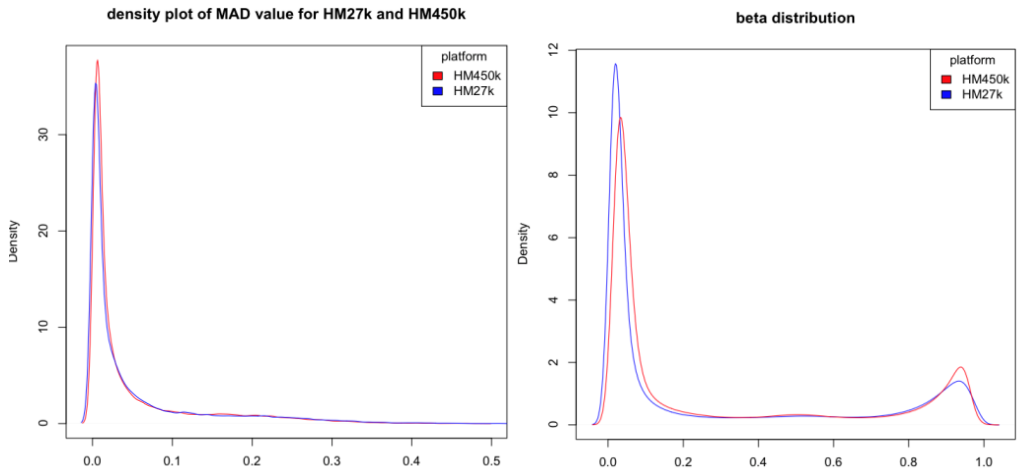


Figure 22: Distribution of two platforms

Left: Distribution of the MAD value for HM27k and HM450k. Right: β-density distribution for HM27k and HM450k among the AML samples.

*For both HM27k and HM450k platform, 1k and 2k probes were selected with the highest MAD value for the CNMF clustering analysis, respectively. Four subsets were generated. They were called k27_mad1k, k27_mad2k, k450_mad1k, and k450_mad2k. CNMF clustering analysis with data from each platform was used to generate the gold standard. Comparing the probes selected from k27_mad1k and k450_mad1k, 100% (1000/1000) of probes overlapped. Comparing the probes selected from k27_mad2k and k450_mad2k, 82.85% (1657/1000) of probes overlapped. This indicated that even though these two platforms have different β value distributions, the most variable probes have high concordance. Based on the cophenetic plot (**Figure 23**) obtained from CNMF clustering analysis, samples have the highest cophenetic coefficient value when k = 6. Therefore, for all comparisons in the AML data set, comparisons of the membership were only made when clusters were equal to 6.*

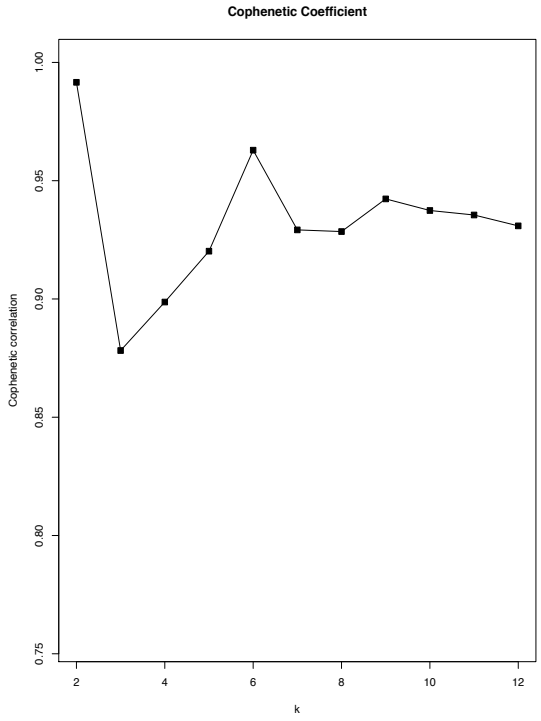


Figure 23: Cophenetic coefficient plot for k27_mad1k

The subgroups' attribution for each sample, designated as "membership," was compared among four data sets. The number of discordant samples is provided in Table 7. The membership

between k27_mad1k and k27_mad2k had the lowest discordance (19/194, 9.79%). The membership between k27_mad2k and k450_mad1k had the highest discordance (31/194, 15.98%). In general, the discordance was 26.2/194 (13.5%). Then the overall gold standards were generated by choosing the most frequent membership among these four data sets. The number of discordant samples between each data set and the overall gold standard is also shown in ***Table 12****.*

Table 12: Number of discordant samples among data sets used to generate gold standard

Data set	*k27_mad1k*	*k27_mad2k*	*k450_mad1k*	*k450_mad2k*
k27_mad1k	*-*	*19*	*25*	*29*
k27_mad2k	*0.097*	*-*	*31*	*29*
k450_mad1k	*0.129*	*0.160*	*-*	*24*
k450_mad2k	*0.149*	*0.149*	*0.124*	*-*
Overall Gold Standard	*7*	*13*	*18*	*22*

Note: The hyphen means not applicable when the cell's corresponding column data set and row data set are the same. The cells on the right of the hyphen are the number of discordant samples comparing the cells' corresponding column data set and row data set. The cells on the left are the percentage of discordant samples comparing the cells' corresponding column data set and row data set.

Next, the simulating data sets were generated based on the data from HM27k and data from HM450k. The data from HM450k reflects four different situations: unadjusted, adjusted by BMIQ, adjusted by PBC, and adjusted by SWAN. The unadjusted situation included three simulated data sets while each adjusted situation included four simulated data sets. All simulated data sets were clustered by CNMF and memberships were obtained. The membership between each simulated data set and the overall gold standard was compared. The number of discordant samples is shown in ***Table 13*** *and* ***Figure 24****. As the data shows, the numbers of discordant samples were large when adjusting the data from HM450k with the SWAN method. This suggests that the SWAN method did not adjust the data for clustering analysis as well as other methods. By comparing the numbers of discordant samples between the SWAN method and the unadjusted situation, SWAN has a significantly higher number of discordant samples than in the unadjusted situations (Wilcox test: p-value = 0.00066). Regardless of the situations (unadjusted*

or adjusted by BMIQ or PBC methods), the number of discordant samples was compared between all data sets with 1k probes and all data sets with 2k probes. There was no significant difference between them, but the data sets with 1k probes have a lower number of discordant samples than data sets with 2k probes. The number of discordant samples was compared between unadjusted, BMIQ, and PBC, and no significant difference was observed (Wilcox test, unadjusted and BMIQ p-value = 0.7; Wilcox test, unadjusted and PBC p-value = 1.0; Wilcox test, BMIQ, and PBC p-value = 0.46.

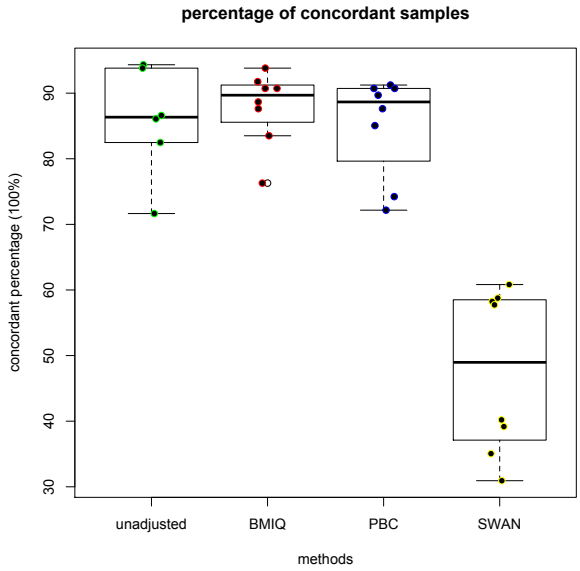


Figure 24: Percentage of concordant samples for evaluated algorithms

Table 13: Membership comparison between simulated data sets and overall gold standard

Subsets	Data set1 Mad1k	Data set1 Mad2k	Data set2 Mad1k	Data set2 Mad2k	Data set3 Mad1k	Data set3 Mad2k	Data set4 Mad1k	Data set4 Mad2k
Unadjusted	27	34	26	55	11	12	NA	NA
BMIQ	24	32	22	46	16	18	18	12
PBC	29	50	24	54	17	18	18	20
SWAN	80	76	134	126	81	82	116	118

For further comparison, the membership of 27k_mad1k and 450k_mad1k was used as the gold standard for simulated data sets (unadjusted, BMIQ, and PBC) with 1k probes. The

membership of 27k_mad2k and 450k_mad2k was used as the gold standard for simulated data sets (unadjusted, BMIQ, and PBC) with 2k probes. The number of discordant samples is shown in ***Table 14****. There was a significant difference in the number of discordant samples between data sets with 1k probes and 2k probes (Wilcox test p-value = 0.0017). However, there was still no significant difference between the three situations unadjusted, BMIQ and PBC (Wilcox test, unadjusted and BMIQ p-value = 0.6092; Wilcox test, unadjusted and PBC p-value = 0.9075; Wilcox test, BMIQ and PBC p-value = 0.692).*

Table 14: Membership comparison between data sets and separate gold standard

	Separate gold standard	*Data set1 Mad1k*	*Data set2 Mad1k*	*Data set3 Mad1k*	*Data set4 Mad1k*
Unadjusted	*27k_mad1k*	*33*	*26*	*12*	*NA*
	450k_mad1k	*21*	*22*	*24*	*NA*
BMIQ	*27k_mad1k*	*27*	*19*	*15*	*17*
	450k_mad1k	*22*	*29*	*30*	*23*
PBC	*27k_mad1k*	*26*	*25*	*12*	*15*
	450k_mad1k	*23*	*33*	*30*	*13*
Unadjusted	*27k_mad2k*	*40*	*57*	*17*	*NA*
	450k_mad2k	*36*	*57*	*25*	*NA*
BMIQ	*27k_mad2k*	*36*	*59*	*9*	*20*
	450k_mad2k	*35*	*51*	*33*	*22*
PBC	*27k_mad2k*	*52*	*58*	*15*	*28*
	450k_mad2k	*51*	*61*	*29*	*29*

Even though there was no significant difference between the unadjusted data sets and adjusted data sets (BMIQ and PBC), the adjusted method BMIQ still had a more stable and smaller number of discordant samples than the unadjusted and adjusted PBC method. Data set adjustment has been suggested by many papers (64, 66, 74, 75). Therefore, the BMIQ adjustment method was used to adjust the GBM DNA methylation data of GBM from HM450k before analysis.

Based on this data-driven analysis, it suggests that BMIQ is the best adjustment method. This can be applied to the HM450k data and then merged with HM27k data to increase the sample size.

****end of the citation****

2.2 Methods

2.2.1 DNA methylation data processing

All gliomas with HM450k data available from TCGA were included. Samples were processed using the UniD package (**Figure 25**): remove probes not mapped to the autosomal chromosomes; remove probes with SNPs within 10 bases of the targeted CpG site (snp-hit) (76); remove probes whose sequence aligns nonspecifically (i.e. aligned to more than one location in the genome; multi-hit) (76). Then in a sample-wise style, probes with bead count less than or equal to 3 or which have not been significantly detected compared to the background (with a detection p-value > 0.05) were set as missing values. Probes with more than 10% missing values across the samples were deleted. Samples with more than 5% missing values across all probes were deleted due to bad quality. The remaining missing values were imputed using k-nearest neighbor (KNN) algorithm.

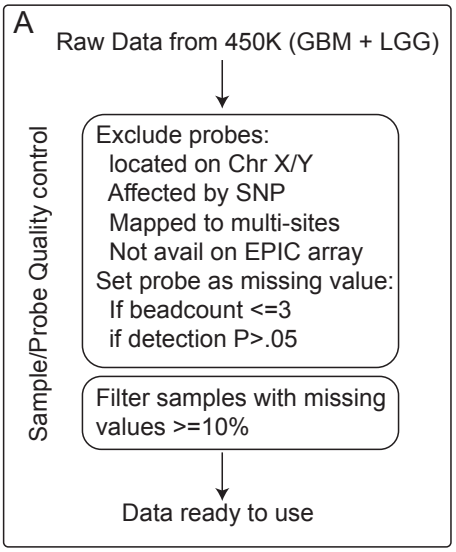


Figure 25: Data processing procedures for binary genomic alterations

2.2.2 Somatic mutation annotation

Genomic alterations, including mutation status of *IDH*, *ATRX*, and *TERT*p, and deletion status of chr1p19q were included. These have either mutant/codel or wild type/intact status. Gliomas in TCGA have been evaluated in multiple assays, including whole-exome sequencing, whole-genome sequencing, Affymetrix SNP6, gene expression microarray, mRNA sequencing, and DNA methylation microarray. Therefore, in my study, I could use either DNA or PCR sequencing data as a reference for mutation status of *IDH*, *ATRX*, and *TERT*p and use the chr1p19q codel status obtained from SNP6 platform as a reference label. Those data were directly extracted from Ceccarelli et al. (15).

2.2.3 Model building

After DNA methylation data processing and reference annotation extraction, I obtained different numbers of samples for each binary genomic alteration (**Table 15**). Then for each binary genomic alteration, all samples were randomly split into three subsets with their binary status as the stratification factor: a training set (60%), a development set (20%), and a test set (20%). For each genomic alteration, four steps (**Figure 26**) were used to build the model: variable selection, parameter tuning, model selection, and model validation.

Table 15: Data sets for binary genomic alterations

Genomic alterations	#total samples	Training samples	Development samples	Test samples	Reference labels
IDH	637	383	127	127	DNA-seq
*TERT*p	298	179	60	59	PCR-seq
ATRX	637	383	128	126	DNA-seq
Chr1p19q	641	385	129	127	SNP6

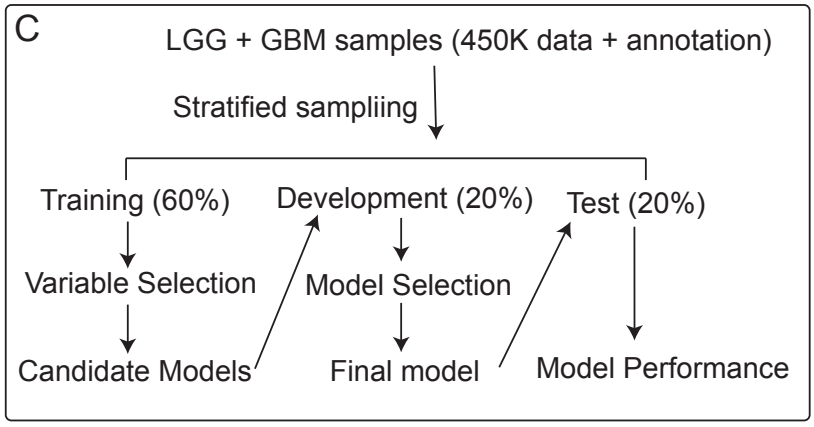


Figure 26: Model building for binary genomic alterations

The training set was used for variable selection and to build candidate models, then the candidate models were applied to the development set. Based on the prediction accuracy of the development set, the final model was selected. The final model was applied to the test set for model performance evaluation.

Elastic net was used to select the most important probes from all available probes. Two parameters are important here: alpha (α) and lambda (λ).

$$\frac{1-\alpha}{2}\left|\left|\beta\right|\right|_2^2 + \alpha||\beta||_1 \ (0 \leq \alpha \leq 1)$$

If alpha = 1, it is a lasso penalty; and if alpha = 0, it is a ridge penalty.

For the training set, the parameter alpha (R package: glmnet (77)) was set from 0.1 to 1, using 0.1 as a step. For each alpha value, 200 lambda values were randomly generated. Among the 200 lambda values, the best lambda value was picked out according to prediction accuracy. For each alpha and lambda value combination, 5-fold CV was applied in the training data set. For each fold among the CV, a set of probes was selected to build the generalized linear model with a nonzero coefficient. By summarizing the selection results among 5-fold by CV, the percentage for each probe selected among the 5-folds was calculated. Therefore, for each alpha value, I obtained a set of probes and its selection percentages. Then I was able to combine the selected probe sets among ten alpha values and calculate their overall selection percentage. Probes were ranked by their selection percentage from high to low.

Based on the probe selection percentage ranking, different sets of top probes, with different numbers of probes were selected from high to low. For example, the top 100 probes, 200 probes, 500 probes, and so on were selected. For each probe set, a generalized linear model was refit with the training set. Parameters were set as follows: alpha was set from 0.1 to 1, with 0.1 as a step; for each alpha value, lambda was set from 0 to 5, with 0.05 as a step. For each probe set, the best alpha and lambda value combination was picked out based on their prediction accuracy in the development set. The final model was built using the selected parameter combination of alpha and lambda and the best performing probe set. The test set was unseen during the whole model building process and was used to evaluate the final model performance. The prediction accuracy was used as the main evaluation metric. Tumor purity obtained from ABSOLUTE (78) was compared by ANOVA among the training, development, and test sets to evaluate the tumor purity's effect on model performance.

2.2.4 Prediction results analysis

A predictive model was built for each genomic alteration. After the final model was determined, all samples were rerun with the model and the methylation-based model prediction status was saved and further compared with the reference annotations from TCGA.

For *ATRX* mutation status, samples were regrouped by the DNA-seq and methylation-based (methyl-based) status. Misclassified samples were cases with discordant mutation status between DNA-seq and the methyl-based model and correctly classified samples were cases with concordant status between DNA-seq and the methyl-based model. To investigate whether the mutation types show any patterns among the misclassified samples, single nucleotide variations (SNVs) were obtained for misclassified samples. SNVs called with MuTect2 (79), VarScan (80), MuSE (81), and Somaticsniper (82) were collected from Genomic Data Commons Data Portal (83) and compared to the misclassified samples.

To validate whether the *ATRX* mutation affects mRNA expression levels, *ATRX* expression levels among subgroups were compared using a two-sample t-test. Because the DNA

methylation level may affect the mRNA expression level, the methylation level of HM450k probes located on *ATRX* were compared among subgroups using ANOVA test.

For chr1p19q codel status, samples were regrouped into four subgroups by comparing the chr1p19q codel status by SNP6 or by methyl-based model. For chr1p19q codel misclassified samples, samples' CNV profiles were derived from HM450k probe data using the R package conumee (84) for visualization and validation.

For gene expression subtype prediction, misclassified samples in the test set (n = 72) were regrouped by their transcriptional subtypes and predicted methylation subtypes. The correctly classified samples and misclassified samples were compared in terms of characteristic genes' CNVs and expression level using the Wilcoxon rank sum test, such as *EGFR*, *NF1*, and *CDKN2A*.

2.2.5 Signature analysis

Signature probes of binary genetic alterations were aligned to the genome and analyzed for genomic context using a two-sample proportion test (function prop.test() in R). The following categories were compared for each signature:

Chromosome enrichment. The number of probes located on each chromosome was summarized and normalized by the total number of probes available on the chromosome. The percentage for each chromosome was calculated by the normalized percent. A proportional test (R function, prop.test) was applied to compare the number of probes and the total number of probes available for each chromosome.

CpG island relationship enrichment. Probes were categorized into following six classes based on their distance to CpG island: CpG island, N_shlef (2 - 4kb from island), S_shlef, N_shore (1 - 2kb from island), S_shore, and unknown. The first class was used if probes were annotated with multiple categories. A proportional test was applied to compare the number of probes and the total number of probes available for each category.

Gene structure enrichment. Probes were categorized into one of the following seven classes based on their relationship to functional gene structure: TSS200, TSS1500, Body, 3'-UTR, 5'-UTR, 1st Exon, and unknown. For probes that could be mapped to multiple gene structure, only the first category was used. The number of probes was compared with the total available probes in each category using proportional test.

Mapping genes. Genes mapped by the probes were summarized for frequency by array annotation. If one probe could map to multiple genes, then the first gene was counted.

Gene ontology (GO) enrichment. Genes mapped by signature probes were analyzed with the DAVID tool (85) for GO enrichment analysis.

For each binary genetic alteration, the DNA methylation probes in the predictive model were compared between the binary status (mutant versus wild type or intact versus codel) using the Wilcoxon rank sum test. The most significantly different probes were selected and used for unsupervised clustering. Predictive signatures of *IDH*, *TERT*p, *ATRX* mutation, and chr1p19q codel were further compared with the G-CIMP signature, previously shown to be correlated with *IDH* status (12). All statistical analyses were performed using R version 3.5.0.

2.2.6 External data validation

For binary genetic alterations, the phase III clinical trial NOA-04 (86) was used as an independent and external validation set. This trial compared the efficacy and safety of radiotherapy followed by chemotherapy at progression to chemotherapy followed by radiotherapy at progression in patients with anaplastic gliomas (n = 115). DNA methylation HM450k data were available for all tumor samples. Most of the tumors were characterized by genomic alterations and the following data served as the reference standard for comparison: targeted resequencing of the amplified mutational hotspot (PCR-seq) for *IDH* (n = 108) and *TERT*p mutation (n = 99); multiplex ligation-dependent probe amplification (MLPA) for chr1p19q codel (n = 99); and immunohistochemistry (IHC) for *ATRX* mutation (n = 96). In addition, *IDH* mutation status was also determined using unsupervised clustering analysis of HM450k data.

Chr1p19q codel status was also obtained by reviewing the CNV profiles derived from HM450k data (n = 115) (R package conumee) (84). The methyl-based binary genetic alterations were predicted using UniD models for all samples and compared to reference standards. The *MGMT* promoter methylation statuses obtained by MS-PCR (87) were compared with MGMT-STP27 predictions.

2.2.7 Comparison to existing CNS methylation-based classification

All gliomas samples from TCGA with HM450k methylation data were run through the CNS methylation-based classification using the online tool. Samples were regrouped into nine subsets by their methyl-based predicted genetic alterations, including *IDH*, *ATRX*, and *TERT*p mutation and chr1p19q codel. The calibrated prediction CNS methylation-based class categories were summarized for all nine subsets. The overall survival time and status were compared among the major subsets. The *MGMT* promoter methylation status was compared using MGMT-STP27 with proportional tests to avoid the survival bias.

2.3 Results

2.3.1 Predicted results

2.3.1.1 sample information

I used 129 GBM and 516 LGG with HM450k data available in the data analysis. After QC and probe filtering, I excluded one LGG sample that did not pass the sample QC from the data set. After probe filtering, the final data set included 644 gliomas samples with 380,010 probes. Among the 644 gliomas samples, 637 of them had *IDH* and *ATRX* mutation status available, 298 samples had *TERT*p mutation status annotated, and 641 of them had chr1p19q codel status annotated.

2.3.1.2 Predictive models building

***IDH* mutation prediction**: I selected 1513 probes in the variable selection step: the top 20, 50, 100, 200, 500, 1,000, and 1,500 probes based on importance were selected and used in

the candidate models (**Figure 27**). The *IDH* mutation final model used 100 probes with alpha = 0, and lambda = 1. The final prediction accuracy was 100% in the test set. The area under the curve (AUC) of the receiver operator characteristic (ROC) was 1.0.

***TERT*p mutation prediction**: I selected 2,325 probes in the variable selection step. The top 50, 100, 200, 500, 1000, 1,500, and 2,000 probes were evaluated in the model selection step (**Figure 28**). The *TERT*p mutation final model used 1,000 probes with the same parameter settings in *IDH* prediction model. The prediction accuracy in the test was 98.3%, with an AUC of 1.0.

***ATRX* mutation prediction**: I used 2,112 probes after variable selection. The top 50, 100, 200, 500, 1,000, and 1,500 probes were evaluated in the model selection step (**Figure 29**). The final model used 500 probes with the same parameter values as *IDH* model. The prediction accuracy in the internal was 90.48% with an AUC of 0.9952.

chr1p19q codel prediction: In variable selection, 1,279 probes showed non-zero importance. Again, the same modeling process was applied except that the top 20, 50, 100, 200, 500, 1,000, and 1,279 probes were selected for model selection steps (**Figure 30**). The final model used the top 100 probes with alpha = 0 and lambda = 0.1 and reached 97.67% accuracy in testing set. The final model has an AUC of 0.9974.

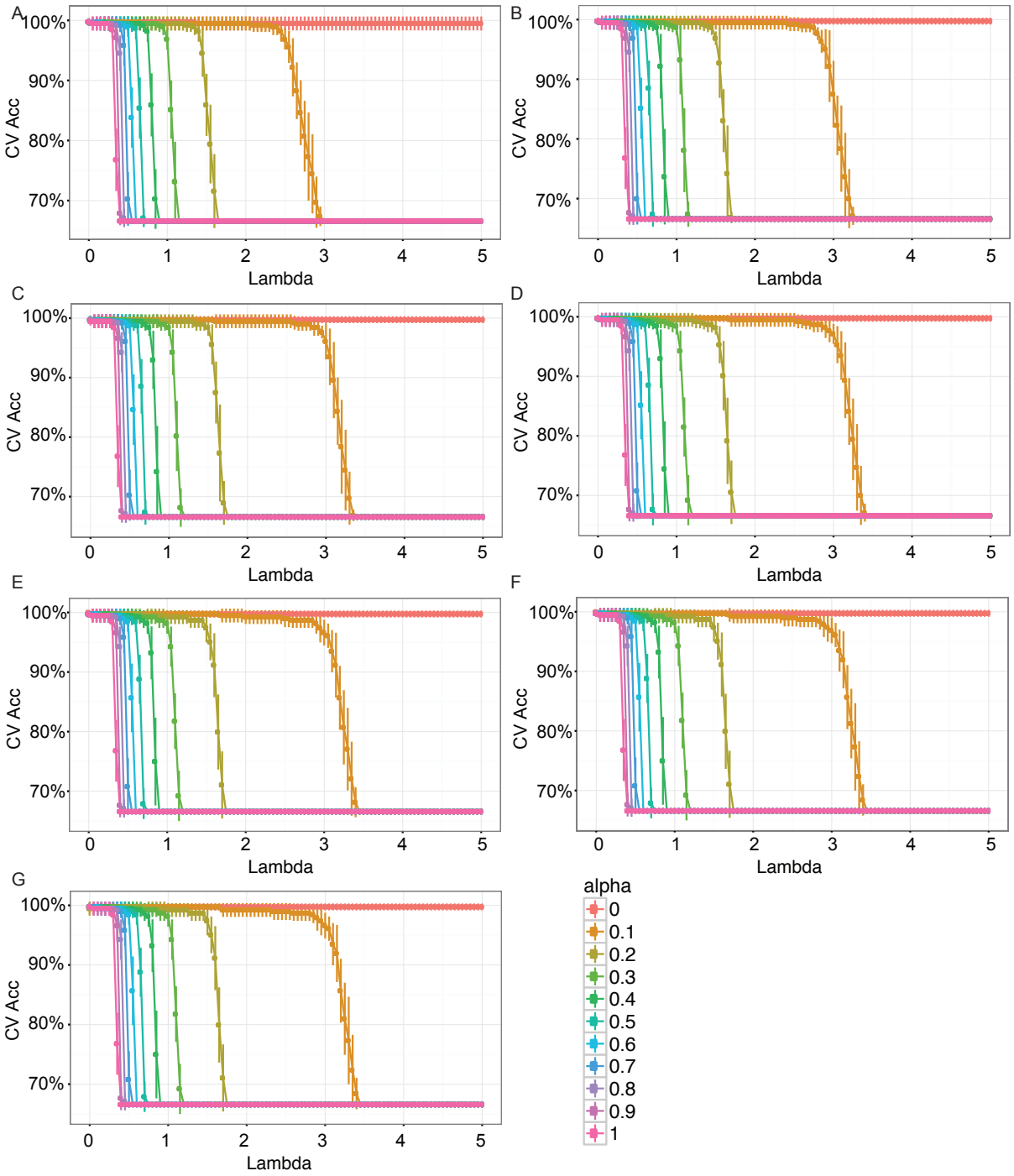


Figure 27: Prediction models' performance of *IDH* mutation in the training set.

