

# Northumbria Research Link

Citation: Thompson, Dean (2022) Transforming cancer molecular diagnostics: Molecular subgrouping of medulloblastoma via lowdepth whole genome bisulfite sequencing. Doctoral thesis, Northumbria University.

This version was downloaded from Northumbria Research Link:  
<https://nrl.northumbria.ac.uk/id/eprint/50105/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>



# Foreword

## Declaration

I hereby declare that this thesis submission, entitled:

***Transforming cancer molecular diagnostics:  
Molecular subgrouping of medulloblastoma via low-depth whole-genome bisulfite sequencing***

consists of work conducted independently by myself throughout the course of the Northumbria PhD Studentship Programme, except when acknowledged within the thesis text.

I confirm that:

- i) No component of this thesis has been submitted as part of a degree or other qualification within Northumbria University or any other academic institution.
- ii) Any published work that is referred to in this document has been attributed to the correct author(s).
- iii) Any assistance provided by others is acknowledged clearly.

Signed:

Dean Thompson  
January 2022

## Acknowledgements

First and foremost, I would like to thank my primary supervisor, Dr. Ed Schwalbe, for guiding me through the last three years and for sharing his bioinformatics expertise with me. Without you I would never have made it past the first year of this journey and would never have fallen in love with bioinformatics during my undergraduate years in the first place.

I would also like to thank Dr. Debbie Hicks, my secondary supervisor for providing me with detailed feedback, having patience with my bad grammar habits and for supporting me when our sequencing partner made a mess of library preparation and put the entire project in jeopardy.

Thanks to Prof. Steven Clifford for offering me a workspace within the Paediatric Brain Tumour Research Group in the Herschel Building at Newcastle University, for aiding me with funding and helping to shape the narrative of my project.

I would like to extend my utmost appreciation to (future Dr) Jemma Castle, who prepared all sample DNA that was sequenced as part of the internal cohort in this project and to Dr. Yura Grabovska, for helping me with DNA methylation array processing and trusting me enough to pass on the responsibility of handling array sample processing on behalf of the group.

Next, I would like to thank Dr. Karen Bower, for helping me get going on the ROCKET high performance computing cluster within Newcastle University. You were patient with me even when I caused backlogs on the server when I was just starting out and provided crucial assistance in helping me learn to set up bespoke package environments within the server architecture.

Finally, I would like to provide an honourable mention to my rock, my 2019 MacBook Pro. Since we joined forces at the start of this journey you have never stopped being there for me. You have kept my most embarrassing coding mistakes private whilst I learned to write in R, Python, and shell script and for that I will never stop being grateful. You have earned your retirement. Rest easy friend.



## Dedication

*To my mum Julie and dad Alan for supporting me endlessly, even though you remind me daily that you don't have a clue what I'm doing.*

*To my stepfather, George, who sadly passed away late in my penultimate year due to Oesophageal Cancer and Covid-19. It pains me that you will never see me graduate, but I know you would be proud of me.*

*And finally, to my partner Amber (and her mum Donna!) for putting up with the brunt of my stress, helping to pick me up when I'm down and for helping me through some of the toughest moments of the past three years.*