Each figure used a different number of probes and applied different alpha and lambda combinations. The x-axis represents the lambda value and y-axis represent the prediction accuracy among the 5-fold CV. Lines with different colors represent the different alpha values as shown in the legend. Figures **A** to **G** show the prediction accuracy using the top 20, 50, 100, 200, 500, 1,000, and 1,500 probes when fitting the model, respectively.

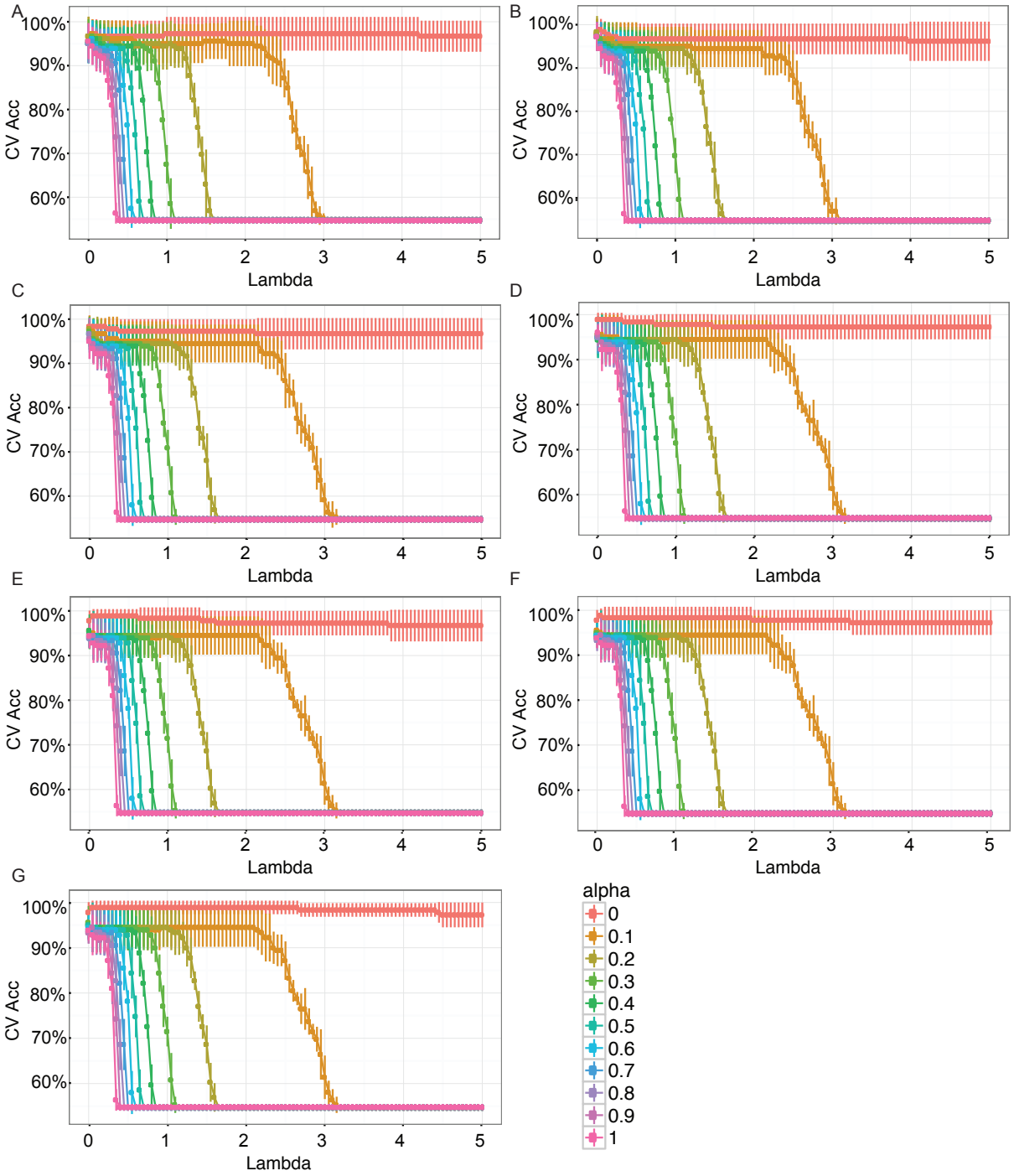


Figure 28: Prediction model's performance of *TERT*p mutation in the training set

Each figure used a different number of probes and applied different alpha and lambda combinations. The x-axis represents the lambda value and y-axis represent the prediction accuracy among the 5-fold CV. Lines with different colors represent the different alpha values as shown in the legend. Figures **A** to **G** show the prediction accuracy using the top 50, 100, 200, 500, 1000, 1,500, and 2,000 probes when fitting the model, respectively.

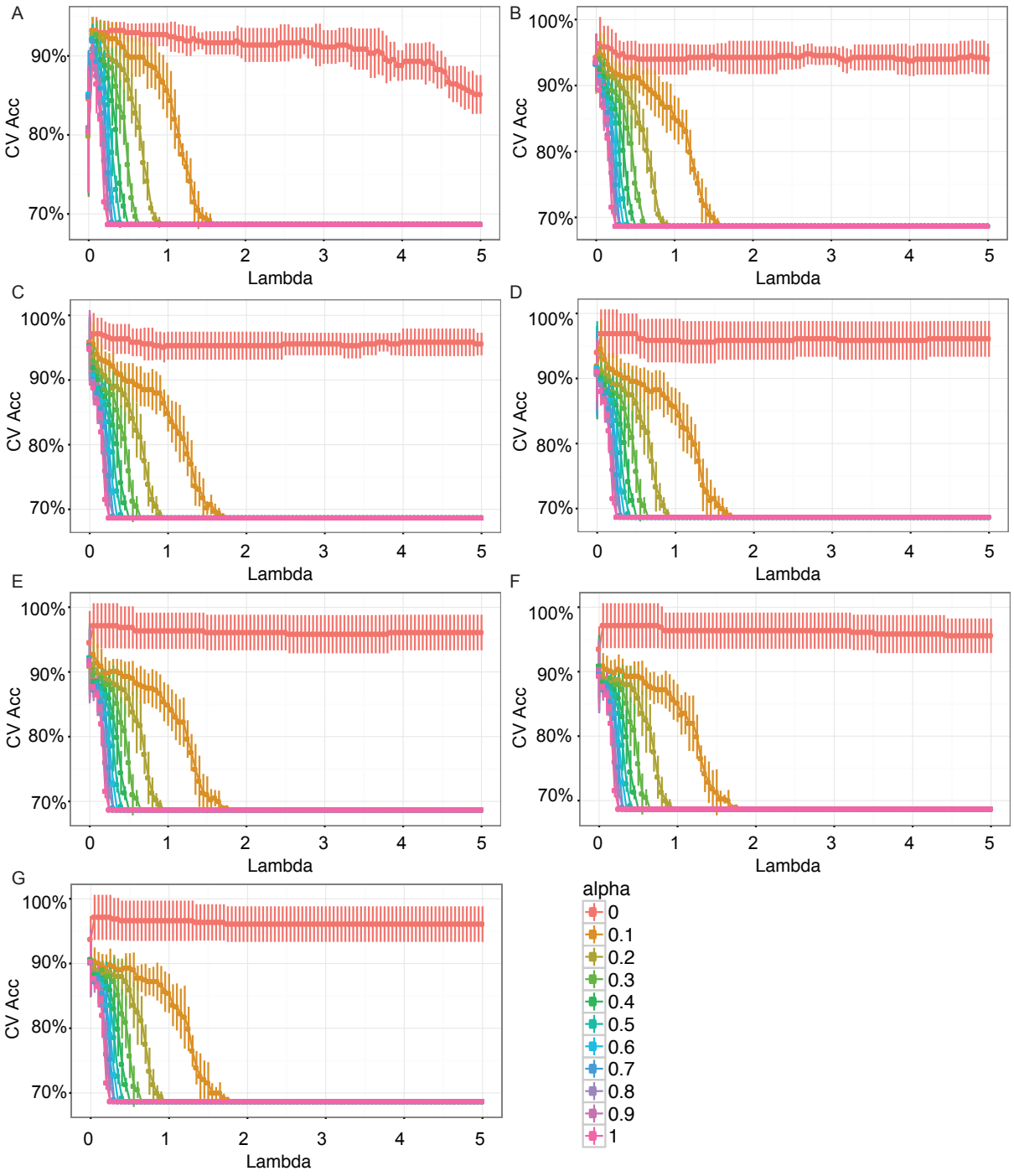


Figure 29: Prediction models' performance of *ATRX* mutation in the training set

Each figure used a different number of probes and applied different alpha and lambda combinations. The x-axis represents the lambda value and y-axis represent the prediction accuracy among the 5-fold CV. Lines with different colors represent the different alpha values as shown in the legend. Figures **A** to **G** show the prediction accuracy using the top 50, 100, 200, 500, 1,000, 1,500, and 2,000 probes when fitting the model, respectively.

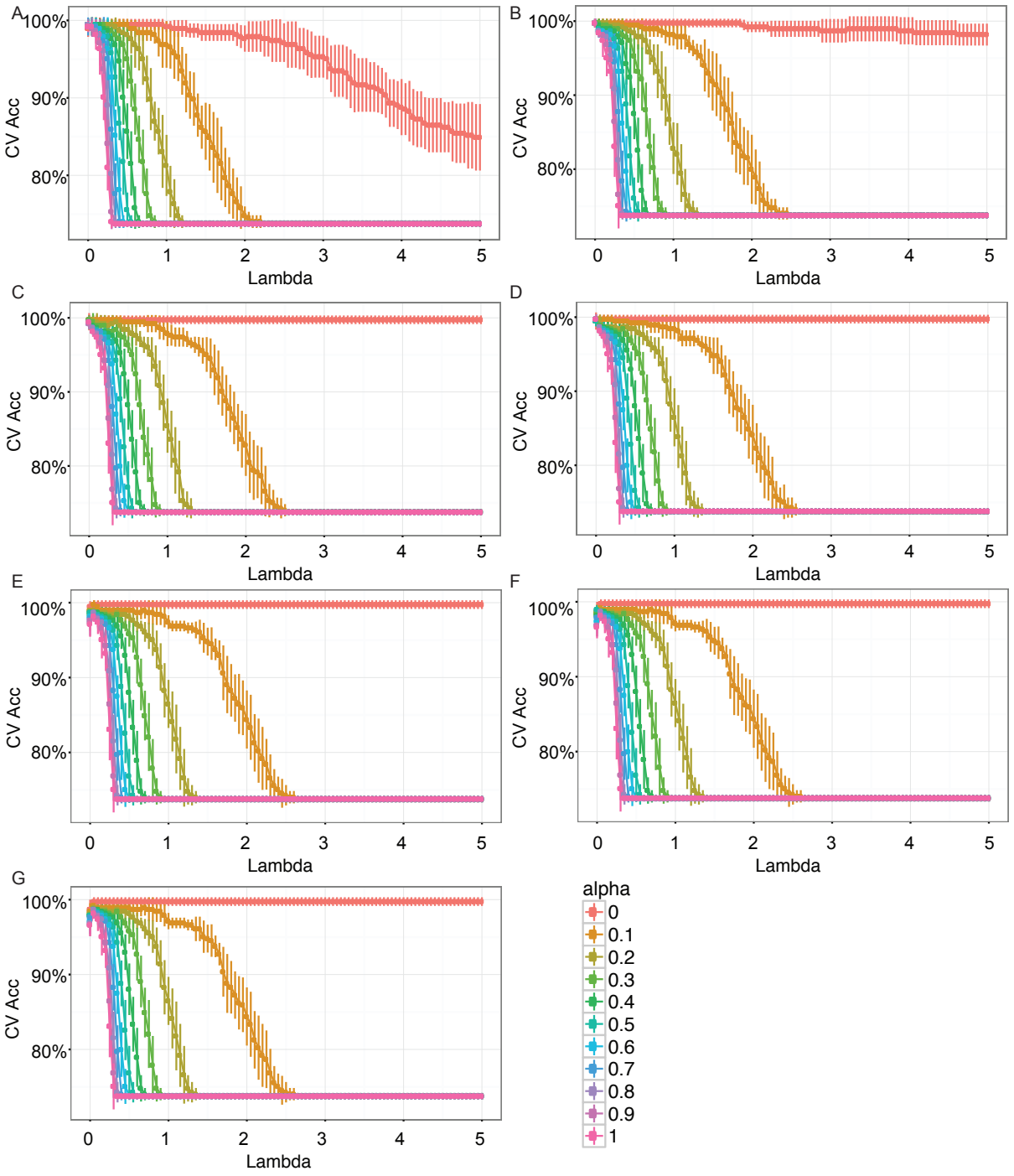


Figure 30: Prediction models' performance of chr1p19q codel in the training set

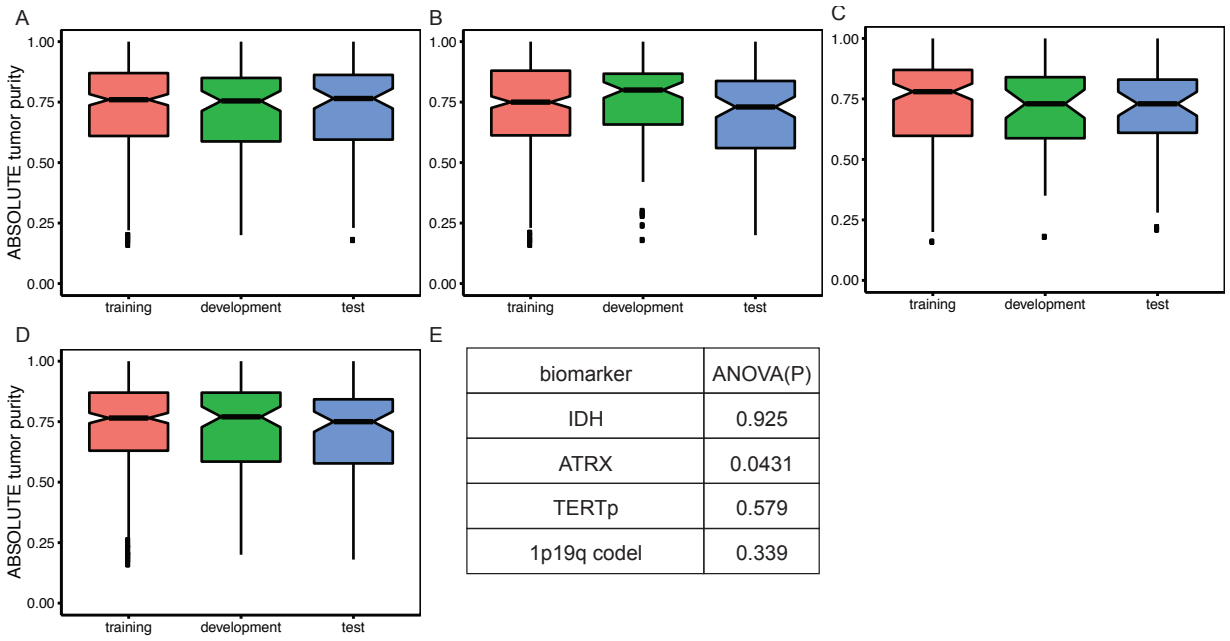
Each figure used a different number of probes and applied different alpha and lambda combinations. The x-axis represents the lambda value and y-axis represent the prediction accuracy among the 5-fold CV. Lines with different colors represent the different alpha values as shown in the legend. Figures **A** to **G** show the prediction accuracy using the top 20, 50, 100, 200, 500, 1,000, and 1,279 probes when fitting the model, respectively.

In summary, for binary genetic alterations, all predictive models achieved high prediction accuracy as shown in **Table 16**. The test set achieved a prediction accuracy of 100%, 98.3%, 90.48%, and 99.21% for *IDH*, *TERT*p, and *ATRX* mutation, and chr1p19q codel status, respectively.

Table 16: Binary genomic alteration predictive model performance summary

Genomic alterations	#probes in model	Reference label	#total samples	prediction accuracy (#samples)		
				training set	development set	test set
IDH	100	DNA-seq	637	100% (383/383)	100% (127/127)	100% (127/127)
*TERT*p	1000	PCR-seq	298	100% (179/179)	96.67% (58/60)	98.3% (58/59)
ATRX	500	DNA-seq	637	97.12% (372/383)	85.16% (109/128)	90.48% (114/126)
chr1p19q	100	SNP6	641	99.74% (384/385)	97.67% (126/129)	99.12% (126/127)

Tumor purity derived by ABSOLUTE algorithm (78) was compared between training, development, and test set for each evaluated genomic alteration. With the exception of tumors with *ATRX* mutation (p-value = 0.043, ANOVA test), no other sample sets showed significant difference in tumor purity (**Figure 31**).



biomarker	ANOVA(P)
IDH	0.925
ATRX	0.0431
TERTp	0.579
1p19q codel	0.339

Figure 31: Tumor purity comparison between training, development, and tests set for binary biomarker.

From **A** to **D**, each figure shows the tumor purity compared between training, development, and test sets for samples used in *IDH*, *ATRX*, *TERT*p, and chr1p19q codel, predictive model development. Y-axis is the tumor purity obtained from ABSOLUTE using whole-exome sequencing data. X-axis is the three subsets data used in model development. Figure **E** shows the P-value of ANOVA-test in comparing the tumor purity between the three subsets.

2.3.2 Predictive signature analysis

By comparing the number of probes enriched for chromosomes after normalization, I found that probes in the *IDH* mutation prediction model were enriched in chromosomes 22 (13.08%) and 21 (8.8%), while probes in the *ATRX* mutation prediction model were enriched in chromosomes 9 (7.2%) and 14 (7.2%). Interestingly, probes in the *TERT*p mutation, chr1p19q codel, and gene expression subtype prediction models were enriched in chromosome 18 (*TERT*p: 8.2%; chr1p19q codel: 16.3%; gene expression subtype: 12.6%) (**Table 17, 18, 19, 20**).

Table 17 Chromosome enrichment analysis for *IDH* mutation prediction signatures

Genomic alteration	CHR	# probes available	# probes	normalized percent	percent % (sum to 1)	p-value (proportion Test)
IDH mutation	1	36923	8	0.022%	3.343%	0.1314
	2	27711	13	0.047%	7.239%	
	3	20279	7	0.035%	5.326%	
	4	16241	2	0.012%	1.900%	
	5	19398	3	0.015%	2.386%	
	6	28392	5	0.018%	2.717%	
	7	23237	3	0.013%	1.992%	
	8	16401	2	0.012%	1.882%	
	9	7783	4	0.051%	7.930%	
	10	18784	6	0.032%	4.929%	
	11	23487	8	0.034%	5.256%	
	12	20018	5	0.025%	3.854%	
	13	9794	2	0.020%	3.151%	
	14	12234	2	0.016%	2.523%	
	15	12402	4	0.032%	4.977%	
	16	17917	3	0.017%	2.584%	
	17	23237	3	0.013%	1.992%	
	18	4939	1	0.020%	3.124%	
	19	21429	8	0.037%	5.761%	
	20	8823	3	0.034%	5.247%	
	21	3507	2	0.057%	8.800%	
	22	7074	6	0.085%	13.088%	

Table 18: Chromosome enrichment analysis for *TERT*p mutation prediction signatures

Genomic alteration	CHR	# probes available	# probes	normalized percent	percentage (sum to 1)	p-value (proportion Test)
*TERT*p mutation	1	36923	88	0.238%	4.023%	0.0000352
	2	27711	81	0.292%	4.934%	
	3	20279	60	0.296%	4.994%	
	4	16241	45	0.277%	4.677%	
	5	19398	54	0.278%	4.699%	
	6	28392	74	0.261%	4.399%	
	7	23237	85	0.366%	6.175%	
	8	16401	44	0.268%	4.528%	
	9	7783	18	0.231%	3.904%	
	10	18784	32	0.170%	2.876%	
	11	23487	51	0.217%	3.665%	
	12	20018	41	0.205%	3.457%	
	13	9794	25	0.255%	4.309%	
	14	12234	28	0.229%	3.863%	
	15	12402	33	0.266%	4.491%	
	16	17917	38	0.212%	3.580%	
	17	23237	65	0.280%	4.722%	
	18	4939	24	0.486%	8.202%	
	19	21429	49	0.229%	3.860%	
	20	8823	44	0.499%	8.418%	
	21	3507	5	0.143%	2.407%	
	22	7074	16	0.226%	3.818%	

Table 19: Chromosome enrichment analysis for *ATRX* mutation prediction signatures

Genomic alteration	CHR	# probes available	# probes	normalized percent	percent % (sum to 1)	p-value (proportion Test)
ATRX mutation	1	36923	56	0.152%	5.347%	0.01635
	2	27711	37	0.134%	4.708%	
	3	20279	18	0.089%	3.129%	
	4	16241	18	0.111%	3.908%	
	5	19398	28	0.144%	5.089%	
	6	28392	30	0.106%	3.725%	
	7	23237	32	0.138%	4.855%	
	8	16401	27	0.165%	5.804%	
	9	7783	16	0.206%	7.248%	
	10	18784	14	0.075%	2.628%	
	11	23487	30	0.128%	4.503%	
	12	20018	28	0.140%	4.932%	
	13	9794	3	0.031%	1.080%	
	14	12234	25	0.204%	7.205%	
	15	12402	17	0.137%	4.833%	
	16	17917	33	0.184%	6.494%	
	17	23237	26	0.112%	3.945%	
	18	4939	3	0.061%	2.142%	
	19	21429	33	0.154%	5.430%	
	20	8823	15	0.170%	5.994%	
	21	3507	3	0.086%	3.016%	
	22	7074	8	0.113%	3.987%	

Table 20: Chromosome enrichment analysis for chr1p19q codel prediction signatures

Genomic alteration	CHR	# probes available	# probes	normalized percent	percent % (sum to 1)	p-value (proportion Test)
1p19q codel	1	36923	3	0.008%	1.305%	0.003735
	2	27711	6	0.022%	3.477%	
	3	20279	6	0.030%	4.751%	
	4	16241	6	0.037%	5.932%	
	5	19398	6	0.031%	4.966%	
	6	28392	11	0.039%	6.221%	
	7	23237	13	0.056%	8.983%	
	8	16401	7	0.043%	6.853%	
	9	7783	1	0.013%	2.063%	
	10	18784	8	0.043%	6.838%	
	11	23487	4	0.017%	2.735%	
	12	20018	4	0.020%	3.208%	
	13	9794	1	0.010%	1.639%	
	14	12234	3	0.025%	3.937%	
	15	12402	1	0.008%	1.295%	
	16	17917	1	0.006%	0.896%	
	17	23237	4	0.017%	2.764%	
	18	4939	5	0.101%	16.255%	
	19	21429	3	0.014%	2.248%	
	20	8823	5	0.057%	9.099%	
	21	3507	0	0.000%	0.000%	
	22	7074	2	0.028%	4.540%	

Summarizing the dispersion of probes by chromosome, I found most of the probes were enriched on CpG islands. Among the four predictive models, the *IDH* predictive signature showed the highest percentage of CpG islands (76%) (**Table 21**). For the other three predictive signatures, about 32% probes were located on CpG islands (**Table 22, 23, 24**).

Table 21: CpG island relationship enrichment for *IDH* mutation prediction signatures

Genomic alteration	Relation to CpG island	# probes available	# probes	Normalization to total #probes	Percentage (sum to 1)	p-value (proportion test)
IDH mutation	Island	120312	89	0.074%	75.9958%	<2.2e-16
	N_Shelf	18946	0	0.000%	0.0000%	
	N_Shore	50338	5	0.010%	10.2043%	
	S_Shelf	16765	0	0.000%	0.0000%	
	S_Shore	39412	5	0.013%	13.0332%	
	not categorized	134237	1	0.001%	0.7653%	

Table 22: CpG island relationship enrichment for *TERT*p mutation prediction signatures

Genomic alteration	Relation to CpG island	# probes available	# probes	Normalization to total #probes	Percentage which sum to 1	p-value (proportion test)
*TERT*p mutation	Island	120312	536	0.446%	33.104%	<2.2e-16
	N_Shelf	18946	30	0.158%	11.766%	
	N_Shore	50338	100	0.199%	14.762%	
	S_Shelf	16765	31	0.185%	13.740%	
	S_Shore	39412	74	0.188%	13.952%	
	not categorized	134237	229	0.171%	12.676%	

Table 23: CpG island relationship enrichment for *ATRX* mutation prediction signatures

Genomic alterations	Relation to CpG island	# probes available	# probes	Normalization to total #probes	Percentage which sum to 1	p-value (proportion test)
ATRX mutation	Island	120312	261	0.217%	31.313%	<2.2e-16
	N_Shelf	18946	18	0.095%	13.713%	
	N_Shore	50338	71	0.141%	20.359%	
	S_Shelf	16765	6	0.036%	5.166%	
	S_Shore	39412	54	0.137%	19.777%	
	not categorized	134237	90	0.067%	9.677%	

Table 24: CpG island relationship enrichment for chr1p19q prediction signatures

Genomic alterations	Relation to CpG island	# probes available	# probes	Normalization to total #probes	Percentage which sum to 1	p-value (proportion test)
chr1p19q codel	Island	120312	53	0.044%	32.777%	0.0001692
	N_Shelf	18946	1	0.005%	3.927%	
	N_Shore	50338	14	0.028%	20.693%	
	S_Shelf	16765	4	0.024%	17.752%	
	S_Shore	39412	7	0.018%	13.215%	
	not categorized	134237	21	0.016%	11.640%	

Among the 100 probes with *IDH* predictive signatures, 45% of the probes (45/100) were located in the promoter region (including TSS200, TS1500, and 1st Exon) (**Table 25**). In total 65 genes were mapped by those 100 probes. Among all genes, CASP8 and FADD like apoptosis regulator (*CFLAR*) genes had four probes mapped and nuclear receptor subfamily 4, group A, member 1 (*NR4A1*) had three probes mapped. Applying those 65 genes to the DAVID (85, 88) (version 6.7) for functional annotation, the top GOs enriched pathways were regulation of apoptosis (p-value = 0.0052), regulation of programmed cell death (p-value = 0.0056), and regulation of cell death (p-value = 0.0057) (**Table 26**). Genes related to those GOs were *CFLAR*, tumor necrosis factor receptor superfamily member 6 (*FAS*); potassium voltage-gated channel interacting protein 3 (*KCNIP3*); death effector domain containing 2 (*DEDD2*); lectin, galactoside-binding, soluble, 1 (*LGALS1*); *NR4A1*; proline dehydrogenase 1 (*PRODH*); retinoic acid receptor, gamma (*RARG*); and erb-b2 receptor tyrosine kinase 2 (*ERBB2*). Those GOs were not significant after p-value adjustment due to the small gene set.