## Table of Contents

<b>Foreword</b>	<b>1</b>
<b>Declaration</b>	<b>1</b>
<b>Acknowledgements</b>	<b>2</b>
<b>Dedication</b>	<b>3</b>
<b>Abstract</b>	<b>9</b>
<b>List of abbreviations</b>	<b>10</b>
<b>List of figures</b>	<b>14</b>
<b>List of tables</b>	<b>17</b>
<b>Chapter 1 - Introduction</b>	<b>19</b>
<b>1.1 – Cancer: Overview</b>	<b>20</b>
<b>1.2 – Cancer: Epidemiology</b>	<b>20</b>
<b>1.3 – Drivers of tumorigenesis</b>	<b>21</b>
1.3.1 – Resisting apoptosis	25
1.3.2 – Invasion & metastasis	27
1.3.3 – Angiogenesis	28
1.3.4 – Genome instability & mutation	29
1.3.5 – Dereglulation of metabolism	30
1.3.6 – Immune system evasion	32
1.3.7 – Inflammation	32
1.3.8 – Sustained proliferation	33
1.3.9 – Cell immortality	33
1.3.10 – Cancer stem cells	34
1.3.11 – Enabling Characteristics	35
<b>1.4 – Oncogenes &amp; tumour suppressor genes</b>	<b>36</b>
<b>1.5 – Epigenetics</b>	<b>36</b>
1.5.1 – DNA methylation	37
1.5.2 – Histone modifications	40
1.5.3 – Chromatin remodelling	40
1.5.4 – Methylation & cancer	41
<b>1.6 – Tumours of the central nervous system</b>	<b>43</b>
1.6.1 – Types of brain tumour	43
1.6.2 – Paediatric brain tumours	44
<b>1.7 – Medulloblastoma</b>	<b>45</b>
1.7.1 – Epidemiology	46
1.7.2 – Clinical presentation/symptoms	47
1.7.3 – Histology	48
1.7.3.1 – Classic histology	50
1.7.3.2 – Desmoplastic/nodular	50
1.7.3.3 – Medulloblastoma with extensive nodularity (MBEN)	50
1.7.3.4 – Large cell/anaplastic	51
1.7.4 – Medulloblastoma is comprised of distinct molecular entities	51
1.7.4.1 – WNT medulloblastoma	52
1.7.4.2 – WNT signalling & implicated genes	53
1.7.4.3 – SHH medulloblastoma	55

1.7.4.4– SHH signalling & implicated genes	56
1.7.4.5 – Group 3 and Group 4 medulloblastoma	57
1.7.4.6 – Identifications of subtypes within Group 3 and Group 4 medulloblastoma	59
1.7.5 – Diagnosis of medulloblastoma	61
1.7.6 – Determination of molecular subgroup	63
1.7.7 – Treatment: General protocols	64
1.7.7.1 – Assignment of risk	65
1.7.8 – Complications of treatment	66
<b>1.8 – Next generation sequencing &amp; cancer research</b>	<b>67</b>
<b>1.9 – Assessing DNA methylation</b>	<b>68</b>
<b>1.10– Limitations of DNA methylation microarrays / advantages of methylation sequencing</b>	<b>71</b>
<b>1.11 – Summary &amp; aims</b>	<b>73</b>
<b>Chapter 2 – Materials &amp; Methods</b>	<b>75</b>
<b>2.1 – Study Cohorts</b>	<b>76</b>
<b>2.2 – Sample preparation</b>	<b>89</b>
2.2.1 – Library preparation & sequencing	90
2.2.2 – Sequencing data quality control (external)	90
<b>2.3 – Bioinformatic techniques</b>	<b>90</b>
2.3.1 – Array data processing: normalisation	91
2.3.2 – Alignment of bisulfite treated DNA	92
2.3.3 - Imputation	93
2.3.3.1 – Imputation of array beta values via k-nearest neighbours	94
2.3.3.2 – Imputation of WGBS beta values via BoostMe	95
2.3.4 – Dimensionality reduction / visualisation techniques	96
2.3.4.1 – Principal component analysis	96
2.3.4.2 – t-distributed stochastic neighbour embedding	97
2.3.4.3 – Uniform manifold approximation and projection	97
2.3.5 – DNA methylation-based classification	98
2.3.5.1 – Elastic net	99
2.3.5.2 – Random forest	99
2.3.5.3 – Extreme gradient boosting	99
2.3.5.4 – Support vector machines	100
2.3.5.5 – Logic regression	100
2.3.6 – NGS and methylation array-based copy number estimation	101
<b>Chapter 3 – Establishment of a robust processing methodology for the clinical analysis of WGBS samples sequenced at low-pass</b>	<b>102</b>
<b>3.1 – Introduction</b>	<b>103</b>
<b>3.2 – Aims</b>	<b>107</b>
<b>3.3 – Materials &amp; Methods</b>	<b>108</b>
3.3.1 – Analysis Overview	109
3.3.2 – Raw data processing – WGBS	109
3.3.2.1 – Computing resources/software	109
3.3.2.2 – Quality control & read trimming	109
3.3.2.3 – Alignment and deduplication	111
3.3.2.4 – Methylation bias assessment & methylation data extraction	111
3.3.3 – External cohort data handling	113
3.3.4 – Methylation data processing	113
3.3.4.1 - Computing resources / software	113
3.3.4.2 – QC & filtering of DNA methylation microarray data	113

3.3.4.3 – QC & filtering of WGBS methylation data	114
3.3.5 – Data downsampling	115
3.3.6 – Imputation via BoostMe	115
3.3.7 – Dataset summary	116
3.3.8 – Liftover & interplatform correlation analysis	117
<b>3.4 – Results &amp; discussion</b>	<b>118</b>
3.4.1 – Establishment of a robust processing pipeline for clinical low-depth WGBS samples	118
3.4.1.1 – Quality control analysis of sequencing data	118
3.4.1.2 – Quality control analysis of methylation data – WGBS	124
3.4.1.3 – Differences between internal and external cohort downsampling approaches	125
3.4.1.4 – Quality control analysis of methylation data - Array	129
3.4.1.5 – Methylation values calculated from WGBS sample data display increased bimodal distribution	129
3.4.2 – Machine learning imputation predicts robust, accurate methylation values in low-pass WGBS samples	132
3.4.2.1 – Model performance	132
3.4.2.2 – Potential for future optimisation	138
3.4.3 – WGBS platform produces highly correlated methylation data to corresponding CpGs located on the array platform at low-sequencing depths	139
3.4.3.1 – Cross-platform correlation analysis	139
3.4.3.2 – Filter recommendations/summary	141
3.4.4 – Conclusions/chapter summary	143
<b>Chapter 4 – Molecular subgrouping of medulloblastoma via low-depth whole genome bisulfite sequencing</b>	<b>145</b>
<b>4.1 – Introduction</b>	<b>146</b>
<b>4.2 – Aims &amp; objectives</b>	<b>153</b>
<b>4.3 – Materials &amp; methods</b>	<b>154</b>
4.3.1 – Analysis Overview	155
4.3.2 – Combined cohort description	155
4.3.3 – Computing resources/software	157
4.3.4 – Analysing methylation variability	157
4.3.5 – Dimensionality reduction/visualisation	158
4.3.6 – Applying WGBS sample data to existing DNA methylation microarray classifiers to predict the subgroups of medulloblastoma	159
4.3.7 – Performance assessment of validated multi-class models trained using WGBS derived methylation data	159
<b>4.4 – Results &amp; discussion</b>	<b>162</b>
4.4.1 – CpGs of greater variability are present within the medulloblastoma methylome that are not assessed by the DNA methylation microarray platform	162
4.4.2 – WGBS derived variable CpGs are differentially distributed across chromosomes compared to the array platform	164
4.4.3 – Molecular subgroups of medulloblastoma are recapitulated when visualising cohorts sequenced via WGBS	165
4.4.4 – Medulloblastoma samples sequenced via low-pass WGBS can successfully be assigned into subgroup using pre-existing DNA methylation microarray classifiers	175
4.4.5 – Validated classifier models trained using low-pass WGBS perform equally to array trained models	180
4.4.5.1 – Elastic net	180
4.4.5.2 – Random forest	183
4.4.5.3 – Extreme gradient boosting	186
4.4.5.4 – Support vector machine	189
4.4.5.5 – Logic regression	192
4.4.5.6 – Model limitations	195