Table 25: Gene structure enrichment for *IDH* mutation prediction signatures

Genomic Alterations	relation to gene structure	# probes available	#probes	normalized #probe	Percentage (sum to 1)	**p-value (proportion test)**
IDH mutation	TSS200	41774	20	0.048%	22.055%	0.00004473
	TSS1500	55088	13	0.024%	10.871%	
	Body	126827	23	0.018%	8.354%	
	3'UTR	13641	0	0.000%	0.000%	
	5'UTR	33719	14	0.042%	19.127%	
	1st Exon	18147	12	0.066%	30.462%	
	not categorized	90814	18	0.020%	9.131%	

Table 26: GO enrichment analysis results for *IDH* mutation prediction signatures

Category	Term	Count	P-Value	Benjamini-adjust p-value
GOTERM_BP_FAT	sensory organ development	6	0.001	0.51
GOTERM_BP_FAT	regulation of apoptosis	9	0.0052	0.84
GOTERM_BP_FAT	regulation of programmed cell death	9	0.0056	0.72
GOTERM_BP_FAT	regulation of cell death	9	0.0057	0.63
GOTERM_BP_FAT	muscle cell differentiation	4	0.008	0.67
GOTERM_BP_FAT	positive regulation of apoptosis	6	0.015	0.82
GOTERM_BP_FAT	positive regulation of programmed cell death	6	0.015	0.78
GOTERM_BP_FAT	positive regulation of cell death	6	0.016	0.74
GOTERM_BP_FAT	negative regulation of cell activation	3	0.018	0.76
GOTERM_BP_FAT	regulation of lymphocyte differentiation	3	0.019	0.74
GOTERM_BP_FAT	regulation of cell activation	4	0.022	0.74
GOTERM_BP_FAT	induction of apoptosis	5	0.023	0.74
GOTERM_BP_FAT	induction of programmed cell death	5	0.023	0.71
GOTERM_BP_FAT	cartilage development	3	0.026	0.73
GOTERM_BP_FAT	regulation of cell size	4	0.033	0.78
GOTERM_BP_FAT	regulation of T cell differentiation in the thymus	2	0.037	0.8
GOTERM_BP_FAT	embryonic eye morphogenesis	2	0.04	0.81
GOTERM_BP_FAT	homeostatic process	7	0.041	0.8
GOTERM_BP_FAT	regulation of cell proliferation	7	0.049	0.84

For *TERT*p predictive signatures, most of the probes were located at the body (29.1%) (**Table 27**). Probes were mapped to 612 genes in total, and the most frequently mapped gene was isthmin 1 (*ISM1*) with ten probes. The second most frequently mapped gene was atlastin GTPase 3 (*ATL3*) with seven probes. For gene functional annotation, the most significant GOs were all related to the regulation of transcription (**Table 28**).

Table 27: Gene structure enrichment for *TERT*p mutation prediction signatures

Genomic Alterations	relation to gene structure	# probes available	#probes	normalized #probe	Percentage (sum to 1)	p-value (proportion test)
*TERT*p mutation	TSS200	41774	162	0.388%	16.700%	< 2.2e-16
	TSS1500	55088	130	0.236%	10.162%	
	Body	126827	291	0.229%	9.881%	
	3'UTR	13641	96	0.704%	30.306%	
	5'UTR	33719	34	0.101%	4.342%	
	1st Exon	18147	79	0.435%	18.747%	
	not categorized	90814	208	0.229%	9.863%	

Table 28: GO enrichment analysis results for *TERT*p mutation prediction signatures

Category	Term	Count	P-Value	Benjamini-adjust p-value
GOTERM_BP_FAT	regulation of transcription	131	9.100E-08	0.00022
GOTERM_BP_FAT	regulation of transcription, DNA-dependent	97	2.300E-07	0.00028
GOTERM_BP_FAT	regulation of RNA metabolic process	97	6.400E-07	0.00052
GOTERM_BP_FAT	regulation of transcription from RNA polymerase II promoter	49	2.400E-06	0.0014
GOTERM_BP_FAT	negative regulation of transcription, DNA-dependent	30	5.700E-06	0.0027
GOTERM_BP_FAT	negative regulation of transcription from RNA polymerase II promoter	25	6.600E-06	0.0027
GOTERM_BP_FAT	negative regulation of RNA metabolic process	30	7.800E-06	0.0027
GOTERM_BP_FAT	negative regulation of macromolecule biosynthetic process	39	9.200E-06	0.0028
GOTERM_BP_FAT	negative regulation of biosynthetic process	40	1.100E-05	0.003
GOTERM_BP_FAT	negative regulation of cellular biosynthetic process	39	1.600E-05	0.0039
GOTERM_BP_FAT	negative regulation of transcription	34	1.800E-05	0.004
GOTERM_BP_FAT	negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	35	6.700E-05	0.014
GOTERM_BP_FAT	negative regulation of nitrogen compound metabolic process	35	8.800E-05	0.016
GOTERM_BP_FAT	negative regulation of gene expression	34	1.100E-04	0.019
GOTERM_BP_FAT	negative regulation of macromolecule metabolic process	44	1.400E-04	0.022
GOTERM_BP_FAT	transcription	98	1.700E-04	0.025
GOTERM_BP_FAT	pattern specification process	22	1.800E-04	0.025
GOTERM_BP_FAT	positive regulation of gene expression	36	3.500E-04	0.046
GOTERM_BP_FAT	positive regulation of transcription	35	4.200E-04	0.053

For *ATRX* predictive signature, most of the probes were located at the body (13.9%) (**Table 29**). In total, probes were mapped to 333 genes. Gene F-box protein 6 (*FBXO6*) had five probes mapped, and cathepsin F (*CTSF*) gene was mapped by four probes. Using the DAVID gene functional annotation, genes were significantly enriched in development related GOs, including embryonic development and neuron development (**Table 30**). I further compared the 70 probes overlapping the *ATRX* predictive signature and *TERT*p predictive signature; 52 genes

were mapped. Most of the genes were enriched during transcription regulation. The most frequently mapped gene was *FBXO6* and phosphodiesterase 7B (*PDE7B*). The top GOs enriched for those overlapping genes was cell - cell signaling (**Table 31**).

Table 29: Gene structure enrichment for *ATRX* mutation prediction signatures

Genomic Alterations	relation to gene structure	# probes available	#probes	normalized #probe	Percentage (sum to 1)	p-value (proportion test)
ATRX mutation	TSS200	41774	71	0.170%	14.037%	< 2.2e-16
	TSS1500	55088	67	0.122%	10.044%	
	Body	126827	139	0.110%	9.051%	
	3'UTR	13641	55	0.403%	33.299%	
	5'UTR	33719	10	0.030%	2.449%	
	1st Exon	18147	46	0.253%	20.934%	
	not categorized	90814	112	0.123%	10.185%	

Table 30: GO enrichment analysis results for *ATRX* mutation prediction signatures

Category	Term	Count	P-Value	Benjamini-adjust p-value
GOTERM_BP_FAT	chordate embryonic development	18	6.100E-05	0.097
GOTERM_BP_FAT	embryonic development ending in birth or egg hatching	18	6.800E-05	0.055
GOTERM_BP_FAT	neuron fate commitment	7	7.600E-05	0.042
GOTERM_BP_FAT	endocrine system development	8	1.800E-04	0.072
GOTERM_BP_FAT	regionalization	12	6.200E-04	0.19
GOTERM_BP_FAT	embryonic skeletal system development	7	2.100E-03	0.44
GOTERM_BP_FAT	pattern specification process	13	2.300E-03	0.43
GOTERM_BP_FAT	gland development	9	2.300E-03	0.39
GOTERM_BP_FAT	regulation of cell development	11	2.900E-03	0.42
GOTERM_BP_FAT	regulation of phosphorylation	18	2.900E-03	0.39
GOTERM_BP_FAT	skeletal system development	14	3.600E-03	0.42
GOTERM_BP_FAT	muscle organ development	11	3.600E-03	0.4
GOTERM_BP_FAT	neuron differentiation	17	3.800E-03	0.39
GOTERM_BP_FAT	muscle organ morphogenesis	3	4.200E-03	0.4

Table 31: overlapped probes between *ATRX* and *TERT*p predictive signature mapped genes enrichment GO

Category	Term	Count	P-Value	Benjamini-adjust p-value
GOTERM_BP_FAT	cell-cell signaling	11	8.400E-04	0.7
GOTERM_BP_FAT	regulation of cell communication	15	2.000E-03	0.77
GOTERM_BP_FAT	positive regulation of cellular component biogenesis	6	2.100E-03	0.64
GOTERM_BP_FAT	regulation of signaling	15	2.400E-03	0.57
GOTERM_BP_FAT	regulation of transcription from RNA polymerase II promoter	11	3.600E-03	0.64
GOTERM_BP_FAT	cellular component assembly	13	5.400E-03	0.72
GOTERM_BP_FAT	regulation of cellular component biogenesis	7	6.200E-03	0.72
GOTERM_BP_FAT	positive regulation of developmental process	8	6.700E-03	0.7
GOTERM_BP_FAT	regulation of multicellular organismal development	10	7.700E-03	0.71
GOTERM_BP_FAT	positive regulation of multicellular organismal process	9	8.400E-03	0.7
GOTERM_BP_FAT	transcription from RNA polymerase II promoter	10	1.100E-02	0.76
GOTERM_BP_FAT	cellular component biogenesis	13	1.300E-02	0.78
GOTERM_BP_FAT	positive regulation of macromolecule biosynthetic process	9	1.600E-02	0.83
GOTERM_BP_FAT	protein secretion	5	1.600E-02	0.81
GOTERM_BP_FAT	regulation of cellular component organization	11	1.800E-02	0.83
GOTERM_BP_FAT	positive regulation of transcription from RNA polymerase II promoter	7	2.000E-02	0.84
GOTERM_BP_FAT	regulation of signal transduction	12	2.000E-02	0.82

For chr1p19q codel prediction signature, 44% of probes mapped to the promoter region, including TSS200, TSS1500, and 1st Exon (**Table 32**). Four probes mapped to gene *ATL3* and fibroblast growth factor receptor 2 (*FGFR2*). The top two GOs were regulation of cellular protein metabolic process (p-value = 0.001) and bone morphogenetic proteins (*BMP*) signaling pathway (p-value = 0.0098). However, no GOs were significant after Benjamin p-value adjustment (**Table 33**).

Table 32: Gene structure enrichment for chr1p19q codel prediction signatures

Genomic Alterations	relation to gene structure	# probes available	#probes	normalized #probe	Percentage (sum to 1)	p-value (proportion test)
chr1p19q codel	TSS200	41774	16	0.038%	17.652%	0.005519
	TSS1500	55088	19	0.034%	15.896%	
	Body	126827	24	0.019%	8.722%	
	3'UTR	13641	2	0.015%	6.757%	
	5'UTR	33719	15	0.044%	20.503%	
	1st Exon	18147	9	0.050%	22.858%	
	not categorized	90814	15	0.017%	7.613%	

Table 33: GO enrichment analysis results for chr1p19q codel mutation prediction signatures

Category	Term	Count	P-Value	Benjamini-adjust p-value
GOTERM_BP_FAT	regulation of cellular protein metabolic process	8	1.000E-03	0.44
GOTERM_BP_FAT	BMP signaling pathway	3	9.800E-03	0.94
GOTERM_BP_FAT	regulation of protein modification process	5	1.800E-02	0.96
GOTERM_BP_FAT	regulation of protein amino acid phosphorylation	4	2.100E-02	0.95
GOTERM_BP_FAT	tissue morphogenesis	4	2.300E-02	0.93
GOTERM_BP_FAT	enzyme linked receptor protein signaling pathway	5	2.800E-02	0.93
GOTERM_BP_FAT	negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	6	2.900E-02	0.91
GOTERM_BP_FAT	negative regulation of nitrogen compound metabolic process	6	3.100E-02	0.89
GOTERM_BP_FAT	negative regulation of macromolecule biosynthetic process	6	3.700E-02	0.91
GOTERM_BP_FAT	negative regulation of cellular biosynthetic process	6	4.100E-02	0.9
GOTERM_BP_FAT	negative regulation of biosynthetic process	6	4.400E-02	0.9
GOTERM_BP_FAT	positive regulation of cellular protein metabolic process	4	4.500E-02	0.88
GOTERM_BP_FAT	morphogenesis of an epithelium	3	4.600E-02	0.87
GOTERM_BP_FAT	transmembrane receptor protein serine/threonine kinase signaling pathway	3	4.800E-02	0.86
GOTERM_BP_FAT	positive regulation of protein metabolic process	4	4.900E-02	0.85
GOTERM_BP_FAT	ossification	3	5.800E-02	0.88
GOTERM_BP_FAT	bone development	3	6.500E-02	0.89

For each binary genetic alteration, samples were clustered into two subgroups with the highest ranked significantly different probes and showed high consistency with the known genomic alteration (**Figure 32A**). By comparing the signature probes of *IDH*, *TERT*p, *ATRX*, and chr1p19q codel with the G-CIMP signature, I found that no probes overlapped among these five probe signatures (**Figure 32B**). The lack of overlap between *ATRX* and *TERT*p mutation

signatures was consistent with the mutually exclusive nature of *ATRX* and *TERT*p in telomere maintenance(89).

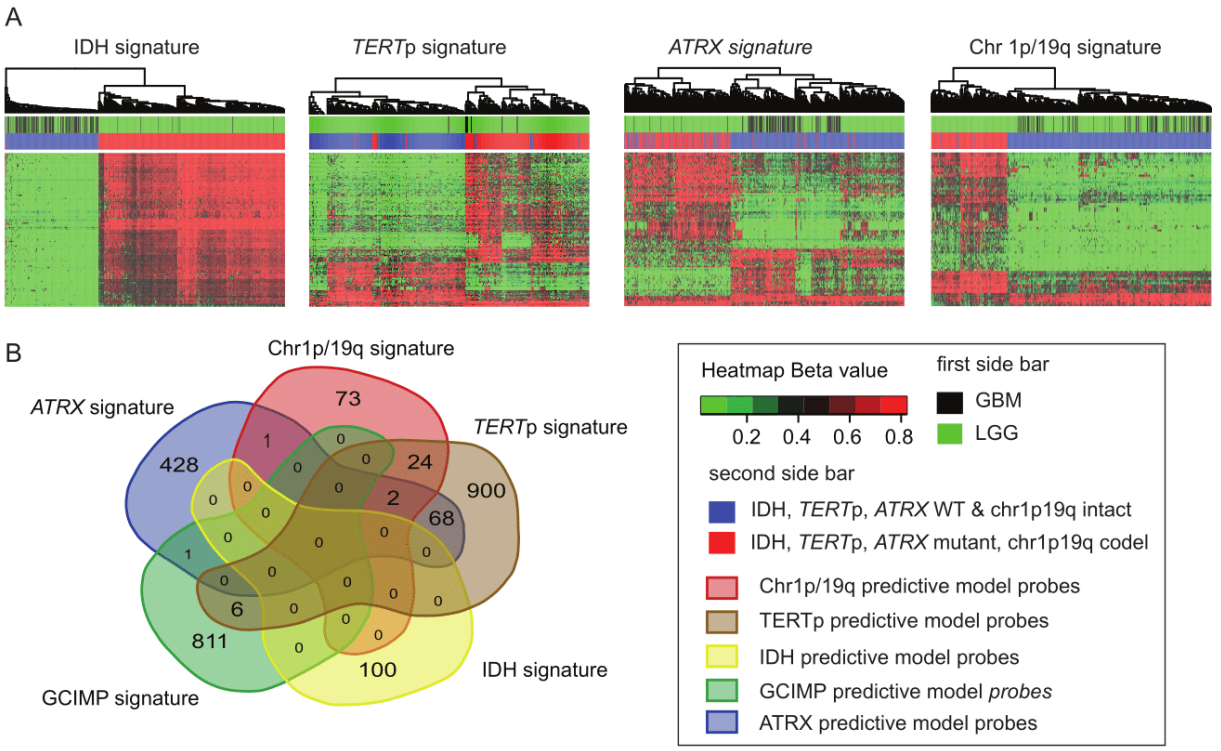


Figure 32: Methyl-based predictive signature analysis for binary genomic alterations

A Heatmaps of DNA methylation level (β value): samples are in columns and DNA methylation probes are in rows. The two top sidebars show the sample source and genetic alteration status. From left to right, each image shows the most significant probes in signatures of *IDH* (number of probes = 100), *TERT*p (number of probes = 200), *ATRX* (number of probes = 90), and chr1p19q codel (number of probes = 70). **B** In this Venn diagram, the predictive model probes of each binary genetic alteration (number of probes: *IDH* = 100, *TERT*p = 1,000, *ATRX* = 500, chr1p19q codel = 100) and the GCIMP probes (number of probes = 818) identified in published paper (12) were compared with each other. Different colors represent different probe sets. The overlapping blocks between any probes sets indicate the overlapping probes. The number within the diagram indicates the number of probes within a specific block.

2.3.3 Prediction results analysis

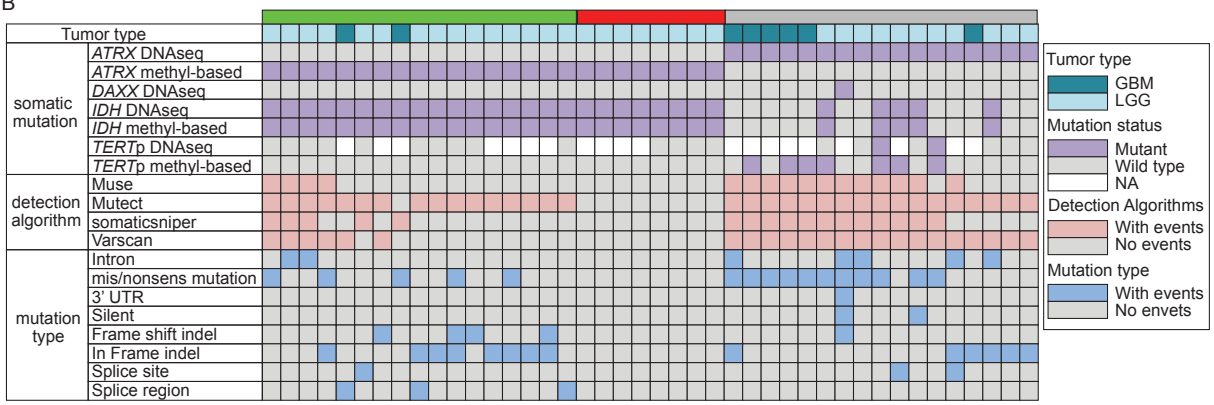
For *ATRX* mutation status, 42 samples were misclassified by the methylation-based model. To investigate the mutation types (such as missense, nonsense, frameshift, and so on) of those samples, the SNV information was collected and compared among misclassified samples.

Five sets (sets 1 to 5) were formed based on the DNA-seq based *ATRX* status (mutant or wild type), methyl-based *ATRX* status (mutant or wild type), and SNV information (with mutation detected or not detected) (**Figure 33A**). Twenty-five samples were classified as wild type by DNA-seq but mutant by methyl-based model (**Table 34**). Among these 25 samples, 17 samples (set 2) showed at least 1 mutation call from the 4 SNV algorithm results and 8 samples (set 3) had no mutation calls according to the SNVs (**Figure 33B**). For set 4, samples with *TERT*p mutation status, 3 of the 8 were *TERT*p mutant and wild type for *ATRX*. All samples misclassified as *ATRX* mutant by methyl-based model harbored *IDH* mutation while all samples misclassified as wild type by methyl-based model were *IDH* wild type (**Figure 33B**). A mutation type shift occurred between set 2 (*ATRX* DNA-seq wild type, methyl-based mutant, and with mutation calls by reviewing SNV information) and set 4 (*ATRX* DNA-seq mutant, methyl-based wild type): the enriched mutations shifted from frameshift indels and in-frame indels to intron, missense and nonsense which may not lead to loss of function. No significant differences in *ATRX* gene expression were observed between sets for which methylation results agreed, even when the sequencing result was discordant between those sets. Conversely, when the methylation results were discordant, even when the sequencing results were in an agreement, a significant difference in expression was observed (**Figure 33CD**). The DNA methylation level of probes located on *ATRX* did not show significant differences among the three subsets with the exception of a single probe (**Figure 34**).

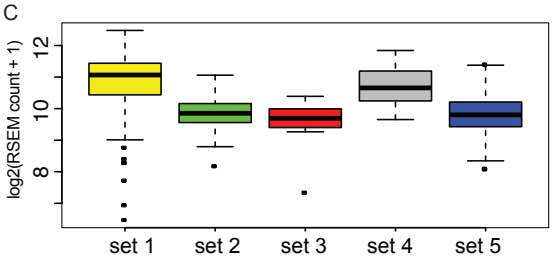
A

Subsets		set 1	set 2	set 3	set 4	set 5
ATRX	DNA-seq	WT	WT	WT	Mutant	Mutant
	methyl-based	WT	Mutant	Mutant	WT	Mutant
Mutation calls		NA	Yes	No	Yes	NA
No. of cases		413	17	8	17	182

B



C



D

P.value*	set 2	set 3	set 4	set 5
set 1	3.97e-7	5.54e-05	0.2177	<2.2e-16
set 2	NA	0.3503	0.0012	0.837
set 3	NA	NA	0.0015	0.4367
set 4	NA	NA	NA	1e-4

*p-value: wilcoxon rank sum test

Figure 33: Investigation of misclassified samples for *ATRX* prediction

A All samples (n = 637) used to build the methyl-based predictive model for *ATRX* are classified into five subgroups based on DNA-seq and methyl-based *ATRX* mutation status and on whether they had mutation calls by reviewing the SNV information. **B** The misclassified 42 samples are shown with their tumor type, mutation status, mutation calling algorithms, and detailed mutation type. **C** Boxplots show the *ATRX* gene expression level for each set. By applying Wilcoxon rank sum test, set 2 and set 3 samples show significantly lower ATRX expression level than set 1 (set 2 versus set 1: the estimated difference was 1.134 and 95%CI was 0.67 to 1.47, p-value = 3.97×10^{-7}, set 3 versus set 1: the estimated difference was 1.33 and 95%CI was 0.83-1.81, p-value = 5.54×10^{-5}) and set4 showed significantly higher *ATRX* expression level compared to set 5 (the estimated difference was 0.87 and 95% CI 0.47 to 1.27, p-value = 1×10^{-4}). **D** T-test was used to compare *ATRX* gene expression level between every two subsets. The p-value is provided for each comparison in the table.

Table 34: *ATRX* prediction results analysis for misclassified samples

sample	ATRX status		Find in algorithms (Yes/No)				#detected times	IDH status
	DNA-seq	Methyl	Muse	somaticsniper	Varscan	Mutect		
TCGA-WY-A85C-01A	WT	mutant	No	No	No	No	0	Mutant
TCGA-WY-A859-01A	WT	mutant	No	No	No	No	0	Mutant
TCGA-WH-A86K-01A	WT	mutant	No	No	No	No	0	Mutant
TCGA-S9-A7R4-01A	WT	mutant	No	No	No	No	0	Mutant
TCGA-P5-A5F4-01A	WT	mutant	No	No	No	No	0	Mutant
TCGA-DH-5143-01A	WT	mutant	No	No	No	No	0	Mutant
TCGA-DB-5277-01A	WT	mutant	No	No	No	No	0	Mutant
TCGA-CS-5393-01A	WT	mutant	No	No	No	No	0	Mutant
TCGA-26-1442-01A-01D	WT	mutant	No	Yes	No	No	1	Mutant
TCGA-CS-6665-01A-11D	WT	mutant	No	No	No	Yes	1	Mutant
TCGA-HT-7606-01A-11D	WT	mutant	No	No	No	Yes	1	Mutant
TCGA-HT-8018-01A-11D	WT	mutant	No	No	No	Yes	1	Mutant
TCGA-HW-A5KM-01A	WT	mutant	No	No	No	Yes	1	Mutant
TCGA-S9-A7R3-01A-11D	WT	mutant	No	No	No	Yes	1	Mutant
TCGA-S9-A89Z-01A-11D	WT	mutant	No	No	No	Yes	1	Mutant
TCGA-TQ-A7RF-01A	WT	mutant	No	No	No	Yes	1	Mutant
TCGA-TQ-A7RW-01A	WT	mutant	No	No	No	Yes	1	Mutant
TCGA-TQ-A8XE-01A	WT	mutant	No	No	No	Yes	1	Mutant
TCGA-06-6389-01A-11D	WT	mutant	No	No	Yes	Yes	2	Mutant
TCGA-DB-A4XB-01A	WT	mutant	No	Yes	No	Yes	2	Mutant
TCGA-S9-A7J0-01A-11D	WT	mutant	No	No	Yes	Yes	2	Mutant
TCGA-DU-8167-01A-11D	WT	mutant	Yes	No	Yes	Yes	3	Mutant
TCGA-DB-A4X9-01A-11D	WT	mutant	Yes	Yes	Yes	Yes	4	Mutant
TCGA-HT-7601-01A-11D	WT	mutant	Yes	Yes	Yes	Yes	4	Mutant
TCGA-HW-8321-01A	WT	mutant	Yes	Yes	Yes	Yes	4	Mutant
TCGA-06-A5U0-01A-11D	mutant	WT	No	No	Yes	Yes	2	WT
TCGA-DU-7298-01A-11D	mutant	WT	No	No	Yes	Yes	2	Mutant
TCGA-HT-7469-01A-11D	mutant	WT	No	No	Yes	Yes	2	WT
TCGA-HT-7857-01A-11D	mutant	WT	No	No	Yes	Yes	2	WT
TCGA-DU-5852-01A-11D	mutant	WT	No	Yes	Yes	Yes	3	WT
TCGA-S9-A89V-01A-11D	mutant	WT	Yes	No	Yes	Yes	3	WT
TCGA-06-5858-01A-01D	mutant	WT	Yes	Yes	Yes	Yes	4	WT
TCGA-06-6388-01A-12D	mutant	WT	Yes	Yes	Yes	Yes	4	WT
TCGA-06-6391-01A-11D	mutant	WT	Yes	Yes	Yes	Yes	4	WT
TCGA-4W-AA9T-01A-11D	mutant	WT	Yes	Yes	Yes	Yes	4	WT
TCGA-74-6575-01A-11D	mutant	WT	Yes	Yes	Yes	Yes	4	WT
TCGA-DH-5144-01A-01D	mutant	WT	Yes	Yes	Yes	Yes	4	Mutant
TCGA-DU-6392-01A-11D	mutant	WT	Yes	Yes	Yes	Yes	4	WT
TCGA-FG-5963-01A-11D	mutant	WT	Yes	Yes	Yes	Yes	4	WT
TCGA-FG-7638-01B-11D	mutant	WT	Yes	Yes	Yes	Yes	4	Mutant
TCGA-FG-A713-01A-11D	mutant	WT	Yes	Yes	Yes	Yes	4	Mutant
TCGA-HT-7880-01A-11D	mutant	WT	Yes	Yes	Yes	Yes	4	Mutant

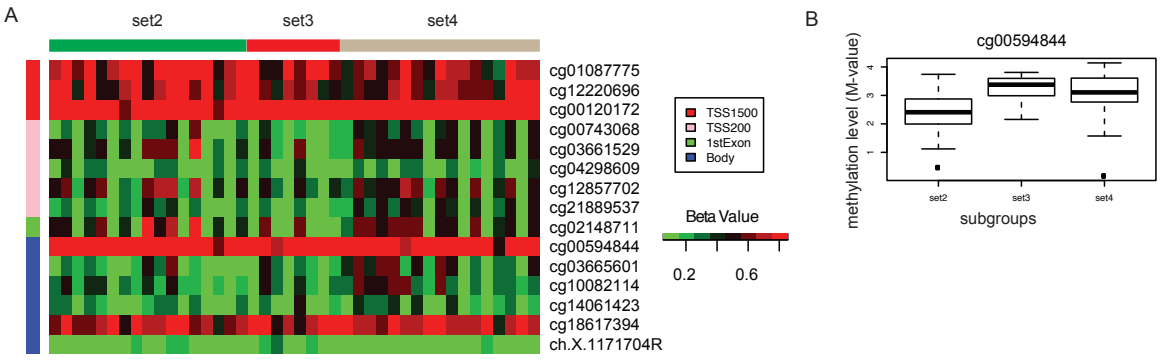


Figure 34: Investigation of HM450k probes located on *ATRX*.