<b>4.5 – Summary &amp; conclusion</b>	<b>196</b>
<b>Chapter 5 – Generating diagnostic readouts for medulloblastoma via low-depth whole genome bisulfite sequencing</b>	<b>198</b>
<b>5.1 – Introduction</b>	<b>199</b>
<b>5.2 – Aims &amp; objectives</b>	<b>202</b>
<b>5.3 – Materials &amp; methods</b>	<b>203</b>
5.3.1 – Analysis Overview	203
5.3.2 – Cohort description	204
5.3.3 – Chromosomal copy number estimation	205
5.3.3.1 – Computing resources/software	205
5.3.3.2 – Establishment of an optimised reference-free CNV calling pipeline for use with low-pass WGBS sample data	206
5.3.3.3 – Validation of <i>MYC/MYCN</i> amplifications	208
5.3.3.4 – Copy number variation analysis of matched DNA methylation microarray cohort	210
5.3.4 – Variant calling of low-pass WGBS cohort	210
5.3.5 – Differential methylation analysis	211
5.3.5.1 – DMR detection of low-pass WGBS samples via metilene	211
5.3.5.2 – DMR detection of array data using DMRcate	212
5.3.5.2 – Subgroup-specific DMR annotation and gene ontology analysis	212
<b>5.4 – Results &amp; discussion</b>	<b>213</b>
5.4.1 – Application of an optimised reference-free CNVKit pipeline to low-pass WGBS samples successfully identifies large regions of aneuploidy with excellent concordance to current array methodologies	213
5.4.2 – Validated amplifications of <i>MYC</i> and <i>MYCN</i> are identified with improved sensitivity	217
5.4.3 – Analytical limitations of copy number analysis	221
5.4.4 – Reliable variant calling is beyond the capability of low-pass WGBS	222
5.4.5 – Greater numbers of differentially methylated regions are identified from WGBS data compared matched array	223
5.4.6 – Subgroup-specific differential methylation is occurring throughout the methylome that is missed by the array platform	226
5.4.7 – Ontology analysis of genes affected by differential methylation identifies greater numbers of enriched subgroup-specific biological processes in WGBS compared to DNA methylation microarray	230
5.4.8 – NMF-directed splitting of subtype VII is supported by differential methylation analysis	231
<b>5.5 – Summary &amp; conclusion</b>	<b>232</b>
<b>Chapter 6 – Summary &amp; Discussion</b>	<b>234</b>
<b>6.1 - Introduction</b>	<b>235</b>
<b>6.2 – High quality, non-missing methylation data can be generated for clinical samples sequenced via low-pass whole genome bisulfite sequencing that is highly correlated to matched array sample data</b>	<b>238</b>
<b>6.3 – Medulloblastoma sample data sequenced via low-pass WGBS can seamlessly be integrated into pre-existing array trained classification models</b>	<b>240</b>
<b>6.4 – CpGs of greater variability are present throughout the medulloblastoma methylome that contribute to increased visual separation of medulloblastoma subgroups</b>	<b>242</b>
<b>6.5 – Classifiers trained using low pass WGBS samples to predict the molecular subgroups of medulloblastoma perform equally to array-trained models</b>	<b>244</b>

<b>6.6 – Low-pass WGBS can be used to generate clinically relevant readouts in medulloblastoma</b>	<b>245</b>
<b>6.7 – Future work</b>	<b>247</b>
6.7.1 – Developing additional methods of imputation for use with WGBS that are less memory-intensive	247
6.7.2 – Establishment of user-friendly software for data handling of samples sequenced via low-pass WGBS	247
6.7.3 – Enabling WGBS sample compatibility with the DFKZ paediatric brain tumour classifier	248
6.7.4 – Novel ‘omics profiling of medulloblastoma subtypes via WGBS	248
<b>6.8 – Implications of Results</b>	<b>250</b>
<b>6.9 – Concluding remarks</b>	<b>250</b>
<b>7 – References</b>	<b>251</b>
<b>Chapter 8 – Appendix</b>	<b>289</b>
<b>8.1 – Cohort data</b>	<b>290</b>
8.1.1 – Sequencing quality control	290
8.1.1.1 – WGBS Sequencing summary statistics – ICGC Cohort	290
8.1.2 – Methylation quality control	292
8.1.2.1 – NMB Methylation QC summary via RnBeads2.0	292
8.1.2.2 – ICGC methylation QC summary via RnBeads 2.0	294
8.1.3 – Imputation	296
8.1.3.1 - BoostMe model performance – ICGC Cohort – Effect of downsampling on RMSE	296
8.1.3.2 - BoostMe model performance – NMB Cohort – Effect of downsampling on RMSE	298
8.1.3.3 – BoostMe model performance – ICGC cohort – Validation & test set performance for ICGC cohort (30x)	300
8.1.3.4 - BoostMe model performance – ICGC cohort – Validation & test set performance for ICGC cohort (10x – 3x filter)	302
8.1.4 – Applying WGBS sample data to pre-existing array trained classifier models to predict the molecular subgroups of medulloblastoma	304
8.1.4.1 – 4-group classifier results	304
8.1.4.2 – 7-group classifier results	305
8.1.5 – CpG variability analysis	307
8.1.5.1 – CpG variability (Array vs Liftover cohort)	307
8.1.5.2 – CpG variability (Top 100,000 CpGs) – WGBS vs Array	308
8.1.5.3 – Chromosomal distribution of top 10,000 most variable features (WGBS / Liftover / Array cohorts)	309
* denotes significantly different proportion compared to array – calculated using two-proportion z-test	310
8.1.6 – Copy number analysis	310
8.1.6.1 – Example conumee segmentation results – NMB_362	310
8.1.6.2 – Example CNVkit segmentation results – NMB_362	312
8.1.7 – Differential methylation analysis	314
8.1.7.1 – Gene ontology analysis – Subtype VII - VII_1 vs VII_2 DMRs (Array)	315
8.1.7.2 – Gene ontology analysis – Subtype VII - VII_1 vs VII_2 DMRs (WGBS)	316

## Abstract

**INTRODUCTION:** International consensus recognises four molecular subgroups of medulloblastoma, each with distinct molecular features and clinical outcomes. Assigning molecular subgroup is typically achieved via the Illumina DNA methylation microarray. Given the rapidly-expanding WGS capacity in healthcare institutions, there is an unmet need to develop platform-independent, sequence-based subgrouping assays.

Whole genome bisulfite sequencing (WGBS) enables the assessment of genome-wide methylation status at single-base resolution. To date, its routine application for subgroup assignment has been limited, due to high economic cost and sample input requirements and currently no optimised pipeline exists that is tailored for handling samples sequenced at low-pass (i.e., 1-10x depth).

**METHODOLOGY:** Two datasets were utilised; 36 newly-sequenced low-depth (10x) and 42 publicly available high-depth (30x) WGBS medulloblastoma and cerebellar samples, all with matched DNA methylation microarray data. We applied imputation to low-pass WGBS data, assessed inter-platform correlation and identified molecular subgroups by directly integrating WGBS sample data with pre-existing array-trained models. We developed machine learning WGBS-based classifiers and compared performance against microarray. We optimised reference-free aneuploidy detection with low-pass WGBS and assessed concordance with microarray-derived aneuploidy calls.

**RESULTS:** We optimised a pipeline for processing and analysis of low-pass WGBS data, suitable for routine molecular subgrouping and aneuploidy assessment. Using down-sampling, we showed that subgroup assignment remains robust at low depths and identified additional regions of differential methylation that are not assessed by methylation microarray. WGBS data can be integrated into existing array-trained models with high assignment probabilities, and WGBS-derived classifier performance measures exceeded microarray-derived classifiers.

**CONCLUSION:** We describe a platform-independent WGBS assay for molecular subgrouping of medulloblastoma. It performs equivalently to array-based methods at increasingly comparable cost (currently ~\$396 vs ~\$584) and provides proof-of-concept for routine clinical adoption using standard WGS technology. Finally, the full methylome enabled elucidation of additional biological heterogeneity that has hitherto been inaccessible.