A The heatmap shows the methylation beta value of HM450k probes located at the *ATRX* gene region (TSS200, TSS1500, 1st Exon, and body) for samples in set 2, set 3, and set 4. Each column represents one sample while each row represents one probe. The column-sidebar indicates the subset samples belong to while the row-sidebar indicates the annotations for probes. **B** The boxplot shows the comparison of DNA methylation level of probe cg00594844 among the set 2, set 3, and set 4. The y-axis shows the DNA methylation level (M-value) and the x-axis shows the three sets. This is the only probe that showed different methylation levels (p-value = 0.0213, ANOVA test) among the three subsets.

For the chr1p19q codel prediction model, samples were classified into four different sets (set 1 to 5) by their methyl-based *IDH* status, SNP6-based chr1p19q status, and methyl-based chr1p19q status (**Figure 35A**). I observed clear concordance between the methyl-based chr1p19q codel status and other known somatic mutations for different sets. Five samples were misclassified when comparing methyl-based status to SNP6-based status (**Figure 35B**). The CNV profile of chr1 and chr19 were derived from the HM450k methylation data using R package conumee(84) (**Figure 35C**). Four over five samples were misclassified as codel and one sample was misclassified as non-codel by methylation model. The deletion in the TCGA-CS-5394 and TCGA-FG-7637 which match with methyl-based model prediction is clear to see. The CNV profile pattern of the other three samples is not obvious; therefore, it is difficult to determine their status.

A

set	#cases	IDH	chr 1p/19q status		tumor grade		somatic mutation status	*TERT*p		*CIC**	*FUBP1*^	*ATRX*		*TP53*
		methyl-based	SNP6-based	methyl-based	LGG	GBM		PCR-seq	methyl-based	DNA-seq	DNA-seq	DNA-seq	methyl-based	DNA-seq
set1	214	WT	non-codel	non-codel	94	120	Mutant	43	174	0	0	13	1	31
							WT	19	40	186	186	197	213	155
							NA	152	0	28	28	4	0	28
set2	255	Mut	non-codel	non-codel	249	6	Mutant	7	7	223	224	182	208	212
							WT	140	248	3	2	70	47	14
							NA	108	0	29	29	3	0	29
set3	4	Mut	non-codel	codel	4	0	Mutant	0	2	1	1	0	0	0
							WT	2	2	2	2	4	4	3
							NA	2	0	1	1	0	0	1
set4	1	Mut	codel	non-codel	1	0	Mutant	0	1	0	0	0	0	0
							WT	1	0	1	1	1	1	1
							NA	0	0	0	0	0	0	0
set5	167	Mut	codel	codel	167	0	Mutant	85	166	98	40	3	0	6
							WT	1	1	54	112	164	167	146
							NA	81	0	15	15	0	0	15

B

	1p19q status		*IDH* status		*ATRX* status		*TERT*p status		*FUBP1*	*CIC*	*TP53*
	SNP6	methyl-based	DNA-seq	methyl-based	DNA-seq	methyl-based	PCR-seq	methyl-based	DNA-seq	DNA-seq	DNA-seq
TCGA.CS.5394	intact	codel	Mut	Mut	WT	WT	WT	WT	Mut	Mut	WT
TCGA.FG.7637	intact	codel	Mut	Mut	WT	WT	WT	WT	WT	WT	WT
TCGA.S9.A6WI	intact	codel	Mut	Mut	WT	WT	NA	Mut	WT	WT	WT
TCGA.VM.A8CA	intact	codel	Mut	Mut	WT	WT	NA	Mut	NA	NA	NA
TCGA.HT.8010	codel	intact	Mut	Mut	WT	WT	WT	Mut	WT	WT	WT

C

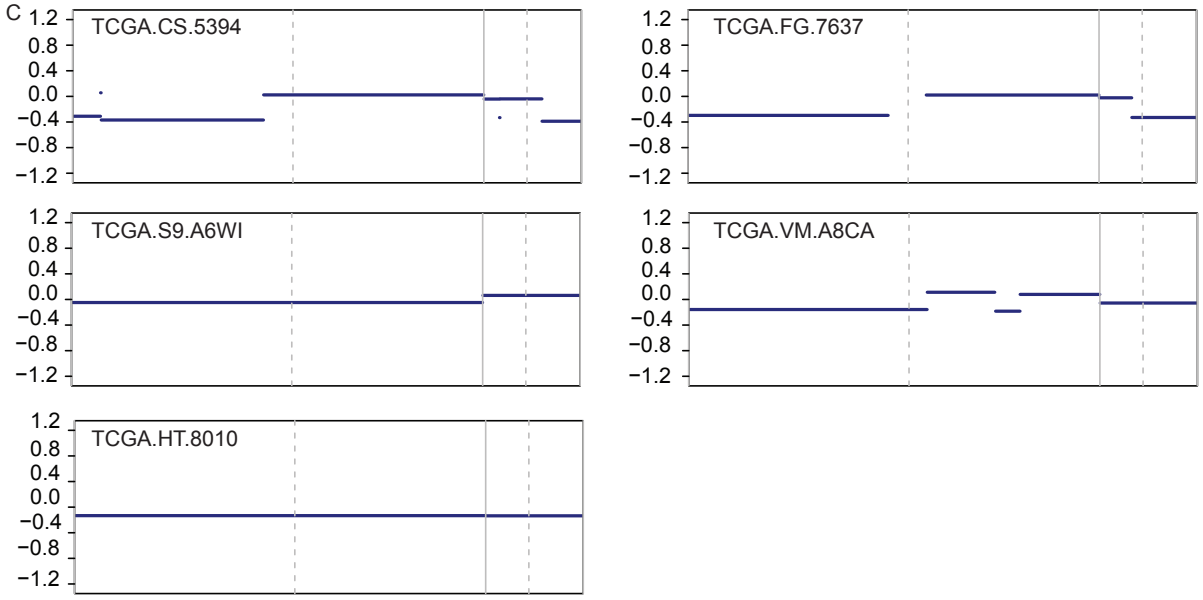


Figure 35: Investigation of misclassified samples for chr1p19q codel prediction model

A Genetic characterization of subsets (set 1 to set 6) according to chr1p19q codel status determined by SNP6 and methyl-based prediction. Set 1 samples (*IDH* wild type and chr1p19q intact from both SNP6-based and methyl-based) show frequent *TERT*p mutations and majority wild type status of *CIC*, *FUBP1*, *ATRX*, and *TP53*. Set 2 samples (*IDH* mutation and chr1p19q intact by both SNP6-based and methyl-based) show frequent mutation for *CIC*, *FUBP1*, *ATRX*, and *TP53*, and lacked *TERT*p mutation. Set 5 (*IDH* mutation and chr1p19q codel from both SNP6-based and methyl-based status) show frequent *TERT*p, *CIC*, and *FUBP1* mutations but not *ATRX* or *TP53* mutation. Two out of four samples in set 3 (methyl-based chr1p19q codel but

SNP6 CNV intact) show consistency mutation profile with set2. **B** Genetic characterization of the five misclassified samples by comparing methyl-based chr1p19q status to SNP6-based chr1p19q status. **C** CNV profile of chr1 and chr19 of the five misclassified samples were derived from HM450k data using R package conumee.

2.3.4 Model validation

The prediction accuracy for each biomarker in the NOA-04 cohort was: 89.9% (98/109) for *IDH* mutation by PCR-seq and 99.10% (114/115) for *IDH* mutation by unsupervised clustering analysis, 82.8% (82/99) for *TERT*p mutation by PCR-seq, 92.7% (89/96) for *ATRX* mutation by IHC, and 88.89% (88/99) for chr1p19q status by MLPA and 95.65% (110/115) for chr1p19q status by HM450k-based CNV profiles (**Table 35**). In terms of *IDH* mutation status, 11 samples were misclassified by methyl-based prediction: nine of them were predicted as wild type by PCR-seq and mutant by the methyl-based model. *MGMT* methylation status comparisons are shown in **Table 36**.

Table 35: Binary genomic predictive model validation using NOA-04 data set

Genomic alteration	#samples	reference label	prediction Accuracy
IDH	108	PCR-seq	89.90%
	115	HM450k clustering	99.10%
*TERT*p	99	PCR-seq	82.80%
ATRX	96	IHC	92.70%
chr1p19q	99	MLPA	89.99%
	115	HM450k derived CNV	95.65%

Table 36: *MGMT* promoter methylation status comparison between MS-PCR and MGMT-STP27 in NOA04 samples

MGMT-STP27 results	Methylation-specific PCR results	
	methylated	unmethylated
Unmethylated	3	13
methylated	69	29

2.3.5 Comparison to existing CNS methylation-based classification

Nine subsets were formed based on methyl-based biomarker classification (n = 644) and the majority of tumors fell into five of the nine subsets (**Figure 35A**). A subset of gliomas (39/644, set 8) with both wild type *ATRX* and *TERT*p suggested that alternative mechanisms existed to maintain their telomere length. CNS methylation-based classifications are summarized for each of the nine subsets (**Figure 36A**). Discordant samples between the two classification systems were described in detail in **Figure 36B-E**. First row of **Figure 36B**: 40/644 of all gliomas were classified as "CONTR" categories (classified-normal) which are normal brain tissue according to CNS methylation-based classification, while the remaining cases were classified into "tumor" categories (classified-tumor). The ABSOLUTE tumor purity showed significant differences between classified-normal samples and classified-tumor samples in **Figure 36C** ($P < 1\times10^{-5}$, T-Test). However, 48 classified-tumor samples also showed tumor purity equal to or less than the median tumor purity of classified-normal samples. Second row of **Figure 36B**: all CONTR, HEMI (methylation class control tissue, hemispheric cortex) in subgroups 1 to 4 are expected to be *IDH* wild type and all have been detected with *IDH* mutation by DNA sequencing. Third row of **Figure 36B**: Twelve samples in Grp2 (CONTR, HEMI; A IDH; and A IDH, HG) were classified as either normal brain normal tissue or *IDH* wild type glioma without chr1p19q codel while their CNV profile from SNP6 showed clear chr1p19q codel (**Figure 36D**). Fourth row of **Figure 36B**: SFT, HMPC (methylation class solitary fibrous tumor / hemangiopericytoma) samples are expected to have a euploid genome while TCGA-19-5951 in Grp7 showed significant chr10 loss and chr19p and chr20 amplification (**Figure 36E**)**.** Fifth row of **Figure 36B**: A IDH, HG in Grp8 were expected to be *IDH* mutant by CNS classification but, in fact, were wild type by sequencing. Sixth row: Two samples from adult patients (TCGA-06-5858 and TCGA-06-6698) were classified as IHG (infantile hemispheric glioma) by CNS methylation-based classification which is typically limited to infants. The log-rank test among the five enriched subgroups (Grp 1, 2, 3, 7, and 8) showed no significant difference in survival compared to Grp1, Grp2, and Grp3 (**Figure 36F and 36G**).

This indicates all patients with *IDH* mutant tumors have similar survival time regardless of their tumors' *ATRX*, *TERT*p, or chr1p19q status. *IDH* wild type glioma enriched subsets (Grp7 and Grp8), showed significantly worse survival from *IDH* mutant samples. Moreover, Grp7 demonstrated poorer survival compared with Grp8, whose samples harbored *TERT*p mutations, indicated the negative prognostic significance of *TERT*p mutation in the absence of *IDH* or *ATRX* mutation. Grp7 (75/176, 42.61%) also included a significantly higher percentage ($P < 0.01$) of tumors with *MGMT* promoter methylation compared to Grp8 (7/39, 17.95%).

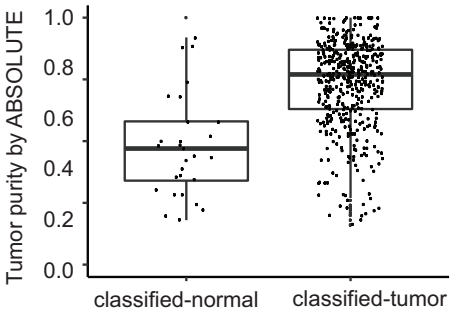
A

	IDH	*ATRX*	*TERT*p	chr1p19q	#cases	CNS methylation-based calibrated predicted classes
Grp1	Mutant	Mutant	WT	non-codel	208	A IDH(184) A IDH,HG(20) CONTR,HEMI(4)
Grp2	Mutant	WT	Mutant	codel	168	O IDH(156) CONTR,HEMI(6) A IDH(3) A IDH,HG(3)
Grp3	Mutant	WT	WT	non-codel	41	A IDH(33) A IDH,HG(6) CONTR,HEMI(2)
Grp4	Mutant	WT	Mutant	non-codel	7	A IDH(4) CONTR,HEMI(3)
Grp5	Mutant	WT	WT	codel	3	O IDH(3)
Grp6	Mutant	Mutant	Mutant	non-codel	1	A IDH(1)
Grp7	WT	WT	Mutant	non-codel	176	GBM,MES(71) GBM,RTK II (67) GBM,RTK I (24) CONTR,HEMI(9) CONTR,INFLAM(1) CONTR,REACT(1) GBM,RTK III (1) SFT,HMPC (1) SUBEPN,PF(1)
Grp8	WT	WT	WT	non-codel	39	CONTR,HEMI(7) CONTR,CEBM(2) CONTR,WM(1) CONTR,INFLAM(2) CONTR,HYPTHAL(1) CONTR,PONS(1) GBM,G34(2) GBM,MID(2) GBM,MYCN(1) GBM<RTK II (1) LGG,GG(4) LGG,MYC(1) LGG,PA PF(2) LGG,PA/GG ST(1) A IDH,HG(1) ANA,PA(1) CNS,NB,FOXR2(3) IHG(2) DMG,K27(1) HGNET,BCOR(1) PLEX,PEDB(1) PTPR,B(1)
Grp9	WT	Mutant	WT	non-codel	1	ANA PA(1)

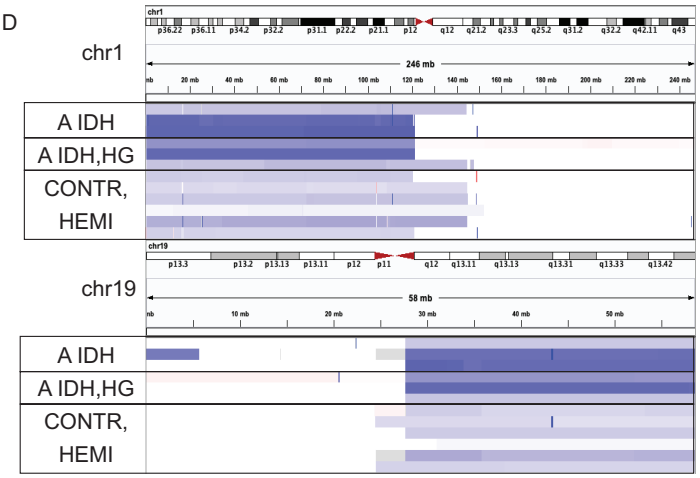
B

Discordant samples	Detail information
All CONTR samples (40)	gliomas classifed as normal brain tissue, tumor purity can't explain (**C**)
Grp1:CONTR,HEMI(4) Grp2:CONTR,HEMI(6) Grp3:CONTR,HEMI(2) Grp4:CONTR,HEMI(3)	Classified as normal samples, expected to be *IDH* wild type, but identified as *IDH* mutant samples (**A**)
Grp2:CONTR,HEMI(6) A IDH(3) A IDH,HG(3)	Not expect chr1p19q co-del, but CNV profile shown codel (**D**)
Grp7: SFT,HMPC (1)	Expect euploidy genome, but shown chr10 del and chr19.20 amp (**E**)
Grp8: A IDH, HG (1)	Classified as *IDH* mutant category, but identified as *IDH* wild type (**A**)
Grp9: IHG (2)	IHG happens in infant. Two cases were diagnosed at their 45 and 53 ys

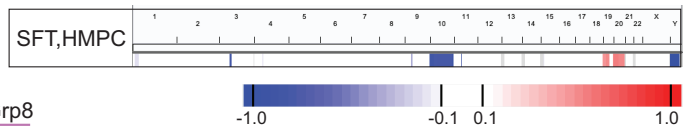
C



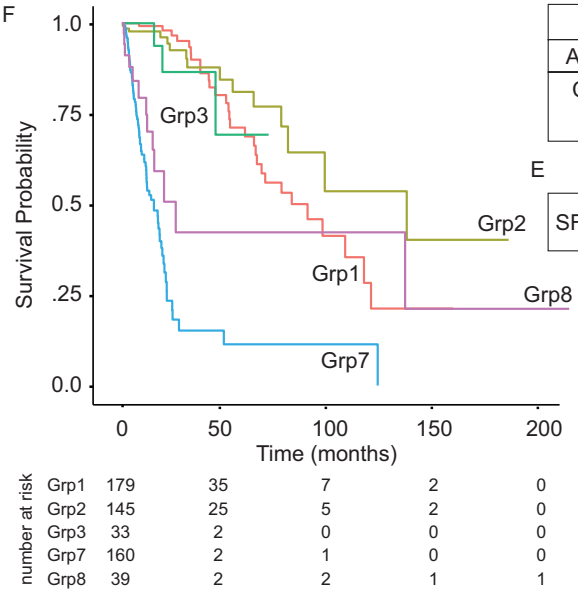
D



E



F



G

P-value	Grp1	Grp2	Grp3	Grp7
Grp2	0.3	NA	NA	NA
Grp3	0.6	0.5	NA	NA
Grp7	<2e-16	<2e-16	3e-6	NA
Grp8	5e-5	5e-6	0.006	0.03

Figure 36: DNA methyl-based glioma classification.

A All samples with HM450k data available were classified into nine subgroups according to their methyl-based genetic alterations. The number of samples for each subset is also provided. The CNS methylation-based calibrated categories are summarized for each subgroup in the rightmost column. **B** By comparing the predicted annotation between both methyl-based and CNS methylation-based classification, all discordant samples were picked out with detailed information. The left column shows the discordant samples and their subgroup belonging in **A**. The right column shows the rationale why they were discordant samples. **C** Boxplot showing the comparison of ABSOLUTE tumor purity between the samples classified into CONTR categories (classified-normal) and samples classified as tumor categories (classified-tumor). Each point represents one sample. **D** SNP6-based CNV profile of the 12 samples in Grp2 (CONTR, HEMI; A IDH; A IDH, HG) clearly showing chr1p19q codel. The upper and lower panel show the profiles of chr1 and chr19, respectively. **E** Whole genome CNV profile derived SNP6 array for the sample classified as SFT, HMPC in Grp7. **F** Kaplan-Meier plot with overall survival time (months) for the five enriched subgroups in **A** (Grp1, 2, 3, 7, and 8). The risk table is provided below. **G** Log-rank test was used to compare every two subgroups. P-values were provided in the table.

Chapter 3

3 Gene expression subtype prediction

As I described in the first chapter, gene expression subtypes were identified in GBM which classifies GBM into three subgroups: CL, MES, and PN. Samples belong to each subgroup have their distinguished genetic characterizations, such as high expressed gene sets, amplified or deleted genes in terms of copy number, and mutant genes. Because there are three subtypes, it is a multi-class problem to build a predictive model for gene expression subtypes.

3.1 Methods

3.1.1 DNA methylation data

To increase the number of samples available for model building, I included all GBM in TCGA with HM27k and HM450k data available. TCGA level 3 data were directly used for samples with HM27k data. Samples with HM450k data were processed with UniD pipeline and further normalized with BMIQ (66). HM450k probes belong to the following categories were deleted: those missing from the EPIC platform, those that were multi-hit (76) or snp-hit (76), those located on chromosome X or Y, and those with ≥5% missing values in the data set. Then the retained HM450k probes were intersected with HM27k probes. Only probes existing in both platforms were kept for the following analysis. The retained probes belonging to following categories were deleted: (1) those with missing values in ≥5 samples, (2) those not located on CpG island, and (3) those not mapped to a known gene. Samples without gene expression data (Agilent 244K) were also excluded. For each probe, the Spearman correlation coefficient value was calculated between the methylation level and corresponding gene expression level. Probes with an absolute correlation coefficient value ≥ 0.1 were included (**Figure 37**).

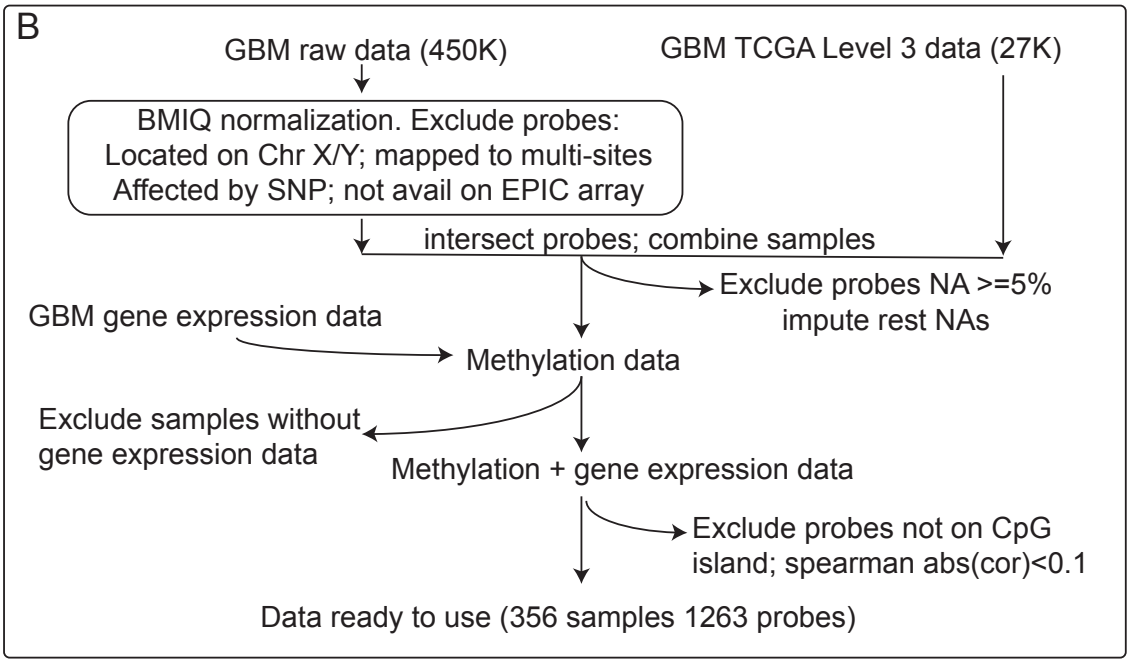


Figure 37: Data processing of DNA methylation data for gene expression subtype prediction

3.1.2 Gene expression subtypes annotation

The gene expression subtype and empirical p-value calculated by permutations were obtained for each sample according to the published algorithm (9). A probability of each sample belonging to each subtype was calculated as shown below (the CL subtype was used as an example). This per sample per subtype probability was used to calculate the sum of different probabilities, which is one of the metrics to evaluate model performance.

$$Prob_C = \frac{1 - P.value_C}{(1 - P.value_C) + (1 - P.value_M) + (1 - P.value_P)}$$

$$= \frac{1 - P.value_C}{3 - (P.value_C + P.value_M + P.value_P)}$$

3.1.3 Model building

For gene expression subtype predictions, subtype and probability were calculated using gene expression data and were used as a reference for model building (9). Samples were randomly stratified into training (60%), development (20%), and test (20%) sets with gene expression subtype as the stratification factor. The Fselector (R package) built within the mlr

package(90) in R was applied to the training set and three different evaluation metrics were calculated: IG, GR, and SU. Probes were ranked by each evaluation metric from high to low. Then the rank sum was added up for each probe. Probes were then sorted by the rank sum from high to low. The top probes were those that showed up as the most important in terms of their response variable.

With the selected probes, different probe sets were selected based on different quantile values (top 100%, 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, and 10%) according to importance rankings. In total, 21 machine learning algorithms (**Table 37**) were fitted and evaluated in the training set. For each tested algorithm, a training set with different probe sets was applied 100 times using 5-fold CV with seeds from 1 to 100. The prediction accuracy and sum of different probabilities were calculated for each fold of CV. The prediction accuracy was calculated by comparing the predicted gene expression subtype (predicted subtype) and the assigned subtypes were obtained from gene expression data (the "real" subtype). The sum of different probabilities was calculated by the sum of the square of the different probability of each subtype (C: classical, M: mesenchymal, P: proneural), as shown below. For each algorithm and each selected probe set, the prediction accuracy and sum of probability difference from 100 times of 5-fold CV were summarized.

$$diff.prob_C = (real.probability_C - predicted.probability_C)^2$$

$$sum\ of\ different\ probabilities = diff.prob_C + diff.prob_M + diff.prob_P$$

Top candidates with the best prediction accuracy and sum of different probabilities in the training set were applied to the development set. The final algorithm was selected based on its prediction accuracy in the development set (**Figure 38**). With the final algorithm determined, all samples from the training and development sets were used to build the final model. Then the final model was applied to the test set to evaluate its performance.

Table 37: Twenty-one machine learning algorithms and R package application

Algorithms/function	R package
boosting	adabag
C50	C50
cforest	party
ctree	party
cvglmnet	glmnet
earth	earth
evtree	evtree
gbm	gbm
glmnet	glmnet
Ibk	Rweka
J48	Rweka
Jrip	Rweka
Kknn	kknn
ksvm	kernlab
lda	MASS
naiveBayes	e1071
OneR	Rweka
PART	Rweka
randomForest	randomForest
randomForestSRC	randomForestSRC
ranger	ranger

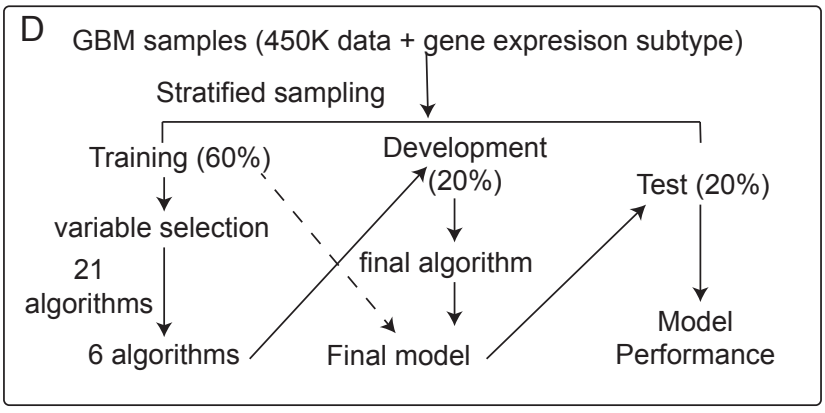


Figure 38: Model building process for gene expression subtypes prediction

3.1.4 Signature analysis

For signature analysis, the same genomic context comparison in the binary genomic alterations were applied for gene expression signature, except for the CpG island relationship comparison because all probes were already filtered by CpG island.

3.1.5 Prediction results analysis

For gene expression subtype prediction, misclassified samples in the test set (n = 72) were regrouped by their transcriptional subtypes and methyl-based predicted subtypes. The correctly classified and misclassified samples were compared in terms of CNV and gene expression for each transcriptional subtype using the Wilcoxon rank sum test.

3.1.6 Model validation

To validate the gene expression subtype prediction, TCGA LGG gene expression subtypes were compared between those determined by DNA methyl-based model and those obtained directly using gene expression profile (transc-based) (9). Histopathological and genomic characteristics were compared between methyl-based and transc-based subtype determinations using χ-square or Fisher's exact tests.

3.2 Results

3.2.1 Model building

GBM samples with HM450k data were processed as described above, which left 129 samples with 407,067 probes. Samples from the HM27k platform included 287 samples with 23,578 probes. After data integration, there were 416 samples with 20,720 probes available. After probe filters, 9,519 probes were kept for correlation evaluation. I only kept the samples with gene expression information available, which led to 1,263 probes and 356 samples making up the final data set. In the training set, 212 samples were used for probe importance evaluation, and 985 probes were assigned zero importance and then were excluded from the analysis. According to the sum of rankings of three important metrics, the top 100%, 90%, 80%, 70%, 60%,

50%, 40%, 30%, 20%, and 10% of the 278 probes were used for the 21 machine learning algorithms (**Figure 39**). The misclassification rate of each machine learning algorithm and each probe set is summarized in **Figure 40**. The sum of different of probabilities for each machine learning algorithm and each probe set is summarized in **Figure 41 and 42**. In terms of the probe sets, it is clear that the prediction accuracy did not affect by the number of probes involved and the average sum of different probabilities decreased as the number of probes increased. To reach to the best-predicted probability, I used all 278 probes.

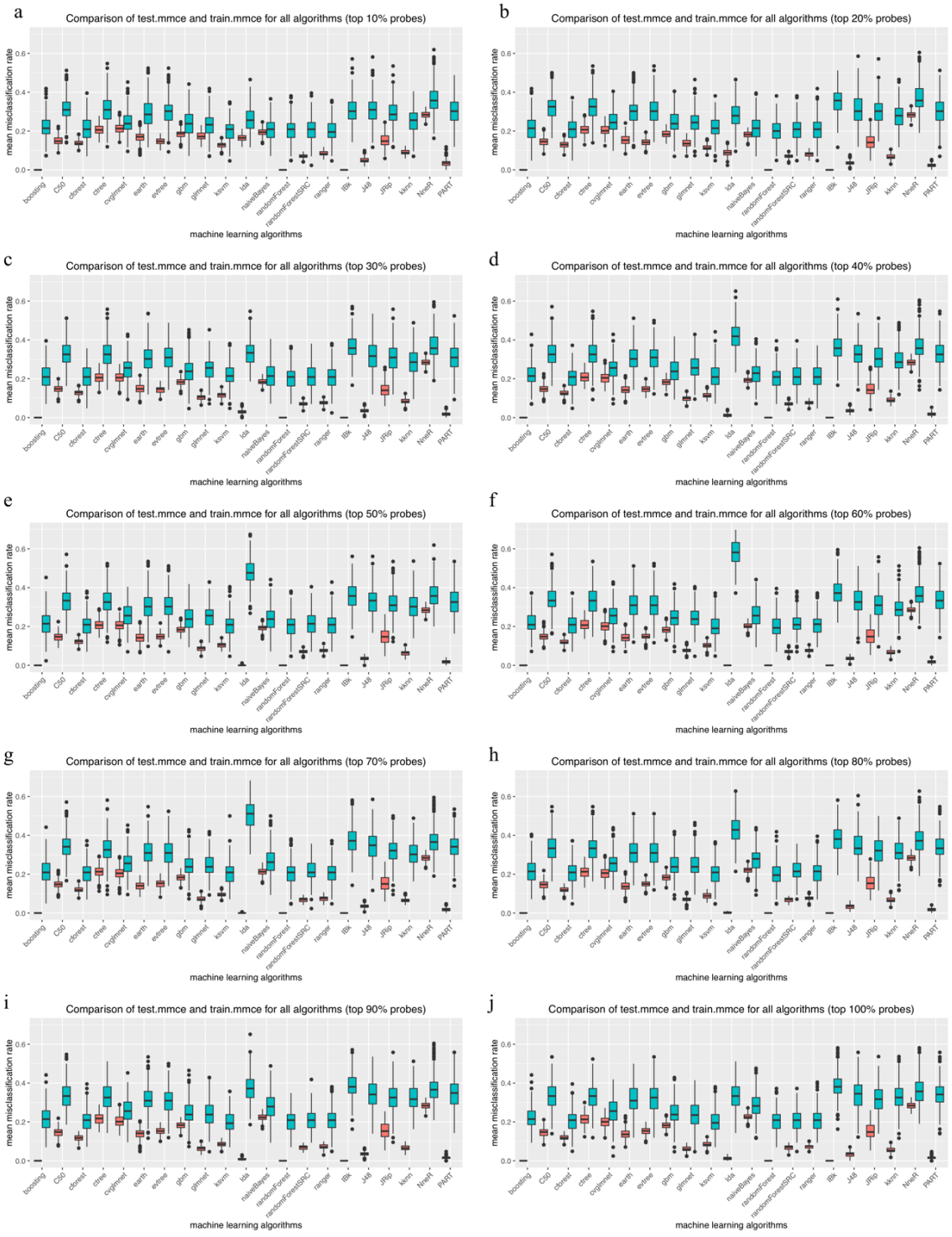


Figure 39: Comparison of misclassified rate of TCGA gene expression subtype of 5-fold corss validation tests using top quantile probes

For each machine learning algorithms show in the x-axis, two boxplots are shown: the red boxplot shows the misclassification rate when building the model with ranodm four-folds of the traning data set and the green boxplot show the misclassification rate when applying the model to the left one-fold of the traning data set. The y-axis is the averaged misclassification rate among 5-

fold CV. Figures **a** to **j** represent the top 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% probes, respectively.

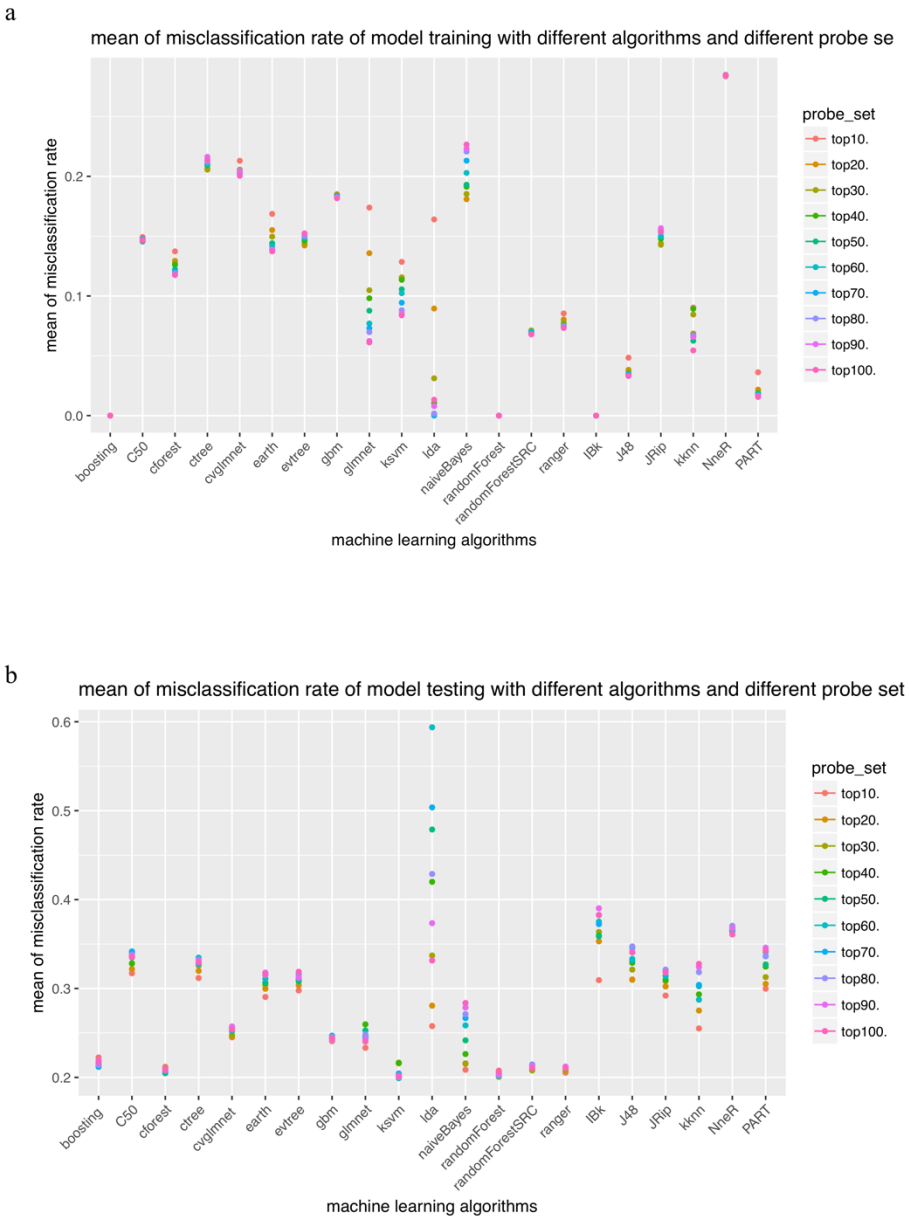


Figure 40: Summarization of the averaged misclassification rate for the 5-fold CV using top quantile probes.

The x-axis shows the twenty-one machine learning algorithms, the y-axis shows the averaged misclassification rate. The different colored dots represent the summarized the mean of misclassification rate calculated with different probe set. **a.** The averaged misclassification rate of the model building using random four folds data in the training set. **b.** The averaged misclassification rate when applying the model to predict the left one-fold data in the training set.

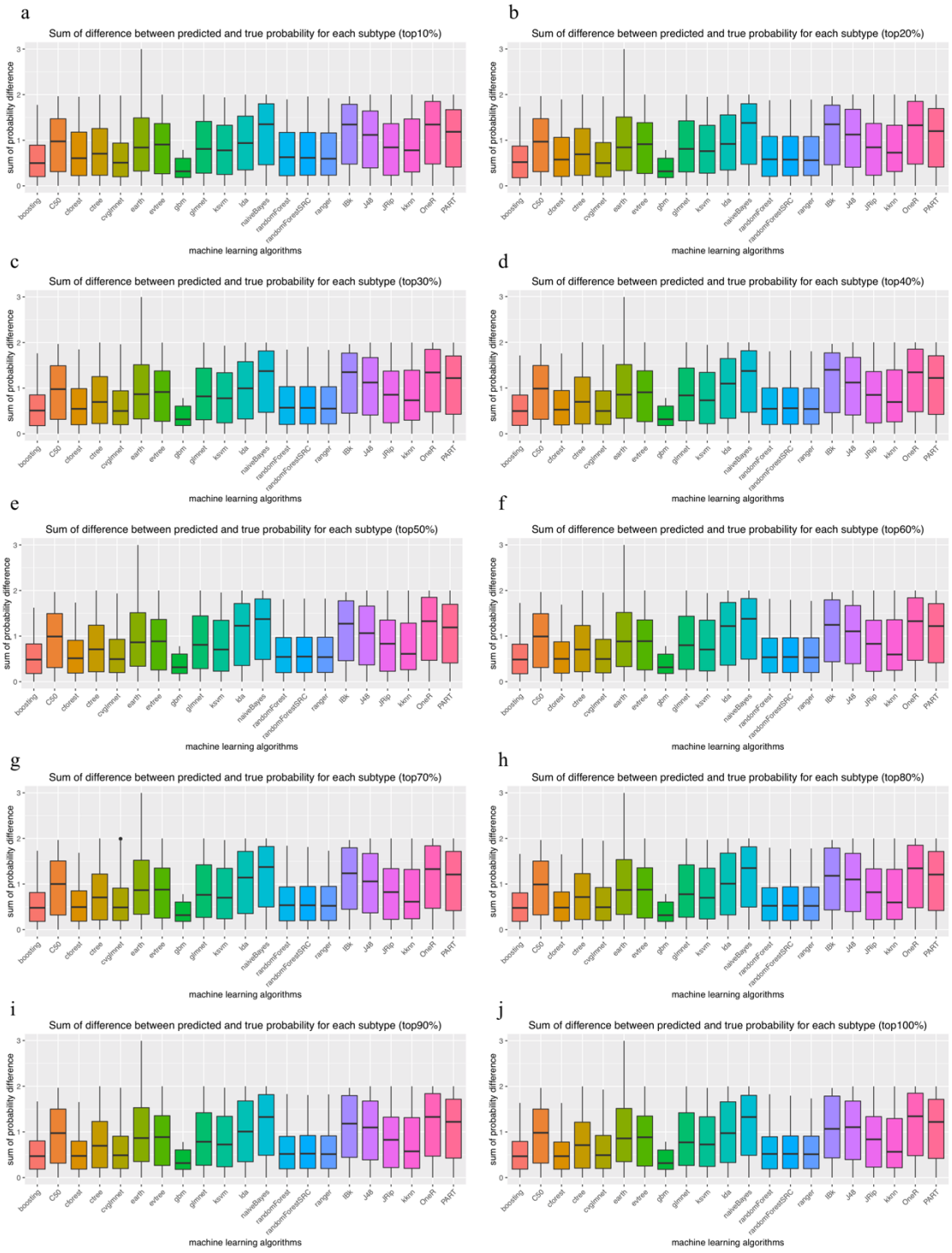


Figure 41: The summarized sum of different probability among twenty-one machine learning algorithms using different top quantile probe sets.

The x-axis show the evaluated algorithms and the y-axis show the sum of different probability. Each boxplot represent the summary of the sum of different probability in the test-fold among the 5-fold CV among hundreds repeated calculations. Figures **a** to **j** show the results of top 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% quantile probe sets, respectively.

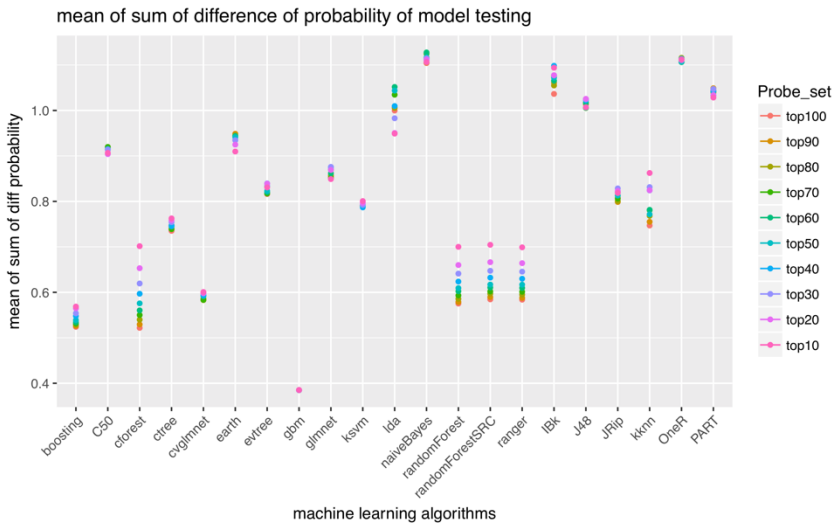


Figure 42: Summarized averaged sum of different probability among 5-fold CV using different top quantile probe sets

The x-axis represents the twenty-one machine learning algorithms and the y-axis represents the averaged sum of different probability. Each colored dot represents the averaged sum of different probability using different probe sets.

The gene expression subtype prediction model utilized 278 probes after probe selection. Based on the prediction accuracy and sum of different probabilities, six candidate algorithms were selected among the 21 machine learning algorithms. Random forest outperformed the other five candidate algorithms and was selected as the final algorithm (**Table 38**). The final random forest model was refitted with training (n = 212) and development (n = 72) sets and achieved a prediction accuracy of 72.2% (52/72) in the test set. For probes used in predictive models refer to the UniD package.

Table 38: Six candidate algorithms performance in development set for gene expression subtype prediction.

Algorithm	Predicted accuracy	averaged sum of different probabilities per sample
Boosting	68.06%	0.203081
cvglmnet	70.83%	0.198657
cforest	70.83%	0.203108
randomForest	73.61%	0.195134
randomForestSRC	73.61%	0.197521
ranger	73.61%	0.195947

3.2.2 Predictive signature analysis

The gene expression subtype prediction signature was enriched in chromosome 18 (**Table 39**) and most of the probes were mapped to the 1st Exon (39.95%) and 5'-UTR regions (25.8%) (**Table 40**). Four probes were mapped to the gene suppressor of cytokine signaling 2 (*SOCS2*) and three probes were mapped to *ERBB2* and retinol binding protein 1 (*RBP1*). The two most significant GOs were correlated with neuron development and differentiation (**Table 41**).

Table 39: Chromosome enrichment analysis for gene expression subtypes

Genomic alterations	CHR	# probes available	# probes	normalized percent	percent % (sum to 1)	p-value (proportion Test)
Gene expression subtype	1	36923	26	0.070%	3.652%	9.486E-10
	2	27711	17	0.061%	3.182%	
	3	20279	26	0.128%	6.650%	
	4	16241	5	0.031%	1.597%	
	5	19398	15	0.077%	4.011%	
	6	28392	13	0.046%	2.375%	
	7	23237	13	0.056%	2.902%	
	8	16401	10	0.061%	3.163%	
	9	7783	15	0.193%	9.997%	
	10	18784	7	0.037%	1.933%	
	11	23487	19	0.081%	4.196%	
	12	20018	7	0.035%	1.814%	
	13	9794	4	0.041%	2.118%	
	14	12234	15	0.123%	6.360%	
	15	12402	6	0.048%	2.509%	
	16	17917	14	0.078%	4.053%	
	17	23237	19	0.082%	4.241%	
	18	4939	12	0.243%	12.602%	
	19	21429	9	0.042%	2.178%	
	20	8823	16	0.181%	9.406%	
	21	3507	5	0.143%	7.395%	
	22	7074	5	0.071%	3.666%	

Table 40: Gene structure enrichment for gene expression subtypes signature

Genomic Alterations	relation to gene structure	# probes available	#probes	normalized #probe	Percentage (sum to 1)	p-value (proportion test)
gene expression subtype	TSS200	41774	38	0.091%	11.990%	< 2.2e-16
	TSS1500	55088	54	0.098%	12.920%	
	Body	126827	62	0.049%	6.443%	
	3'UTR	13641	3	0.022%	2.899%	
	5'UTR	33719	66	0.196%	25.799%	
	1st Exon	18147	55	0.303%	39.948%	
	not categorized	90814	0	0.000%	0.000%	

Table 41: GO enrichment analysis results for gene expression subtype signature

Category	Term	Count	P-Value	Benjamini-adjust p-value
GOTERM_BP_FAT	neuron differentiation	24	4.300E-08	6.80E-05
GOTERM_BP_FAT	neuron development	17	2.100E-05	1.60E-02
GOTERM_BP_FAT	behavioral fear response	5	4.400E-05	2.30E-02
GOTERM_BP_FAT	behavioral defense response	5	4.400E-05	2.30E-02
GOTERM_BP_FAT	fear response	5	1.200E-04	4.70E-02

3.2.3 Predictive results analysis

For the gene expression subtype predictive model, samples in the test set (n = 72) were categorized according to methyl-based and transc-based gene expressions subtype (**Figure 43A**). Samples discordant between the two methods showed significant differences in copy number variation and gene expression level compared to samples with concordant subtypes. I examined alterations reported as enriched in specific subtypes in discordant samples to determine which classification approach showed the highest association with these characteristic alterations (using the Wilcoxon rank sum test) (**Figure 43BC**). This further supports methyl-based classification among discordant cases classified as CL by transcription and MES by methylation or classified as MES by transcription and CL by methylation.

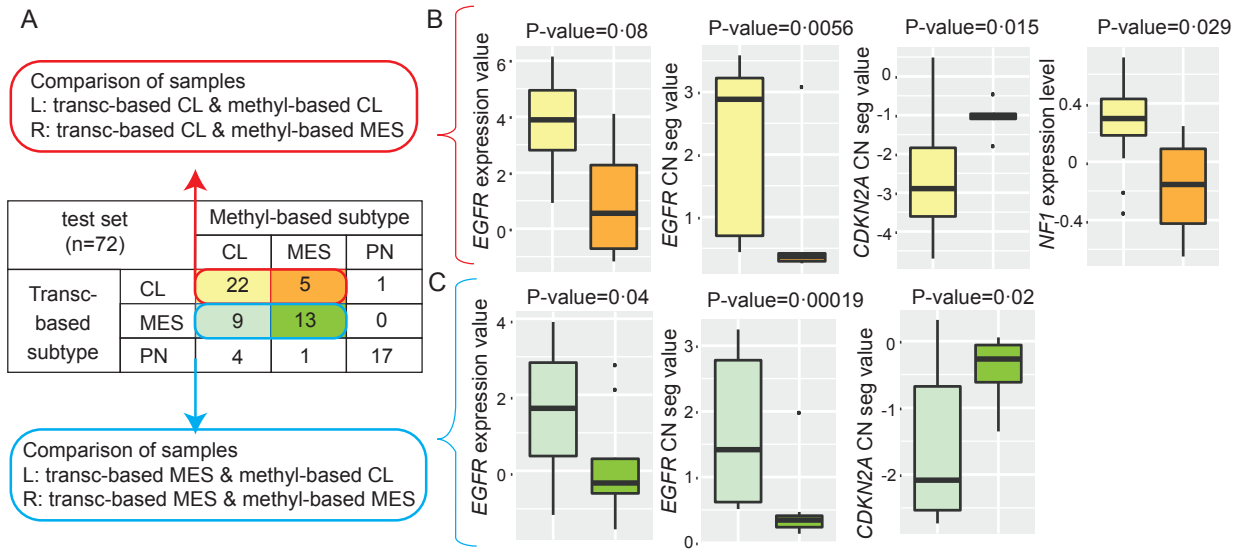


Figure 43: Methylation-based gene expression subtype predictions analysis in test set

A Confusion matrix of test set (n = 72) based on gene expression subtype of transc-based prediction and methyl-based prediction. (L: left; R: Right). **B** Transc-based CL and methyl-based CL samples were compared with transc-based CL and methyl-based MES samples for gene expression level (*EGFR*, *NF1*) and copy number (CN) segmentation level (*EGFR*, *CDKN2A*). For all transc-based CL samples, methyl-based MES sample shows lower *EGFR* expression (the estimated difference is 2.56 and 95% CI is -0.17-5.2, p-value = 0.08), lower EGFR amplification (the estimated difference is 1.28 and 95% CI is 0.14-2.86, p-value = 0.0056), higher CDKN2A/CDKN2B amplification (the estimated difference is -1.83 and 95% CI is -2.64 to -0.55 p-value = 0.015), and lower NF1 expression level (the estimated difference is 0.44 and 95% CI is 0.04 to 0.92, p-value = 0.029) than methyl-based CL samples. **C** Transc-based MES and methyl-based CL samples were compared with transc-based and methyl-based MES samples for gene expression level (*EGFR*) and CN segmentation level (*EGFR*, *CDKN2A*). For transc-based MES samples, methyl-based CL subtype show higher *EGFR* expression (the estimated difference is 1.42 and 95% CI is 0.04 to 3.25, p-value = 0.04), higher *EGFR* amplification (the estimated difference is 0.68 and 95% CI is 0.27 to 2.44, p-value = 1.9×10^{-4}), and lower *CDKN2A*/*CDKN2B* amplification (the estimated difference is -1.45 and 95% CI is -2.23 to -0.18, p-value = 0.02) compared to methyl-based MES samples.

3.2.4 Predictive model validation

The gene expression subtypes predicted by methyl-based and transc-based algorithms demonstrated large differences from the classification results of the PN subtype in TCGA LGG samples (**Table 42**); 422/486 (86.8%) samples were classified as PN in the methyl-based subtype and only 228/486 (46.9%) were classified in the transc-based subtype. Histology, chr1p19q codel, *MGMT* methylation status, *IDH1*, *TP53*, *ATRX*, and *PTEN* mutation all showed statistically significant associations with both methyl-based subtypes and transc-based subtypes

(**Figure 44, Table 43**). However, a significant association of the *NF1* mutation was observed only in methyl-based subtype.

Table 42: gene expression subtype validation using TCGA-LGG samples

validation data set TCGA-LGG		methyl-based subtype		
		CL	MES	PN
transc-based subtype	CL	21	19	109
	MES	0	24	85
	PN	0	0	228

Figure 44: TCGA LGG samples' gene expression subtype prediction and other genomic alteration profiles

Heatmap of TCGA-LGG samples (n = 486) with gene expression subtypes, histology, chr1p19q codel, *MGMT* promoter methylation, somatic mutations, and CNV.

Table 43: Chi-square test between genomic alteration and methyl-based and transc-based gene expression for TCGA-LGG samples

	gene expression subtype source	subgroups	CL	MSE	PN	p.value
histology*	methyl-based subtype	astrocytoma	10 (58.82%)	28 (75.68%)	111 (31.90%)	1.09E-06
		oligoastrocytoma	5 (29.41%)	3 (8.1%)	91 (26.15%)	
		oligodendroglioma	2 (11.76%)	6 (16.22%)	146 (41.95%)	
		sum	17	37	348	
	transc-based subtype	astrocytoma	52 (41.94%)	51 (54.83%)	46 (24.86%)	6.63E-11
		oligoastrocytoma	39 (31.45%)	26 (27.96%)	34 (18.38%)	
		oligodendroglioma	33 (26.61%)	16 (17.20%)	105 (56.76%)	
		sum	124	93	185	
chr1p19q codel^	methyl-based subtype	codel	0 (0%)	1 (2.5%)	144 (38.4%)	1.85E-09
		non-codel	19 (100%)	39 (97.5%)	231 (61.6%)	
		sum	19	40	375	
	transc-based subtype	codel	23 (17.04%)	11 (11.11%)	111 (55.50%)	3.35E-19
		non-codel	112 (82.96%)	88 (88.89%)	89 (44.50%)	
		sum	135	99	200	
MGMT^	methyl-based subtype	methylated	10 (47.62%)	17 (39.53%)	373 (88.39%)	6.51E-15
		unmethylated	11 (52.38%)	26 (60.47%)	49 (11.61%)	
		sum	21	43	422	
	transc-based subtype	methylated	108 (72.48%)	79 (72.48%)	213 (93.42%)	2.52E-09
		unmethylated	41 (27.52%)	30 (27.52%)	15 (6.58%)	
		sum	149	99	228	
IDH1 mut^	methyl-based subtype	mutant	0 (0%)	2 (5%)	336 (89.36%)	3.04E-42
		WT	19 (100%)	38 (95%)	40 (10.64%)	
		sum	19	40	376	
	transc-based subtype	mutant	88 (64.71%)	69 (70.41%)	181 (90.05%)	1.60E-08
		WT	48 (35.29%)	29 (29.59%)	20 (9.95%)	
		sum	136	98	201	
IDH2 mut^	methyl-based subtype	mutant	0 (0%)	0 (0%)	19 (5.05%)	0.3305438
		WT	19 (100%)	40 (100%)	357 (94.95%)	
		sum	19	40	376	
	transc-based subtype	mutant	2 (1.47%)	4 (4.08%)	13 (6.47%)	0.07515979
		WT	134 (98.53%)	94 (95.92%)	188 (93.53%)	
		sum	136	98	201	
mut.NF1^	methyl-based subtype	mutant	1 (5.26%)	12 (30%)	11 (2.93%)	2.16E-07
		WT	18 (94.74%)	28 (70%)	365 (97.07%)	
		sum	19	40	376	
	transc-based subtype	mutant	10 (7.35%)	8 (8.16%)	6 (2.99%)	0.07672947
		WT	126 (92.65%)	90 (91.84%)	195 (97.01%)	
		sum	136	98	201	

	gene expression subtype source	subgroups	CL	MSE	PN	p.value
mut.TP53^	methyl-based subtype	mutant	0 (0%)	5 (12.5%)	207 (55.05%)	2.39E-12
		WT	19 (100%)	35 (87.5%)	169 (44.95%)	
		sum	19	40	376	
	transc-based subtype	mutant	70 (51.47%)	64 (65.31%)	78 (38.81%)	6.62E-05
		WT	66 (48.53%)	34 (34.69%)	123 (61.19%)	
		sum	136	98	201	
mut.ATRX^	methyl-based subtype	mutant	0 (0%)	2 (5%)	170 (45.21%)	5.55E-11
		WT	19 (100%)	38 (95%)	206 (54.79%)	
		sum	19	40	376	
	transc-based subtype	mutant	54 (39.71%)	53 (54.08%)	65 (32.34%)	0.001618116
		WT	82 (60.29%)	45 (45.92%)	136 (67.66%)	
		sum	136	98	201	
mut.CIC^	methyl-based subtype	mutant	0 (0%)	0 (0%)	94 (25%)	3.04E-06
		WT	19 (100%)	40 (100%)	282 (75%)	
		sum	19	40	376	
	transc-based subtype	mutant	8 (5.88%)	5 (5.01%)	81 (40.30%)	1.71E-18
		WT	128 (94.12%)	93 (94.90%)	120 (59.70%)	
		sum	136	98	201	
mut.PIK3CA^	methyl-based subtype	mutant	2 (10.53%)	6 (15%)	28 (7.45%)	0.1918081
		WT	17 (89.47%)	34 (85%)	348 (92.55%)	
		sum	0	0	376	
	transc-based subtype	mutant	12 (8.82%)	10 (10.20%)	14 (6.97%)	0.5739994
		WT	124 (91.18%)	88 (89.80%)	187 (93.03%)	
		sum	136	98	201	
mut.PIK3R1^	methyl-based subtype	mutant	1 (5.26%)	0 (0%)	18 (4.79%)	0.3898719
		WT	18 (94.74%)	40 (100%)	358 (95.21%)	
		sum	19	40	376	
	transc-based subtype	mutant	5 (3.68%)	2 (2.04%)	12 (5.97%)	0.3025682
		WT	131 (96.32%)	96 (97.96%)	189 (94.03%)	
		sum	136	98	201	
mut.PTEN^	methyl-based subtype	mutant	5 (26.32%)	7 (17.50%)	6 (1.60%)	1.06E-07
		WT	14 (73.68%)	33 (82.05%)	370 (98.40%)	
		sum	19	40	376	
	transc-based subtype	mutant	11 (8.09%)	6 (6.12%)	1 (0.50%)	0.000303116
		WT	125 (91.91%)	92 (93.88%)	200 (99.50%)	
		sum	136	98	201	
mut.RB1^	methyl-based subtype	mutant	0 (0%)	3 (7.50%)	1 (0.27%)	4.83E-03
		WT	19 (100%)	37 (92.50%)	375 (99.73%)	
		sum	19	40	376	
	transc-based subtype	mutant	1 (0.74%)	2 (2.04%)	1 (0.50%)	0.3576105
		WT	135 (99.26%)	96 (97.96%)	200 (99.50%)	
		sum	136	98	201	

	gene expression subtype source	subgroups	CL	MSE	PN	p.value
mut.NOTCH1^	methyl-based subtype	mutant	0 (0%)	0 (0%)	35 (9.31%)	0.04250279
		WT	19 (100%)	40 (100%)	341 (90.69%)	
		sum	19	40	376	
	transc-based subtype	mutant	5 (3.68%)	5 (5.01%)	25 (12.44%)	0.007861747
		WT	131 (96.32%)	93 (94.90%)	176 (87.56%)	
		sum	136	98	201	
mut.FUBP1^	methyl-based subtype	mutant	0 (0%)	0 (0%)	36 (9.57%)	0.03725316
		WT	19 (100%)	40 (100%)	340 (90.43%)	
		sum	19	40	376	
	transc-based subtype	mutant	4 (2.94%)	2 (2.04%)	30 (14.93%)	1.40E-05
		WT	132 (97.06%)	96 (97.96%)	171 (85.07%)	
		sum	136	98	201 (24.78%)	
cnv.CDKN2A*	methyl-based subtype	Hemi-del	4 (19.05%)	8 (18.60%)	101 (24.05%)	3.66E-27
		Homo-del	14 (66.67%)	19 (44.19%)	20 (4.76%)	
		Low-amp	0 (0%)	2 (4.65%)	7 (1.67%)	
		Neutral	3 (14.29%)	14 (32.56%)	292 (69.52%)	
		sum	21	43	420	
	transc-based subtype	Hemi-del	32 (21.48%)	25 (22.94%)	56 (3.98%)	0.000626091
		Homo-del	28 (18.79%)	16 (14.68%)	9 (3.98%)	
		Low-amp	4 (2.68%)	1 (0.92%)	4 (1.77%)	
		Neutral	85 (57.05%)	67 (61.47%)	157 (69.47%)	
		sum	149	109	226	
cnv.EGFR*	methyl-based subtype	Hemi-del	0 (0%)	1 (2.33%)	3 (0.71%)	4.73E-43
		Homo-del	15 (71.43%)	12 (27.91%)	8 (1.90%)	
		Low-amp	6 (28.57%)	17 (39.53%)	57 (13.57%)	
		Neutral	0 (0%)	13 (30.23%)	352 (83.81%)	
		sum	21	43	420	
	transc-based subtype	Hemi-del	1 (0.67%)	1 (0.92%)	2 (0.88%)	3.25E-07
		Homo-del	26 (17.45%)	7 (6.42%)	2 (0.88%)	
		Low-amp	26 (17.45%)	22 (20.18%)	32 (14.16%)	
		Neutral	96 (64.43%)	79 (72.48%)	190 (84.07%)	
		sum	149	109	226	

	gene expression subtype source	subgroups	CL	MSE	PN	p.value
cnv.PDGFRA*	methyl-based subtype	Hemi-del	0 (0%)	2 (4.65%)	55 (13.10%)	0.1453379
		Homo-del	0 (0%)	3 (6.98%)	14 (3.33%)	
		Low-amp	1 (4.76%)	2 (4.65%)	8 (1.90%)	
		Neutral	20 (95.24%)	36 (83.72%)	343 (81.67%)	
		sum	21	43	420	
	transc-based subtype	Hemi-del	11 (7.38%)	6 (5.50%)	40 (17.70%)	0.01534663
		Homo-del	5 (3.36%)	4 (3.67%)	8 (3.54%)	
		Low-amp	2 (1.34%)	3 (2.75%)	6 (2.65%)	
		Neutral	131 (87.92%)	96 (88.07%)	172 (76.11%)	
		sum	149	109	226	
cnv.CCND2*	methyl-based subtype	Hemi-del	1 (4.76%)	1 (2.33%)	17 (4.05%)	0.5311001
		Homo-del	1 (4.76%)	0 (0%)	22 (5.24%)	
		Low-amp	0 (0%)	5 (11.63%)	39 (9.29%)	
		Neutral	19 (90.48%)	37 (86.05%)	342 (81.43%)	
		sum	21	43	420	
	transc-based subtype	Hemi-del	8 (5.37%)	2 (1.83%)	9 (3.98%)	0.1067253
		Homo-del	5 (3.36%)	7 (6.42%)	11 (4.87%)	
		Low-amp	11 (7.38%)	17 (15.60%)	16 (7.08%)	
		Neutral	125(83.89%)	83 (76.15%)	190 (84.07%)	
		sum	149	109	226	
cnv.FGFR2*	methyl-based subtype	Hemi-del	21 (100%)	32 (74.42%)	74 (17.62%)	5.92E-25
		Homo-del	0 (0%)	0 (0%)	1 (0.24%)	
		Low-amp	0 (0%)	0 (0%)	9 (2.14%)	
		Neutral	0 (0%)	11 (25.58%)	336 (80.00%)	
		sum	21	43	420	
	transc-based subtype	Hemi-del	54 (36.24%)	37 (33.94%)	36 (15.93%)	0.00039857
		Homo-del	0 (0%)	0 (0%)	1 (0.44%)	
		Low-amp	3 (2.01%)	2 (1.83%)	4 (1.77%)	
		Neutral	92 (61.74%)	70 (64.22%)	185 (81.86%)	
		sum	149	109	226	
cnv.CDK4*	methyl-based subtype	Hemi-del	1 (4.76%)	1 (2.33%)	52 (12.38%)	6.37E-05
		high-amp	4 (19.05%)	4 (9.30%)	10 (2.38%)	
		Homo-del	0 (0%)	0 (0%)	1 (0.24%)	
		Low-amp	1 (4.76%)	5 (11.63%)	11 (2.62%)	
		Neutral	15 (71.43%)	33 (76.74%)	346 (82.38%)	
		sum	21	43	420	
	transc-based subtype	Hemi-del	19 (12.75%)	16 (14.68%)	19 (8.41%)	0.1116996
		high-amp	7 (4.70%)	5 (4.59%)	6 (2.65%)	
		Homo-del	0 (0%)	0 (0%)	1 (0.44%)	
		Low-amp	4 (2.68%)	8 (7.34%)	5 (2.21%)	
		Neutral	119 (79.87%)	80 (73.39%)	195 (86.28%)	
		sum	149	109	226	

	gene expression subtype source	subgroups	CL	MSE	PN	p.value
cnv.MDM2*	methyl-based subtype	Hemi-del	1 (4.76%)	1 (2.33%)	60 (14.29%)	1.07E-06
		high-amp	2 (9.52%)	1 (2.33%)	1 (0.24%)	
		Homo-del	0 (0%)	0 (0%)	1 (0.24%)	
		Low-amp	1 (4.76%)	5 (11.63%)	7 (1.67%)	
		Neutral	17 (80.95%)	36 (83.72%)	351 (83.57%)	
		sum	21	43	420	
	transc-based subtype	Hemi-del	19 (12.75%)	21 (19.27%)	22 (9.73%)	0.03638587
		high-amp	3 (2.01%)	0 (0%)	1 (0.44%)	
		Homo-del	0 (0%)	0 (0%)	1 (0.44%)	
		Low-amp	4 (2.68%)	6 (5.5%)	3 (1.33%)	
		Neutral	123 (82.55%)	82 (75.23%)	199 (88.05%)	
		sum	149	109	226	
cnv.PTBP1*	methyl-based subtype	hemi-del	0 (0%)	2 (4.65%)	20 (4.76%)	3.29E-08
		high-amp	0 (0%)	0 (0%)	20 (4.76%)	
		low-amp	16 (76.19%)	7 (16.28%)	74 (17.62%)	
		neutral	5 (23.81%)	34 (79.07%)	306 (72.86%)	
		sum	21	43	420	
	transc-based subtype	hemi-del	4 (2.68%)	9 (8.26%)	9 (3.98%)	0.03295496
		high-amp	6 (4.03%)	3 (2.75%)	11 (4.87%)	
		low-amp	33 (22.15%)	11 (10.09%)	53 (23.45%)	
		neutral	106 (71.14%)	86 (78.90%)	153 (67.70%)	
		sum	149	109	226	
cnv.MDM4*	methyl-based subtype	hemi-del	0 (0%)	2 (4.65%)	15 (3.57%)	1.96E-11
		high-amp	5 (23.81%)	4 (9.30%)	5 (1.19%)	
		low-amp	4 (19.05%)	7 (16.28%)	19 (4.52%)	
		neutral	12 (57.14%)	30 (69.77%)	381 (90.71%)	
		sum	21	43	420	
	transc-based subtype	hemi-del	2 (1.34%)	3 (2.75%)	12 (5.31%)	0.07422154
		high-amp	8 (5.37%)	4 (3.67%)	2 (0.88%)	
		low-amp	11 (7.38%)	7 (6.42%)	12 (5.31%)	
		neutral	128 (85.91%)	95 (87.16%)	200 (88.50%)	
		sum	149	109	226	
cnv.SOX2*	methyl-based subtype	hemi-del	0 (0%)	2 (4.65%)	21 (5.00%)	0.001437712
		high-amp	0 (0%)	1 (2.33%)	3 (0.71%)	
		low-amp	4 (19.05%)	2 (4.65%)	9 (2.14%)	
		neutral	17 (80.95%)	38 (88.37%)	387 (92.14%)	
		sum	21	43	420	
	transc-based subtype	hemi-del	4 (2.68%)	7 (6.42%)	12 (5.31%)	0.35457
		high-amp	0 (0%)	2 (1.83%)	2 (0.88%)	
		low-amp	7 (4.70%)	3 (2.75%)	5 (2.21%)	
		neutral	138 (92.62%)	97 (88.99%)	207 (91.59%)	
		sum	149	109	226	

Note: m.gene: indicate the mutation status of the gene; cnv.gene: indicate the copy number variations of the gene; *: p-value was calculated with χ-square test; ^: p-value was calculated with fisher exact test

Chapter 4

4 Discussion

4.1 Discussion of Prediction results

The results demonstrated that DNA methylation microarray-based classifiers can accurately predict and deconvolute somatic genomic alterations in gliomas and show improved enrichment for characteristic genomic alterations.

4.1.1 Binary genomic alteration prediction

I found that the methyl-based *ATRX* prediction model identified cases with likely loss of function, even those samples were reported as wild type by sequencing. There exists a mutation type shift between set 2 (DNA-seq wild type, methyl-based mutant, and with mutation calls by reviewing SNV information) and set 4 (DNA-seq mutant, methyl-based wild type): the enriched mutations shifted from frameshift indels and in-frame indels to intron, missense and nonsense which may not lead to loss of function. The comparison between the mutation status of *ATRX*, *IDH*, and *TERT*p and *ATRX* expression level indicated that the methyl-based prediction model more accurately identifies the samples with true *ATRX* loss of function. It is reasonable to speculate that samples in set 3 (DNA-seq wild type, methyl-based mutant, and no mutation calls by reviewing SNV information) may be deactivated by some other mechanisms while *ATRX* mutations at the DNA sequencing level of set 4 (DNA-seq mutant, methyl-based wild type) do not change the function of the protein. Besides it, we can further explore the functional impact score (FIS) (91) for all mutations detected in *ATRX* mutation samples. This may help us to identify the real functional mutations among all detected mutations.

The low prediction accuracy of *TERT*p mutation in NOA-04 samples may due to the limited locus tested in target-sequencing.

When I compared the tumor deconvolution based on the mutation status of *IDH*, *ATRX*, and *TERT*p, I found that the methyl-based annotations provided more precise genomic

characterizations than DNA-seq annotations. For example, the same glioma sets are compared in **Figure 44 A** and **B**. Three samples in **Figure 44A** and one sample in **Figure 44B** show mutation status of both *ATRX* and *TERT*p which seems functionally redundant. Four samples with *IDH* wild type status were identified as *ATRX* mutant, which contradicts to current knowledge that *IDH* mutation is prerequisite of the *ATRX* mutation in **Figure 45A;** no such samples are shown for this condition in **Figure 45B**.

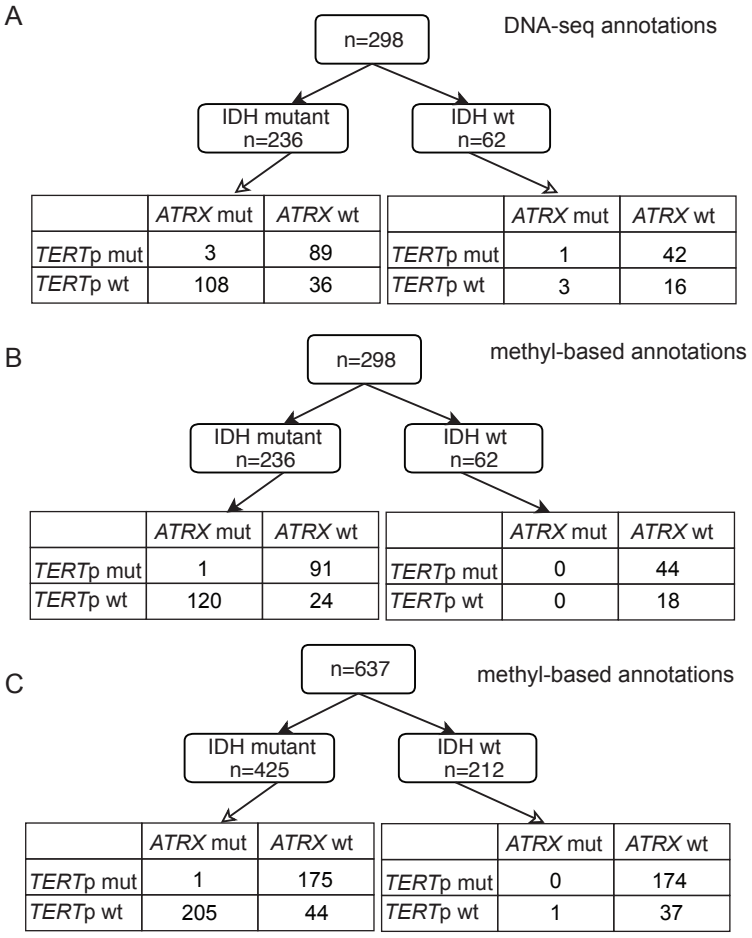


Figure 45: Gliomas deconvolution by mutation status of *IDH*, *ATRX*, and *TERT*p.

A: gliomas are stratified by DNA-seq or PCR-seq annotation of *IDH*, *ATRX*, and *TERT*p mutation status. **B**: the same set of gliomas in A are stratified by methyl-based model predicted *IDH*, *ATRX*, and *TERT*p mutation status. **C**: gliomas with HM450k data available are stratified by methyl-based model predicted *IDH*, *ATRX*, and *TERT*p mutation status.

CBS has been used to derive the CNV profile from HM450k methylation data and was applied with R package (84). However, this method needs to specifically check for the partial

codel in chr1p19q and a threshold needs to be chosen. This is time-consuming and susceptible to considerable inter- and intra-observer variability. The methyl-based predictive model may better or more rapidly provide an objective determination of CNV status.

CNS methylation-based classification (14) has shown great value in standardizing and clarifying brain tumor diagnoses by reducing the inter-observer variance and by classifying tumors previously unclassified by histology. However, this report focused on comprehensive CNS tumor classification based on WHO entities and did not focus on specific genomic alterations within tumor subtypes, including gliomas. My comparison of CNS methylation-based classification and genomic alterations show that it may not properly identify the specific genomic alteration subclasses studies in this report.

4.1.2 Gene expression subtype prediction

Due to tumor heterogeneity and the uneven distribution of MES subtype samples in the external validation data set (CL, 84; MES, 64; PN, 64), the achieved prediction accuracy of gene expression subtype by methylation was not as high as the other genomic alterations. However, by investigating the misclassified samples, I found that the methyl-based classification better aligned with established, characteristic subtype-associated genomic features compared to the transc-based classification. For example, the discordant transc-based CL and methyl-based MES samples (n = 5) did not show the characteristics of CL subtype samples, including *EGFR* amplification and high expression and *CDKN2A* deletion. Instead, they harbored characteristics of the MES subtype, such as *NF1* deletion.

Because there is no easy way to turn RNA-seq into a clinical assay, the gene expression subtype's clinical meaning has been underestimated. For example, PN gliomas are more sensitive to chemo and radiation treatment and NF-silenced gliomas are more sensitive to radiation than temozolomide treatment (92). MES gliomas usually are associate with worse survival rates and occur in older patients (9). It is worth noting that PN and MES transitions exist

during treatment and following recurrence. Overall, with UniD pipeline available, potential therapeutic targets in gene expression subtypes should be easily adopted in the future.

4.2 Research limitation and future direction

4.2.1 Research limitations

There are several limitations to my study. First, using DNAseq results as the *ATRX* mutation gold standard may not be the optimal way; IHC may better capture real *ATRX* deficiency. This may explain why samples in NOA-04 (with IHC as the gold standard) show better prediction accuracy (92.7%) than in TCGA development (85.16%) and in the test set (90.48%). Second, external validation data NOA-04 only include LGG and no GBM. Third, I only validated the UniD pipeline in HM450k data, though UniD pipeline can be applied to the EPIC array as well.

4.2.2 Future direction

With developed DNA methylation signatures for key genomic alterations in gliomas, it is intuitive to come up with the question: what is the relationship between those DNA methylation signatures and key genomic alterations. Whether the DNA methylation profile is caused by the somatic mutation, or the DNA methylation profile is the marker for specific cancer cell lineage. This topic has been discussed more in section 4.4. *ATRX* and *TERT*p mutations are the leading causations for telomere length maintenance as described in the introduction section. But there exist some cancer samples with none of these two mutations or with both of the mutations. Will this related to the epigenetic mechanism? Moreover, the DNA methylation signatures between *ATRX* and *TERT*p mutation show less overlapping, does this indicate the epigenetic independency for *ATRX* and *TERT*p?

One future direction of this research could be expanding predictive relationships to other genomic features or biomarkers. For example, I am currently building the DNA methylation signature for v-Src avian sarcoma viral oncogene homolog (SRC) kinase family activation in glioblastoma. SRC is the first discovered human proto-oncogene and is believed to play an

important role in maintaining neoplastic phenotype and tumor progression. Dasatinib is a multitargeted kinase inhibitor that can inhibit all SRC family kinase (SFK). However, Radiation Therapy Oncology Group (RTOG) 0627 clinical trial that evaluated the dasatinib in recurrent GBM failed. One potential reason is the SRC inhibitor was not specifically targeted for SFK activated patients. As previously discussed related to gene expression subtype, there is no easy way to identify the SRC pathway activation status in patients which leads to the non-precision medicine. Therefore, if the SRC pathway activation status could be identified using the DNA methylation signature, patients could be provided with the drug more specifically and expect better prognoses.

The DNA methylation microarray has many potential benefits with a wide range of application. For example, it can be used with FFPE samples. The sample size can easily be increased by applying the assay to past stored samples without the worry of a long sample collection time. Moreover, once the assay been run, all data can be stored and all future built models can be re-applied to the past sample data and further increase the number of sample data. Importantly, DNA methylation microarray can be easily standardized in different hospitals and labs and once the microarray assay has been verified, all models can be easily verified.

Another future direction is expanding this idea to other tumor types, not just to gliomas. In fact, the model building procedures and methods used in this study are not glioma specific. They can be easily used with other tumor types. For example, the hypermethylation phenotype CIMP, has also been found in colon cancer and been associated with BRAF mutation. I believe, considering all the advantages and promising findings I have discovered with gliomas, DNA methylation signature prediction can benefit research into other tumor types as well and will have not only research but also clinical implications.

4.3 Discussion of current research

DNA methylation's application in tumor research is rising. It has been used a lot as diagnostic and prognostic biomarkers for tumor subtyping and classification. For example, DNA

methylation has been used as the main subtyping feature for medulloblastoma (MB) and ependymoma. MB has been classified into four different entities according to their DNA methylation profiles: WNT, SHH, group3, and group4 (93), while each entity has its special transcriptome profiles, somatic mutations, copy number variations, and clinical outcomes. In addition, MB has been further divided into three or four subgroups within each of the major entities (94) when incorporating the gene expression profiles together with the DNA methylation profiles. With better identification tumor subgroups, we can better design the personalized treatment for each patient.

For this section, I would like to introduce two publications related to my project and discuss how my project is different from theirs.

4.3.1 H. Binder et al.

Paper information: DNA methylation, transcriptome and genetic copy number signatures of diffuse cerebral WHO grade II/III gliomas resolve cancer heterogeneity and development. Acta neuropathologica communications 7.1 (2019): 59.

This paper (95) focused on using the gene expression and DNA methylation profile to stratify LGG samples. They identified four consensus subtypes of LGG and further characterized them in terms of genomic alterations, microenvironment, treatment, and prognosis. The key difference between this paper and my research project is that this paper used unsupervised learning method while my project used supervised learning method for tumor classification. Unsupervised learning methods can provide us a better understanding of the relationship between tumor samples with the data provided. This means the learned relationships between samples can vary a lot by the methods and data. There is no specific method to test it. The most commonly used method to validate unsupervised learning results is to compare the subgroup samples for other tumor features. If the identified subgroups also show differences for other tumor features, we can have more confidence that those samples are actually belonging to different subgroups. Otherwise, the clustering results may not be very useful. In contrast, my study uses

supervised learning with known information. For example, we already know that samples can be stratified by key somatic mutations and those are used as reference in our supervised learning models.

4.3.2 Y Paul et al.

Paper information: Paul, Yashna, et al. "DNA methylation signatures for 2016 WHO classification subtypes of diffuse gliomas." Clinical epigenetics 9.1 (2017): 32.

This paper (96) used the Prediction Analysis of Microarray (PAM) method to identify DNA methylation signatures which can separate following samples: LGG *IDH* wild type versus LGG *IDH* mutant; LGG *IDH* mutant with chr1p19q codel versus LGG *IDH* mutant with chr1p19q intact; and GBM *IDH* wild type versus GBM *IDH* mutant samples. Signatures were validated with TCGA samples and other independent cohorts. This study aimed to use DNA methylation to mimic the 2016 WHO CNS classification for gliomas since both *IDH* mutation and chr1p19q codel are part of the diagnostic criteria.

Though *IDH* and chr1p19q codel were covered in this paper, my study expanded it to more genomic features including *ATRX* and *TERT*p mutation, and gene expression subtypes. The starting point is different: my study is not focused on re-capture the exiting gliomas classification criteria, but is driven by the hypothesis that DNA methylation signatures can reflect different cancer cell lineages. Moreover, this study was limited that you need to know whether your sample is LGG or GBM at first. In fact, recent studies have shown that there is no clear difference of tumor features and patient's characteristics between the LGG and GBM if with *IDH* mutation. In addition, this paper identified 14 CpG sites, 14 CpG sites, and 13 CpG sites to distinguish each sample comparison without providing the predictive models. Without fixed predictive models, signatures can hardly be applied for prediction and may suffer from strong subjective bias. Though model provided, it is possible that the model cannot be applied due to the small sets of CpG sites involved: those probes may not be available due to missing or quality issue.

4.4 Which came first, the chicken or the egg?

We already found the strong associations between DNA methylation profiles and somatic mutations. Those key somatic mutations can be used to identify subgroups and can potentially be used to delineate tumor progression. For example, *IDH* mutation is believed to be an early-stage event and all other alterations, such as *ATRX* or *TERT*p mutation, and chr1p19q codel, happened after it (15). Moreover, *IDH* mutation can lead to the genome-wide DNA methylation profile change (19). Then, the question is, what are the relationships between DNA methylation signatures and other somatic alterations? Does DNA methylation profile identify a specific cancer lineage and this cancer lineage eventually acquire that mutation? Or the specific DNA methylation pattern is caused by specific somatic alteration?

To understand the relationships between DNA methylation profiles and somatic mutation/genomic alterations, we need to understand how the DNA methylation are maintained in normal cells and how the DNA methylation has been reprogrammed in cancer cells.

4.4.1 DNA methylation maintenance

DNA methyltransferase (DNMT) family are enzymes which can add the methyl group to DNA. There are major three types of DNMT in human: DNMT1, DNMT3a and DNMT3b. All DNMTs use the S-adenosyl methionine (SAM) as the donor for methyl group. DNMT1 is the most abundant methyltransferase. It helps to maintain the global DNA methylation through direct interaction with proliferating cell nuclear antigen (PCNA) and ubiquitin-like, containing PHD and RING figure domains 1 (UHRF1). DNMT3a and DNMT3b is more involved in the de novo DNA methylation mechanism during the cell development. Many other factors are also essential to DNA methylation maintenance except the DNMT family, for example, the lysine-specific demethylase 1 (LSD1) which can demethylate both H3K4me and H3K9me of histone 3.

In addition to add methyl group to DNA, it is also important that the existing methyl group can be removed from the DNA through passive and active mechanisms. For passive mechanism, methyl group lost due to the absence of maintenance mechanism along with the replications. For

active mechanism, TET family enzymes get involved to remove or modify the methyl group through oxidizing 5-mC to 5-hmC, and further to 5-fC and 5-caC.

4.4.2 DNA methylation in cancer cell

The epigenetic mechanism is central to cancer initiation, development, and progression. It is common to observe hypomethylation of the cancer genome and oncogenes and hypermethylation of tumor repressor genes. The interplay between these two types of genes, oncogene and tumor suppressor genes, are essential during cancer initiation. The loss function of tumor suppressor genes can be caused in DNA level: gene mutation or loss of heterogeneity; epigenetic level: the change of DNA methylation level in the promoter CpG island regions; and histone posttranslational modification. Many genes were found to be directly or indirectly associated with the DNA methylation and histone modification enzyme (97) (**Table 44 and Table 45**).

Table 44: DNA methylation modifiers in cancer

Gene	Function	Tumor Type	Alteration
DNMT1	DNA methyltransferase	Colorectal, non-small cell lung, pancreatic, gastric, breast cancer	Mutation,overexpression
DNMT3A	DNA methyltransferase	MDS, AML	Overexpression
DNMT3B	DNA methyltransferase	ICF syndrome, SNPs in breast and lung adenoma	Mutation
MBD1/2	Methyl binding protein	Lung and breast cancer	Mutation
TET1	5’ methylcytosine hydroxylase	AML	Chromosome translocation
TET2	5’ methylcytosine hydroxylase	MDS, AML, glioma	Mutation, silencing
IDH1/2	Isocitrate dehydrogenase	Glioma, AML	Mutation
AID	5’ cytidine deaminase	CML	Aberration expression

MBD1/2: methyl-cpg binding domain protein 1 or 2

TET1/2: TET methylcytosine dioxygenase 1 or 2

AID: activation-induced cytidine deaminase

MDS: myelodysplastic syndromes

ICF syndrome: immunodeficiency-centromeric instability-facial anomalies syndrome

CML: chronic myelogenous leukemia

Note: This table is adapted from You, Jueng Soo, and Peter A. Jones. "Cancer genetics and epigenetics: two sides of the same coin?." *Cancer cell* 22.1 (2012): 9-20 with license ID 1005420-1.

Table 45: Histone modification in cancer

Gene	Function	Tumor Type	Alteration
MLL1/2/3	Histone methyltransferase H3K4	Bladder TCC, ALL, AML, non-Hodgkin lymphoma, B cell lymphoma	Translocation, mutation, aberrant expression
BRD4	Bromodomain containing 4	Nuclear protein in testis, midline carcinoma, breast, colon, and AML	Translocation, aberrant expression
EZH2	Histone methyltransferae H3K27	Breast, prostate, bladder, colon, pancreas, liver, gastric, uterine tumors, melanoma, lymphoma, myeloma, and Ewing's sarcoma	Translocation, aberrant expression
ASXL	Enhancer of trithorax and polycomb group (EAP) Additional sex combs like 1	MDS, AML, bohring-Opitz syndrome	Mutation
BMI-1	PRC1 subunit	Ovarian, mantle cell lymphomas and Merkel cell carcinomas	Overexpression
G9a	Histone methyltransferase H3K9	HCC, cervical, uterine, ovarian, and breast cancer	Aberrant expression
PRMT1/5	Protein arginine methyltransferase	Breast/gastric	Aberrant expression
LSD1	Histone demethylase H3K4/H3K9	prostate	Mutation
UTX (KDM6A)	Histone demethylase H3K27	Bladder, breast, kidney, lung, pancreas, esophagus, colon, uterus, brain	Mutation
JARID1B/C	Histone demethylase H3K4/H3K9	Testicular and breast, RCCC	Overexpression
EP300	Histone deacetyltransferase	Breast, colorectal, pancreatic cancer	Mutation
CREBBP	Histone acetyltransferase	Gastric and colorectal, epithelial, ovarian, lung, esophageal cancer	Mutation, overexpression
PCAF	Histone acetyltransferase	Epithelial	Mutation
HDAC2	Histone deacetyltransferase	Colonic, gastric, endometrial cancer	Mutation
SIRT1, HDAC5/7A	Histone deacetyltransferase	Breast, colorectal, prostate cancer	Mutation, aberrant expression

MLL1/2/3: lysine methyltransferase 2A/2D/2C

BRD4: bromodomian containing 4

EZH2: enhancer of Zeste 2 polycomb repressive complex 2 subunit

ASXL: additional sex combs like 2, transcriptional regulator

BMI-1: polycomb group RING finger protein 4

G9a: enchromatic histone lysine methyltransferase 2

PRMT1/5: protein arginine methyltransferase 1

LSD1: lysine demethylase 1A

UTX: lysine demethylase 6A

JARID1B/C: lysine demethylase 5B

EP300: histone acetyltransferase P300

CREBBP: CREB binding protein

PCAF: lysine acetyltransferase 2B

HDAC2: hisone deacetylase 2

SIRT1: NAD-dependent protein deacetylase sirtuin -1

HDAC5/7A: hisone deacetylase 5

HCC: hepatocellular carcinoma

RCCC: renal cell carcinoma

Note: This table is adapted from You, Jueng Soo, and Peter A. Jones. "Cancer genetics and epigenetics: two sides of the same coin?." Cancer cell 22.1 (2012): 9-20 with license ID 1005420-1.

Epigenetic mechanism plays a key role during the tumor initiation. For example, hypermethylation of tumor suppressor gene promoter region are commonly observed in many different cancer types, including *RB*, *BRCA1/2*, and *PTEN*. Some DNA repair genes are also associated with epigenetic silencing, such as *MGMT*. In addition, epigenetic silence can facilitate the selection of mutations in key signaling pathways (97). For instance, in the ovarian carcinoma, mutation frequency is low and large group of genes were silenced through epigenetic mechanism. Furthermore, the cytosine methylation can lead to thymine rather than uracil during the hydrolytic deamination. Therefore, the resulted T:G mismatch is more difficult to repair. The mechanism of this type of point mutation is not fully understood but it may disrupt the gene activity and further disrupt the whole epigenetic regulation mechanism.

In summary, genetic and epigenetic are not independent mechanisms during the tumor initiation and progression, but closely interacted. Until now, no clear evidence shown that the *ATRX* or *TERT*p mutation can cause the DNA methylation profile changes. However, they still hold the potential by changing other key methylation maintenance genes through complex pathway. More interestingly, chr1p19q codel may cause the DNA methylation profile change by the LOH of key genes located on either chr1p or chr19q.

4.5 Conclusion

My classifier demonstrates that DNA methylation signatures accurately predict and deconvolute somatic genomic alterations in human gliomas, thus emphasizing the extensive and significant relationship between cancer's epigenetic signatures and somatic genomic alterations. Given that all predictors are based on a single experimental platform, the Infinium HM BeadChip arrays, the UniD classifier lends itself to the clinical diagnostic setting. The cost of its array, the processing time, and tissue requirements are significantly less than individual sequencing, immunohistochemistry, and copy number assays (for example, fluorescence *in situ* hybridization) currently used clinically with these tumors. Moreover, the Infinium array is suitable for FFPE samples and can be easily adapted for clinical diagnostic tissues. The MGMT-STP27(46) and InfiniumPurify (98) assays, which are based on the same array platform, allow for even further unification of glioma biomarkers into a single assay. Furthermore, the successful development of this DNA methylation-based, infiltrating glioma-specific classifier highlights that methylation-based tumor classification systems can be easily developed for other tumor types, not only for genomic alterations based classification, but for further grading and prognosis, such as for breast(53) or lung cancer (99). For predictive model accessibility, I developed an R package named UniD for rapid determination of biomarker status in gliomas (available on GitHub).

I developed and evaluated a DNA methylation microarray-based classifier that accurately predicts and deconvolutes somatic genomic alterations in gliomas, shows improved enrichment for characteristic genomic alterations and lends itself to clinical diagnostic settings. The array

cost, processing time, and tissue requirements are significantly less than what is currently used clinically in these tumors. Moreover, the Infinium array is suitable for FFPE samples and can be easily adapted for clinical diagnostic tissues.

Bibliography

1. Louis, D. N., A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison. 2016. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* 131: 803-820.
2. Krex, D., B. Klink, C. Hartmann, A. von Deimling, T. Pietsch, M. Simon, M. Sabel, J. P. Steinbach, O. Heese, G. Reifenberger, and others. 2007. Long-term survival with glioblastoma multiforme. *Brain* 130: 2596-2606.
3. Ostrom, Q. T., H. Gittleman, G. Truitt, A. Boscia, C. Kruchko, and J. S. Barnholtz-Sloan. 2018. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2011-2015. *Neuro Oncol* 20: iv1-iv86.
4. Cancer Genome Atlas Research, N., D. J. Brat, R. G. Verhaak, K. D. Aldape, W. K. Yung, S. R. Salama, L. A. Cooper, E. Rheinbay, C. R. Miller, M. Vitucci, O. Morozova, A. G. Robertson, H. Noushmehr, P. W. Laird, A. D. Cherniack, R. Akbani, J. T. Huse, G. Ciriello, L. M. Poisson, J. S. Barnholtz-Sloan, M. S. Berger, C. Brennan, R. R. Colen, H. Colman, A. E. Flanders, C. Giannini, M. Grifford, A. Iavarone, R. Jain, I. Joseph, J. Kim, K. Kasaian, T. Mikkelsen, B. A. Murray, B. P. O'Neill, L. Pachter, D. W. Parsons, C. Sougnez, E. P. Sulman, S. R. Vandenberg, E. G. Van Meir, A. von Deimling, H. Zhang, D. Crain, K. Lau, D. Mallery, S. Morris, J. Paulauskis, R. Penny, T. Shelton, M. Sherman, P. Yena, A. Black, J. Bowen, K. Dicostanzo, J. Gastier-Foster, K. M. Leraas, T. M. Lichtenberg, C. R. Pierson, N. C. Ramirez, C. Taylor, S. Weaver, L. Wise, E. Zmuda, T. Davidsen, J. A. Demchok, G. Eley, M. L. Ferguson, C. M. Hutter, K. R. Mills Shaw, B. A. Ozenberger, M. Sheth, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, B. Ayala, J. Baboud, S. Chudamani, M. A. Jensen, J. Liu, T. Pihl, R. Raman, Y. Wan, Y. Wu, A. Ally, J. T. Auman, M. Balasundaram, S. Balu, S. B. Baylin, R. Beroukhim, M. S. Bootwalla, R. Bowlby, C. A. Bristow, D. Brooks, Y. Butterfield, R. Carlsen, S. Carter, L. Chin, A. Chu, E. Chuah, K.

Cibulskis, A. Clarke, S. G. Coetzee, N. Dhalla, T. Fennell, S. Fisher, S. Gabriel, G. Getz, R. Gibbs, R. Guin, A. Hadjipanayis, D. N. Hayes, T. Hinoue, K. Hoadley, R. A. Holt, A. P. Hoyle, S. R. Jefferys, S. Jones, C. D. Jones, R. Kucherlapati, P. H. Lai, E. Lander, S. Lee, L. Lichtenstein, Y. Ma, D. T. Maglinte, H. S. Mahadeshwar, M. A. Marra, M. Mayo, S. Meng, M. L. Meyerson, P. A. Mieczkowski, R. A. Moore, L. E. Mose, A. J. Mungall, A. Pantazi, M. Parfenov, P. J. Park, J. S. Parker, C. M. Perou, A. Protopopov, X. Ren, J. Roach, T. S. Sabedot, J. Schein, S. E. Schumacher, J. G. Seidman, S. Seth, H. Shen, J. V. Simons, P. Sipahimalani, M. G. Soloway, X. Song, H. Sun, B. Tabak, A. Tam, D. Tan, J. Tang, N. Thiessen, T. Triche, Jr., D. J. Van Den Berg, U. Veluvolu, S. Waring, D. J. Weisenberger, M. D. Wilkerson, T. Wong, J. Wu, L. Xi, A. W. Xu, L. Yang, T. I. Zack, J. Zhang, B. A. Aksoy, H. Arachchi, C. Benz, B. Bernard, D. Carlin, J. Cho, D. DiCara, S. Frazer, G. N. Fuller, J. Gao, N. Gehlenborg, D. Haussler, D. I. Heiman, L. Iype, A. Jacobsen, Z. Ju, S. Katzman, H. Kim, T. Knijnenburg, R. B. Kreisberg, M. S. Lawrence, W. Lee, K. Leinonen, P. Lin, S. Ling, W. Liu, Y. Liu, Y. Liu, Y. Lu, G. Mills, S. Ng, M. S. Noble, E. Paull, A. Rao, S. Reynolds, G. Saksena, Z. Sanborn, C. Sander, N. Schultz, Y. Senbabaoglu, R. Shen, I. Shmulevich, R. Sinha, J. Stuart, S. O. Sumer, Y. Sun, N. Tasman, B. S. Taylor, D. Voet, N. Weinhold, J. N. Weinstein, D. Yang, K. Yoshihara, S. Zheng, W. Zhang, L. Zou, T. Abel, S. Sadeghi, M. L. Cohen, J. Eschbacher, E. M. Hattab, A. Raghunathan, M. J. Schniederjan, D. Aziz, G. Barnett, W. Barrett, D. D. Bigner, L. Boice, C. Brewer, C. Calatozzolo, B. Campos, C. G. Carlotti, Jr., T. A. Chan, L. Cuppini, E. Curley, S. Cuzzubbo, K. Devine, F. DiMeco, R. Duell, J. B. Elder, A. Fehrenbach, G. Finocchiaro, W. Friedman, J. Fulop, J. Gardner, B. Hermes, C. Herold-Mende, C. Jungk, A. Kendler, N. L. Lehman, E. Lipp, O. Liu, R. Mandt, M. McGraw, R. McLendon, C. McPherson, L. Neder, P. Nguyen, A. Noss, R. Nunziata, Q. T. Ostrom, C. Palmer, A. Perin, B. Pollo, A. Potapov, O. Potapova, W. K. Rathmell, D. Rotin, L. Scarpace, C. Schilero, K. Senecal, K. Shimmel, V. Shurkhay, S. Sifri, R. Singh, A. E. Sloan, K. Smolenski, S. M. Staugaitis, R. Steele, L. Thorne, D. P. Tirapelli, A. Unterberg, M.

Vallurupalli, Y. Wang, R. Warnick, F. Williams, Y. Wolinsky, S. Bell, M. Rosenberg, C. Stewart, F. Huang, J. L. Grimsby, A. J. Radenbaugh, and J. Zhang. 2015. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N Engl J Med* 372: 2481-2498.

5. Hegi, M. E., A.-C. Diserens, T. Gorlia, M.-F. Hamou, N. de Tribolet, M. Weller, J. M. Kros, J. A. Hainfellner, W. Mason, L. Mariani, and others. 2005. MGMT gene silencing and benefit from temozolomide in glioblastoma. *New England Journal of Medicine* 352: 997-1003.

6. Barthel, F. P., W. Wei, M. Tang, E. Martinez-Ledesma, X. Hu, S. B. Amin, K. C. Akdemir, S. Seth, X. Song, Q. Wang, T. Lichtenberg, J. Hu, J. Zhang, S. Zheng, and R. G. Verhaak. 2017. Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat Genet* 49: 349-357.

7. Amorim, J. P., G. Santos, J. Vinagre, and P. Soares. 2016. The Role of ATRX in the Alternative Lengthening of Telomeres (ALT) Phenotype. *Genes (Basel)* 7.

8. Napier, C. E., L. I. Huschtscha, A. Harvey, K. Bower, J. R. Noble, E. A. Hendrickson, and R. R. Reddel. 2015. ATRX represses alternative lengthening of telomeres. *Oncotarget* 6: 16543-16558.

9. Wang, Q., B. Hu, X. Hu, H. Kim, M. Squatrito, L. Scarpace, A. C. deCarvalho, S. Lyu, P. Li, Y. Li, F. Barthel, H. J. Cho, Y. H. Lin, N. Satani, E. Martinez-Ledesma, S. Zheng, E. Chang, C. G. Sauve, A. Olar, Z. D. Lan, G. Finocchiaro, J. J. Phillips, M. S. Berger, K. R. Gabrusiewicz, G. Wang, E. Eskilsson, J. Hu, T. Mikkelsen, R. A. DePinho, F. Muller, A. B. Heimberger, E. P. Sulman, D. H. Nam, and R. G. W. Verhaak. 2017. Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell* 32: 42-56 e46.

10. Phillips, H. S., S. Kharbanda, R. Chen, W. F. Forrest, R. H. Soriano, T. D. Wu, A. Misra, J. M. Nigro, H. Colman, L. Soroceanu, P. M. Williams, Z. Modrusan, B. G. Feuerstein, and K. Aldape. 2006. Molecular subclasses of high-grade glioma predict prognosis,

delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 9: 157-173.

11. Verhaak, R. G., K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, D. N. Hayes, and N. Cancer Genome Atlas Research. 2010. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17: 98-110.

12. Noushmehr, H., D. J. Weisenberger, K. Diefes, H. S. Phillips, K. Pujara, B. P. Berman, F. Pan, C. E. Pelloski, E. P. Sulman, K. P. Bhat, R. G. Verhaak, K. A. Hoadley, D. N. Hayes, C. M. Perou, H. K. Schmidt, L. Ding, R. K. Wilson, D. Van Den Berg, H. Shen, H. Bengtsson, P. Neuvial, L. M. Cope, J. Buckley, J. G. Herman, S. B. Baylin, P. W. Laird, and K. Aldape. 2010. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell 2010* 17: 510-522. doi: 510.1016/j.ccr.2010.1003.1017. Epub 2010 Apr 1015.

13. Toyota, M., N. Ahuja, M. Ohe-Toyota, J. G. Herman, S. B. Baylin, and J.-P. J. Issa. 1999. CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences* 96: 8681-8686.

14. Capper, D., D. T. W. Jones, M. Sill, V. Hovestadt, D. Schrimpf, D. Sturm, C. Koelsche, F. Sahm, L. Chavez, D. E. Reuss, A. Kratz, A. K. Wefers, K. Huang, K. W. Pajtler, L. Schweizer, D. Stichel, A. Olar, N. W. Engel, K. Lindenberg, P. N. Harter, A. K. Braczynski, K. H. Plate, H. Dohmen, B. K. Garvalov, R. Coras, A. Holsken, E. Hewer, M. Bewerunge-Hudler, M. Schick, R. Fischer, R. Beschorner, J. Schittenhelm, O. Staszewski, K. Wani, P. Varlet, M. Pages, P. Temming, D. Lohmann, F. Selt, H. Witt, T. Milde, O. Witt, E. Aronica, F. Giangaspero, E. Rushing, W. Scheurlen, C. Geisenberger, F. J. Rodriguez,

A. Becker, M. Preusser, C. Haberler, R. Bjerkvig, J. Cryan, M. Farrell, M. Deckert, J. Hench, S. Frank, J. Serrano, K. Kannan, A. Tsirigos, W. Bruck, S. Hofer, S. Brehmer, M. Seiz-Rosenhagen, D. Hanggi, V. Hans, S. Rozsnoki, J. R. Hansford, P. Kohlhof, B. W. Kristensen, M. Lechner, B. Lopes, C. Mawrin, R. Ketter, A. Kulozik, Z. Khatib, F. Heppner, A. Koch, A. Jouvet, C. Keohane, H. Muhleisen, W. Mueller, U. Pohl, M. Prinz, A. Benner, M. Zapatka, N. G. Gottardo, P. H. Driever, C. M. Kramm, H. L. Muller, S. Rutkowski, K. von Hoff, M. C. Fruhwald, A. Gnekow, G. Fleischhack, S. Tippelt, G. Calaminus, C. M. Monoranu, A. Perry, C. Jones, T. S. Jacques, B. Radlwimmer, M. Gessi, T. Pietsch, J. Schramm, G. Schackert, M. Westphal, G. Reifenberger, P. Wesseling, M. Weller, V. P. Collins, I. Blumcke, M. Bendszus, J. Debus, A. Huang, N. Jabado, P. A. Northcott, W. Paulus, A. Gajjar, G. W. Robinson, M. D. Taylor, Z. Jaunmuktane, M. Ryzhova, M. Platten, A. Unterberg, W. Wick, M. A. Karajannis, M. Mittelbronn, T. Acker, C. Hartmann, K. Aldape, U. Schuller, R. Buslei, P. Lichter, M. Kool, C. Herold-Mende, D. W. Ellison, M. Hasselblatt, M. Snuderl, S. Brandner, A. Korshunov, A. von Deimling, and S. M. Pfister. 2018. DNA methylation-based classification of central nervous system tumours. *Nature*.

15. Ceccarelli, M., F. P. Barthel, T. M. Malta, T. S. Sabedot, S. R. Salama, B. A. Murray, O. Morozova, Y. Newton, A. Radenbaugh, S. M. Pagnotta, S. Anjum, J. Wang, G. Manyam, P. Zoppoli, S. Ling, A. A. Rao, M. Grifford, A. D. Cherniack, H. Zhang, L. Poisson, C. G. Carlotti, Jr., D. P. Tirapelli, A. Rao, T. Mikkelsen, C. C. Lau, W. K. Yung, R. Rabadan, J. Huse, D. J. Brat, N. L. Lehman, J. S. Barnholtz-Sloan, S. Zheng, K. Hess, G. Rao, M. Meyerson, R. Beroukhim, L. Cooper, R. Akbani, M. Wrensch, D. Haussler, K. D. Aldape, P. W. Laird, D. H. Gutmann, T. R. Network, H. Noushmehr, A. Iavarone, and R. G. Verhaak. 2016. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* 164: 550-563.

16. Xu, W., H. Yang, Y. Liu, Y. Yang, P. Wang, S. H. Kim, S. Ito, C. Yang, P. Wang, M. T. Xiao, L. X. Liu, W. Q. Jiang, J. Liu, J. Y. Zhang, B. Wang, S. Frye, Y. Zhang, Y. H. Xu, Q. Y. Lei, K. L. Guan, S. M. Zhao, and Y. Xiong. 2011. Oncometabolite 2-hydroxyglutarate

is a competitive inhibitor of alpha-ketoglutarate-dependent dioxygenases. *Cancer Cell* 19: 17-30.

17. Yen, K., J. Travins, F. Wang, M. D. David, E. Artin, K. Straley, A. Padyana, S. Gross, B. DeLaBarre, E. Tobin, Y. Chen, R. Nagaraja, S. Choe, L. Jin, Z. Konteatis, G. Cianchetta, J. O. Saunders, F. G. Salituro, C. Quivoron, P. Opolon, O. Bawa, V. Saada, A. Paci, S. Broutin, O. A. Bernard, S. de Botton, B. S. Marteyn, M. Pilichowska, Y. Xu, C. Fang, F. Jiang, W. Wei, S. Jin, L. Silverman, W. Liu, H. Yang, L. Dang, M. Dorsch, V. Penard-Lacronique, S. A. Biller, and S. M. Su. 2017. AG-221, a First-in-Class Therapy Targeting Acute Myeloid Leukemia Harboring Oncogenic IDH2 Mutations. *Cancer Discov* 7: 478-493.

18. Popovici-Muller, J., R. M. Lemieux, E. Artin, J. O. Saunders, F. G. Salituro, J. Travins, G. Cianchetta, Z. Cai, D. Zhou, D. Cui, P. Chen, K. Straley, E. Tobin, F. Wang, M. D. David, V. Penard-Lacronique, C. Quivoron, V. Saada, S. de Botton, S. Gross, L. Dang, H. Yang, L. Utley, Y. Chen, H. Kim, S. Jin, Z. Gu, G. Yao, Z. Luo, X. Lv, C. Fang, L. Yan, A. Olaharski, L. Silverman, S. Biller, S. M. Su, and K. Yen. 2018. Discovery of AG-120 (Ivosidenib): A First-in-Class Mutant IDH1 Inhibitor for the Treatment of IDH1 Mutant Cancers. *ACS Med Chem Lett* 9: 300-305.

19. Turcan, S., D. Rohle, A. Goenka, L. A. Walsh, F. Fang, E. Yilmaz, C. Campos, A. W. Fabius, C. Lu, P. S. Ward, C. B. Thompson, A. Kaufman, O. Guryanova, R. Levine, A. Heguy, A. Viale, L. G. Morris, J. T. Huse, I. K. Mellinghoff, and T. A. Chan. 2012. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* 483: 479-483.

20. Noushmehr, H., D. J. Weisenberger, K. Diefes, H. S. Phillips, K. Pujara, B. P. Berman, F. Pan, C. E. Pelloski, E. P. Sulman, K. P. Bhat, R. G. Verhaak, K. A. Hoadley, D. N. Hayes, C. M. Perou, H. K. Schmidt, L. Ding, R. K. Wilson, D. Van Den Berg, H. Shen, H. Bengtsson, P. Neuvial, L. M. Cope, J. Buckley, J. G. Herman, S. B. Baylin, P. W. Laird, K. Aldape, and N. Cancer Genome Atlas Research. 2010. Identification of a CpG island

methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 17: 510-522.

21. Baldewpersad Tewarie, N. M., I. A. Burgers, Y. Dawood, H. C. den Boon, M. G. den Brok, J. H. Klunder, K. B. Koopmans, E. Rademaker, H. B. van den Broek, S. M. van den Bersselaar, J. J. Witjes, C. J. Van Noorden, and N. A. Atai. 2013. NADP+ -dependent IDH1 R132 mutation and its relevance for glioma patient survival. *Med Hypotheses* 80: 728-731.

22. Amankulor, N. M., Y. Kim, S. Arora, J. Kargl, F. Szulzewsky, M. Hanke, D. H. Margineantu, A. Rao, H. Bolouri, J. Delrow, D. Hockenbery, A. M. Houghton, and E. C. Holland. 2017. Mutant IDH1 regulates the tumor-associated immune system in gliomas. *Genes Dev* 31: 774-786.

23. Jafri, M. A., S. A. Ansari, M. H. Alqahtani, and J. W. Shay. 2016. Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies. *Genome Med* 8: 69.

24. Vinagre, J., A. Almeida, H. Populo, R. Batista, J. Lyra, V. Pinto, R. Coelho, R. Celestino, H. Prazeres, L. Lima, M. Melo, A. G. da Rocha, A. Preto, P. Castro, L. Castro, F. Pardal, J. M. Lopes, L. L. Santos, R. M. Reis, J. Cameselle-Teijeiro, M. Sobrinho-Simoes, J. Lima, V. Maximo, and P. Soares. 2013. Frequency of TERT promoter mutations in human cancers. *Nat Commun* 4: 2185.

25. Dilley, R. L., and R. A. Greenberg. 2015. ALTernative Telomere Maintenance and Cancer. *Trends Cancer* 1: 145-156.

26. Chen, Y. J., V. Hakin-Smith, M. Teo, G. E. Xinarianos, D. A. Jellinek, T. Carroll, D. McDowell, M. R. MacFarlane, R. Boet, B. C. Baguley, A. W. Braithwaite, R. R. Reddel, and J. A. Royds. 2006. Association of mutant TP53 with alternative lengthening of telomeres and favorable prognosis in glioma. *Cancer Res* 66: 6473-6476.

27. Heaphy, C. M., A. P. Subhawong, S. M. Hong, M. G. Goggins, E. A. Montgomery, E. Gabrielson, G. J. Netto, J. I. Epstein, T. L. Lotan, W. H. Westra, M. Shih Ie, C. A.

Iacobuzio-Donahue, A. Maitra, Q. K. Li, C. G. Eberhart, J. M. Taube, D. Rakheja, R. J. Kurman, T. C. Wu, R. B. Roden, P. Argani, A. M. De Marzo, L. Terracciano, M. Torbenson, and A. K. Meeker. 2011. Prevalence of the alternative lengthening of telomeres telomere maintenance mechanism in human cancer subtypes. *Am J Pathol* 179: 1608-1615.

28. Serakinci, N., S. F. Hoare, M. Kassem, S. P. Atkinson, and W. N. Keith. 2006. Telomerase promoter reprogramming and interaction with general transcription factors in the human mesenchymal stem cell. *Regen Med* 1: 125-131.

29. Sung, P., and H. Klein. 2006. Mechanism of homologous recombination: mediators and helicases take on regulatory functions. *Nat Rev Mol Cell Biol* 7: 739-750.

30. Xue, Y., R. Gibbons, Z. Yan, D. Yang, T. L. McDowell, S. Sechi, J. Qin, S. Zhou, D. Higgs, and W. Wang. 2003. The ATRX syndrome protein forms a chromatin-remodeling complex with Daxx and localizes in promyelocytic leukemia nuclear bodies. *Proc Natl Acad Sci U S A* 100: 10635-10640.

31. Dyer, M. A., Z. A. Qadeer, D. Valle-Garcia, and E. Bernstein. 2017. ATRX and DAXX: Mechanisms and Mutations. *Cold Spring Harb Perspect Med* 7.

32. Clynes, D., C. Jelinska, B. Xella, H. Ayyub, C. Scott, M. Mitson, S. Taylor, D. R. Higgs, and R. J. Gibbons. 2015. Suppression of the alternative lengthening of telomere pathway by the chromatin remodelling factor ATRX. *Nat Commun* 6: 7538.

33. O'Sullivan, R. J., N. Arnoult, D. H. Lackner, L. Oganesian, C. Haggblom, A. Corpet, G. Almouzni, and J. Karlseder. 2014. Rapid induction of alternative lengthening of telomeres by depletion of the histone chaperone ASF1. *Nat Struct Mol Biol* 21: 167-174.

34. Olshen, A. B., E. S. Venkatraman, R. Lucito, and M. Wigler. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5: 557-572.

35. Cairncross, G., M. Wang, E. Shaw, R. Jenkins, D. Brachman, J. Buckner, K. Fink, L. Souhami, N. Laperriere, W. Curran, and M. Mehta. 2013. Phase III trial of

chemoradiotherapy for anaplastic oligodendroglioma: long-term results of RTOG 9402. *J Clin Oncol* 31: 337-343.

36. van den Bent, M. J., A. A. Brandes, M. J. Taphoorn, J. M. Kros, M. C. Kouwenhoven, J. Y. Delattre, H. J. Bernsen, M. Frenay, C. C. Tijssen, W. Grisold, L. Sipos, R. H. Enting, P. J. French, W. N. Dinjens, C. J. Vecht, A. Allgeier, D. Lacombe, T. Gorlia, and K. Hoang-Xuan. 2013. Adjuvant procarbazine, lomustine, and vincristine chemotherapy in newly diagnosed anaplastic oligodendroglioma: long-term follow-up of EORTC brain tumor group study 26951. *J Clin Oncol* 31: 344-350.

37. Cancer Genome Atlas Research, N. 2012. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489: 519-525.

38. Cancer Genome Atlas Research, N. 2011. Integrated genomic analyses of ovarian carcinoma. *Nature* 474: 609-615.

39. Cancer Genome Atlas, N. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487: 330-337.

40. Cancer Genome Atlas, N. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* 490: 61-70.

41. Cancer Genome Atlas Research, N., C. Kandoth, N. Schultz, A. D. Cherniack, R. Akbani, Y. Liu, H. Shen, A. G. Robertson, I. Pashtan, R. Shen, C. C. Benz, C. Yau, P. W. Laird, L. Ding, W. Zhang, G. B. Mills, R. Kucherlapati, E. R. Mardis, and D. A. Levine. 2013. Integrated genomic characterization of endometrial carcinoma. *Nature* 497: 67-73.

42. Figueroa, M. E., O. Abdel-Wahab, C. Lu, P. S. Ward, J. Patel, A. Shih, Y. Li, N. Bhagwat, A. Vasanthakumar, H. F. Fernandez, M. S. Tallman, Z. Sun, K. Wolniak, J. K. Peeters, W. Liu, S. E. Choe, V. R. Fantin, E. Paietta, B. Lowenberg, J. D. Licht, L. A. Godley, R. Delwel, P. J. Valk, C. B. Thompson, R. L. Levine, and A. Melnick. 2010. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell* 18: 553-567.

43. Stupp, R., W. P. Mason, M. J. van den Bent, M. Weller, B. Fisher, M. J. Taphoorn, K. Belanger, A. A. Brandes, C. Marosi, U. Bogdahn, J. Curschmann, R. C. Janzer, S. K. Ludwin, T. Gorlia, A. Allgeier, D. Lacombe, J. G. Cairncross, E. Eisenhauer, R. O. Mirimanoff, R. European Organisation for, T. Treatment of Cancer Brain, G. Radiotherapy, and G. National Cancer Institute of Canada Clinical Trials. 2005. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med* 352: 987-996.

44. Stupp, R., M. E. Hegi, W. P. Mason, M. J. van den Bent, M. J. Taphoorn, R. C. Janzer, S. K. Ludwin, A. Allgeier, B. Fisher, K. Belanger, P. Hau, A. A. Brandes, J. Gijtenbeek, C. Marosi, C. J. Vecht, K. Mokhtari, P. Wesseling, S. Villa, E. Eisenhauer, T. Gorlia, M. Weller, D. Lacombe, J. G. Cairncross, R. O. Mirimanoff, R. European Organisation for, T. Treatment of Cancer Brain, G. Radiation Oncology, and G. National Cancer Institute of Canada Clinical Trials. 2009. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol* 10: 459-466.

45. Herman, J. G., J. R. Graff, S. Myohanen, B. D. Nelkin, and S. B. Baylin. 1996. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci U S A* 93: 9821-9826.

46. Bady, P., D. Sciuscio, A.-C. Diserens, J. Bloch, M. J. Van Den Bent, C. Marosi, P.-Y. Dietrich, M. Weller, L. Mariani, and F. L. Heppner. 2012. MGMT methylation analysis of glioblastoma on the Infinium methylation BeadChip identifies two distinct CpG regions associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and CIMP-status. *Acta neuropathologica* 124: 547-560.

47. Sharma, S., T. K. Kelly, and P. A. Jones. 2010. Epigenetics in cancer. *Carcinogenesis* 31: 27-36.

48. Ehrlich, M. 2009. DNA hypomethylation in cancer cells. *Epigenomics* 1: 239-259.

49. Feinberg, A. P., M. A. Koldobskiy, and A. Gondor. 2016. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature reviews. Genetics* 17: 284-299.

50. Sahu, A., U. Singhal, and A. M. Chinnaiyan. 2015. Long noncoding RNAs in cancer: from function to translation. *Trends Cancer* 1: 93-109.

51. Brennan, C. W., R. G. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S. R. Salama, S. Zheng, D. Chakravarty, J. Z. Sanborn, S. H. Berman, R. Beroukhim, B. Bernard, C. J. Wu, G. Genovese, I. Shmulevich, J. Barnholtz-Sloan, L. Zou, R. Vegesna, S. A. Shukla, G. Ciriello, W. K. Yung, W. Zhang, C. Sougnez, T. Mikkelsen, K. Aldape, D. D. Bigner, E. G. Van Meir, M. Prados, A. Sloan, K. L. Black, J. Eschbacher, G. Finocchiaro, W. Friedman, D. W. Andrews, A. Guha, M. Iacocca, B. P. O'Neill, G. Foltz, J. Myers, D. J. Weisenberger, R. Penny, R. Kucherlapati, C. M. Perou, D. N. Hayes, R. Gibbs, M. Marra, G. B. Mills, E. Lander, P. Spellman, R. Wilson, C. Sander, J. Weinstein, M. Meyerson, S. Gabriel, P. W. Laird, D. Haussler, G. Getz, and L. Chin. 2013. The somatic genomic landscape of glioblastoma. *Cell* 155: 462-477.

52. Sahm, F., D. Schrimpf, D. Stichel, D. T. W. Jones, T. Hielscher, S. Schefzyk, K. Okonechnikov, C. Koelsche, D. E. Reuss, D. Capper, D. Sturm, H. G. Wirsching, A. S. Berghoff, P. Baumgarten, A. Kratz, K. Huang, A. K. Wefers, V. Hovestadt, M. Sill, H. P. Ellis, K. M. Kurian, A. F. Okuducu, C. Jungk, K. Drueschler, M. Schick, M. Bewerunge-Hudler, C. Mawrin, M. Seiz-Rosenhagen, R. Ketter, M. Simon, M. Westphal, K. Lamszus, A. Becker, A. Koch, J. Schittenhelm, E. J. Rushing, V. P. Collins, S. Brehmer, L. Chavez, M. Platten, D. Hanggi, A. Unterberg, W. Paulus, W. Wick, S. M. Pfister, M. Mittelbronn, M. Preusser, C. Herold-Mende, M. Weller, and A. von Deimling. 2017. DNA methylation-based classification and grading system for meningioma: a multicentre, retrospective analysis. *Lancet Oncol* 18: 682-694.

53. Stefansson, O. A., S. Moran, A. Gomez, S. Sayols, C. Arribas-Jorba, J. Sandoval, H. Hilmarsdottir, E. Olafsdottir, L. Tryggvadottir, J. G. Jonasson, J. Eyfjord, and M. Esteller.

2015. A DNA methylation-based definition of biologically distinct breast cancer subtypes. *Mol Oncol* 9: 555-568.

54. Moran, S., A. Martinez-Cardus, S. Sayols, E. Musulen, C. Balana, A. Estival-Gonzalez, C. Moutinho, H. Heyn, A. Diaz-Lagares, M. C. de Moura, G. M. Stella, P. M. Comoglio, M. Ruiz-Miro, X. Matias-Guiu, R. Pazo-Cid, A. Anton, R. Lopez-Lopez, G. Soler, F. Longo, I. Guerra, S. Fernandez, Y. Assenov, C. Plass, R. Morales, J. Carles, D. Bowtell, L. Mileshkin, D. Sia, R. Tothill, J. Tabernero, J. M. Llovet, and M. Esteller. 2016. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol* 17: 1386-1395.

55. Herceg, Z., and P. Hainaut. 2007. Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis. *Mol Oncol* 1: 26-41.

56. Qiu, J., B. Peng, Y. Tang, Y. Qian, P. Guo, M. Li, J. Luo, B. Chen, H. Tang, C. Lu, M. Cai, Z. Ke, W. He, Y. Zheng, D. Xie, B. Li, and Y. Yuan. 2017. CpG Methylation Signature Predicts Recurrence in Early-Stage Hepatocellular Carcinoma: Results From a Multicenter Study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 35: 734-742.

57. Visvanathan, K., M. S. Fackler, Z. Zhang, Z. A. Lopez-Bujanda, S. C. Jeter, L. J. Sokoll, E. Garrett-Mayer, L. M. Cope, C. B. Umbricht, D. M. Euhus, A. Forero, A. M. Storniolo, R. Nanda, N. U. Lin, L. A. Carey, J. N. Ingle, S. Sukumar, and A. C. Wolff. 2017. Monitoring of Serum DNA Methylation as an Early Independent Marker of Response and Survival in Metastatic Breast Cancer: TBCRC 005 Prospective Biomarker Study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 35: 751-758.

58. Plass, C., S. M. Pfister, A. M. Lindroth, O. Bogatyrova, R. Claus, and P. Lichter. 2013. Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nat Rev Genet* 14: 765-780.

59. Nordlund, J., C. L. Backlin, P. Wahlberg, S. Busche, E. C. Berglund, M. L. Eloranta, T. Flaegstad, E. Forestier, B. M. Frost, A. Harila-Saari, M. Heyman, O. G. Jonsson, R.

Larsson, J. Palle, L. Ronnblom, K. Schmiegelow, D. Sinnett, S. Soderhall, T. Pastinen, M. G. Gustafsson, G. Lonnerholm, and A. C. Syvanen. 2013. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol* 14: r105.

60. Zhou, W., P. W. Laird, and H. Shen. 2017. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* 45: e22.

61. Du, P., X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin. 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics* 11: 587.

62. Dedeurwaerder, S., M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks. 2011. Evaluation of the Infinium Methylation 450K technology. *Epigenomics* 3: 771-784.

63. Touleimat, N., and J. Tost. 2012. Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* 4: 325-341.

64. Maksimovic, J., L. Gordon, and A. Oshlack. 2012. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* 13: R44.

65. Yousefi, P., K. Huen, R. Aguilar Schall, A. Decker, E. Elboudwarej, H. Quach, L. Barcellos, and N. Holland. 2013. Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. *Epigenetics* 8: 1141-1152.

66. Teschendorff, A. E., F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, and S. Beck. 2013. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 29: 189-196.

67. Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 33: 1-22.

68. Bormann, F., M. Rodriguez-Paredes, F. Lasitschka, D. Edelmann, T. Musch, A. Benner, Y. Bergman, S. M. Dieter, C. R. Ball, H. Glimm, H. G. Linhart, and F. Lyko. 2018. Cell-of-Origin DNA Methylation Signatures Are Maintained during Colorectal Carcinogenesis. *Cell Rep* 23: 3407-3418.

69. Kulis, M., A. C. Queiros, R. Beekman, and J. I. Martin-Subero. 2013. Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. *Biochim Biophys Acta* 1829: 1161-1174.

70. Yang, J. 2015. Integrated analysis of DNA methylation profiles in the malignant brain tumor glioblastom.

71. Dedeurwaerder, S., M. Defrance, M. Bizet, E. Calonne, G. Bontempi, and F. Fuks. 2013. A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in bioinformatics*: bbt054.

72. Dedeurwaerder, S., M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks. 2011. Evaluation of the Infinium Methylation 450K technology. *Epigenomics* 3: 771-784.

73. Brunet, J.-P., P. Tamayo, T. R. Golub, and J. P. Mesirov. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences* 101: 4164-4169.

74. Wilhelm-Benartzi, C. S., D. C. Koestler, M. R. Karagas, J. M. Flanagan, B. C. Christensen, K. T. Kelsey, C. J. Marsit, E. A. Houseman, and R. Brown. 2013. Review of processing and analysis methods for DNA methylation array data. *British journal of cancer* 109: 1394-1402.

75. Yousefi, P., K. Huen, R. A. Schall, A. Decker, E. Elboudwarej, H. Quach, L. Barcellos, and N. Holland. 2013. Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. *Epigenetics* 8: 1141-1152.

76. Nordlund, J., C. L. Backlin, P. Wahlberg, S. Busche, E. C. Berglund, M.-L. Eloranta, T. Flaegstad, E. Forestier, B.-M. Frost, and A. Harila-Saari. 2013. Genome-wide signatures

of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome biology* 14: r105.

77. Tibshirani, J. F. a. T. H. a. R. 2010. glmet R package.

78. Carter, S. L., K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, R. Beroukhim, D. Pellman, D. A. Levine, E. S. Lander, M. Meyerson, and G. Getz. 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 30: 413-421.

79. Cibulskis, K., M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 31: 213-219.

80. Koboldt, D. C., K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25: 2283-2285.

81. Fan, Y., L. Xi, D. S. Hughes, J. Zhang, J. Zhang, P. A. Futreal, D. A. Wheeler, and W. Wang. 2016. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol* 17: 178.

82. Larson, D. E., C. C. Harris, K. Chen, D. C. Koboldt, T. E. Abbott, D. J. Dooling, T. J. Ley, E. R. Mardis, R. K. Wilson, and L. Ding. 2012. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 28: 311-317.

83. Grossman, R. L., A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt. 2016. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* 375: 1109-1112.

84. Hovestadt V, Z. M. conumee: Enhanced copy-number variation analysis using Illumina DNA methylation arrays.

85. Huang da, W., B. T. Sherman, and R. A. Lempicki. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4: 44-57.

86. Wick, W., C. Hartmann, C. Engel, M. Stoffels, J. Felsberg, F. Stockhammer, M. C. Sabel, S. Koeppen, R. Ketter, R. Meyermann, M. Rapp, C. Meisner, R. D. Kortmann, T. Pietsch, O. D. Wiestler, U. Ernemann, M. Bamberg, G. Reifenberger, A. von Deimling, and M. Weller. 2009. NOA-04 randomized phase III trial of sequential radiochemotherapy of anaplastic glioma with procarbazine, lomustine, and vincristine or temozolomide. *J Clin Oncol* 27: 5874-5880.

87. Vlassenbroeck, I., S. Califice, A. C. Diserens, E. Migliavacca, J. Straub, I. Di Stefano, F. Moreau, M. F. Hamou, I. Renard, M. Delorenzi, B. Flamion, J. DiGuiseppi, K. Bierau, and M. E. Hegi. 2008. Validation of real-time methylation-specific PCR to determine O6-methylguanine-DNA methyltransferase gene promoter methylation in glioma. *J Mol Diagn* 10: 332-337.

88. Huang da, W., B. T. Sherman, and R. A. Lempicki. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37: 1-13.

89. Eckel-Passow, J. E., D. H. Lachance, A. M. Molinaro, K. M. Walsh, P. A. Decker, H. Sicotte, M. Pekmezci, T. Rice, M. L. Kosel, I. V. Smirnov, G. Sarkar, A. A. Caron, T. M. Kollmeyer, C. E. Praska, A. R. Chada, C. Halder, H. M. Hansen, L. S. McCoy, P. M. Bracci, R. Marshall, S. Zheng, G. F. Reis, A. R. Pico, B. P. O'Neill, J. C. Buckner, C. Giannini, J. T. Huse, A. Perry, T. Tihan, M. S. Berger, S. M. Chang, M. D. Prados, J. Wiemels, J. K. Wiencke, M. R. Wrensch, and R. B. Jenkins. 2015. Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors. *N Engl J Med* 372: 2499-2508.

90. Jones, B. B. a. M. L. a. L. K. a. J. S. a. J. R. a. E. S. a. G. C. a. Z. M. 2016. mlr: Machine Learning in R. *Journal of Machine Learning Research* 17: 1-5.

91. Reva, B., Y. Antipin, and C. Sander. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 39: e118.

92. Behnan, J., G. Finocchiaro, and G. Hanna. 2019. The landscape of the mesenchymal signature in brain tumours. *Brain* 142: 847-866.

93. Schwalbe, E. C., D. Williamson, J. C. Lindsey, D. Hamilton, S. L. Ryan, H. Megahed, M. Garami, P. Hauser, B. Dembowska-Baginska, D. Perek, P. A. Northcott, M. D. Taylor, R. E. Taylor, D. W. Ellison, S. Bailey, and S. C. Clifford. 2013. DNA methylation profiling of medulloblastoma allows robust subclassification and improved outcome prediction using formalin-fixed biopsies. *Acta Neuropathol* 125: 359-371.

94. Cavalli, F. M. G., M. Remke, L. Rampasek, J. Peacock, D. J. H. Shih, B. Luu, L. Garzia, J. Torchia, C. Nor, A. S. Morrissy, S. Agnihotri, Y. Y. Thompson, C. M. Kuzan-Fischer, H. Farooq, K. Isaev, C. Daniels, B. K. Cho, S. K. Kim, K. C. Wang, J. Y. Lee, W. A. Grajkowska, M. Perek-Polnik, A. Vasiljevic, C. Faure-Conter, A. Jouvet, C. Giannini, A. A. Nageswara Rao, K. K. W. Li, H. K. Ng, C. G. Eberhart, I. F. Pollack, R. L. Hamilton, G. Y. Gillespie, J. M. Olson, S. Leary, W. A. Weiss, B. Lach, L. B. Chambless, R. C. Thompson, M. K. Cooper, R. Vibhakar, P. Hauser, M. C. van Veelen, J. M. Kros, P. J. French, Y. S. Ra, T. Kumabe, E. Lopez-Aguilar, K. Zitterbart, J. Sterba, G. Finocchiaro, M. Massimino, E. G. Van Meir, S. Osuka, T. Shofuda, A. Klekner, M. Zollo, J. R. Leonard, J. B. Rubin, N. Jabado, S. Albrecht, J. Mora, T. E. Van Meter, S. Jung, A. S. Moore, A. R. Hallahan, J. A. Chan, D. P. C. Tirapelli, C. G. Carlotti, M. Fouladi, J. Pimentel, C. C. Faria, A. G. Saad, L. Massimi, L. M. Liau, H. Wheeler, H. Nakamura, S. K. Elbabaa, M. Perezpena-Diazconti, F. Chico Ponce de Leon, S. Robinson, M. Zapotocky, A. Lassaletta, A. Huang, C. E. Hawkins, U. Tabori, E. Bouffet, U. Bartels, P. B. Dirks, J. T. Rutka, G. D. Bader, J. Reimand, A. Goldenberg, V. Ramaswamy, and M. D. Taylor. 2017. Intertumoral Heterogeneity within Medulloblastoma Subgroups. *Cancer Cell* 31: 737-754 e736.

95. Binder, H., E. Willscher, H. Loeffler-Wirth, L. Hopp, D. T. W. Jones, S. M. Pfister, M. Kreuz, D. Gramatzki, E. Fortenbacher, B. Hentschel, M. Tatagiba, U. Herrlinger, H. Vatter, J. Matschke, M. Westphal, D. Krex, G. Schackert, J. C. Tonn, U. Schlegel, H. J. Steiger, W. Wick, R. G. Weber, M. Weller, and M. Loeffler. 2019. DNA methylation, transcriptome

and genetic copy number signatures of diffuse cerebral WHO grade II/III gliomas resolve cancer heterogeneity and development. *Acta Neuropathol Commun* 7: 59.

96. Paul, Y., B. Mondal, V. Patil, and K. Somasundaram. 2017. DNA methylation signatures for 2016 WHO classification subtypes of diffuse gliomas. *Clin Epigenetics* 9: 32.

97. You, J. S., and P. A. Jones. 2012. Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell* 22: 9-20.

98. Qin, Y., H. Feng, M. Chen, H. Wu, and X. Zheng. 2018. InfiniumPurify: An R package for estimating and accounting for tumor purity in cancer methylation research. *Genes Dis* 5: 43-45.

99. Zhang, Q. H., X. H. Dai, Z. M. Dai, and Y. N. Cai. 2015. Genome-scale meta-analysis of DNA methylation during progression of lung adenocarcinoma. *Genet Mol Res* 14: 9200-9214.

Vita

Jie Yang was born in Jiangxi, People's Republic of China, the daughter of Wenqing Yang and Juan Tu. After completing her work at Guixi No.1 Middle School, Guixi, Jiangxi, People's Republic of China in 2008, she entered Harbin Medical University in Harbin, Heilongjiang, People's Republic of China. She received the degree of Bachelor of Medicine with a major in preventive medicine from Harbin Medical University in June, 2013. Then she came to United States of America for her graduate education. She received the degree of Master of Science with a major in epidemiology from The University of Texas Health Science Center at Houston School of Public Health in May, 2015. In August of 2015, she entered MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences for her PhD degree in Quantitative Science, bioinformatics track.

Permanent address:

550 First Avenue

New York, NY 10016