**Word Count:** 300

---

## List of abbreviations

APC	Adenomatous polyposis coli
APOBEC2	Apolipoprotein B mRNA editing enzyme catalytic subunit 2
ATP	Adenosine triphosphate
AUPRC	Area under the precision-recall curve
AUROC	Area under the receiver operating characteristic curve
AXIN1	Axin-1
AXIN2	Axin-2
BCOR	BCL6 Interacting co-repressor
CBS	Circular binary segmentation
CBTRUS	Central Brain Tumour Registry of the United States
CDK4	Cyclin-dependent kinase 4
CDK6	Cyclin-dependent kinase 6
CHD	Chromo-domain, helicase, DNA-binding
CLA	Classic medulloblastoma
CNS	Central nervous system
CpG	Cytosine-guanine dinucleotide
CPR	Central pathological review
CSC	Cancer stem cell
CSF	Cerebrospinal fluid
CSI	Craniospinal irradiation
CT	Computerized tomography
CTDNEP1	CTD nuclear envelope phosphatase 1
CTNNB1	Beta-catenin
ddATP	Dideoxynucleotide adenosine triphosphate
ddCTP	Dideoxynucleotide cytosine triphosphate
ddGTP	Dideoxynucleotide guanine triphosphate
ddUTP	Dideoxynucleotide uridine triphosphate
DDX3X	DEAD-box helicase 3 X-linked
DFKZ	Deutsches Krebsforschungszentrum
DMR	Differentially methylated region
DN	Desmoplastic/nodular medulloblastoma
DNA	Deoxyribose nucleic acid
DNMT1	DNA methyltransferase 1
DP	Depth
E-CAD	E-cadherin
EM-seq	Enzymatic Methyl-seq
ENCODE	The Encyclopaedia of DNA Elements



---

FFPE	Formalin-fixed, paraffin embedded
FISH	Fluorescent, in-situ hybridisation
FOXR2	Forkhead box R2
GAB1	GRB2 associated binding protein 1
GATK	Genome Analysis Tool Kit
GEO	Gene Expression Omnibus
GERD	Gastroesophageal reflux disease
GLI1	GLI family zinc finger 1
GLI2	GLI family zinc finger 2
GLUT1	Glucose transporter 1
GQ	Genotype quality
Grp3	Group 3 medulloblastoma
Grp4	Group 4 medulloblastoma
Gy	Gray
H3	Histone H3
H3K4	Histone H3 lysine K4
HOXA5	Homeobox A5
HPC	High performance computing
hTERT	Telomerase reverse transcriptase (human)
IAP	Inhibitor of apoptosis
INO80	INO80 complex ATPase sub-unit
iSWI	Imitation SWI
KBTBD4	Kelch repeat and BTB domain containing 4
KMT2C	Lysine methyltransferase 2C
KMT2D	Lysine methyltransferase 2D
kNN	k-nearest neighbours
LCA	Large cell/anaplastic medulloblastoma
LINE-1	Long interspersed nuclear element-1
MAC	Methylation Array Classifier
MB	Medulloblastoma
MBD-Seq	Methyl-CpG binding domain-based capture and sequencing
MBD2	Methyl-CpG binding domain protein 2
MBEN	Medulloblastoma with extensive nodularity
Me-DIP	Methylated DNA immunoprecipitation sequencing
MLPA	Multiplex ligation-dependent probe amplification
MQ	Mapping quality
MRI	Magnetic resonance imaging
MYC	V-myc myelocytomatosis viral oncogene homolog
MYCN	Neuroblastoma-derived V-Myc avian

---

---

NCBI	National Center for Biotechnology Information
NCL	Newcastle
NGS	Next-generation sequencing
NHS	National Health Service (UK)
P53	Tumour suppressor P53
PBAT	Post-bisulfite adaptor tagging
PCA	Principal component analysis
PCR	Polymerase chain reaction
PD-1	Programmed cell death protein 1
PD-L1	Programmed death-ligand 1
PE150	Paired-end sequencing – 150 base pairs
PKS	Polyketide synthase
PRDM6	PR/SET domain 6
PTCH1	Patched 1
PVT1	PVT1 oncogene
qPCR	Quantitative polymerase chain reaction
RAM	Random access memory
Rb1	Retinoblastoma 1
RMSE	Root mean squared error
RNA	Ribonucleic acid
RRBS	Reduced representation bisulfite sequencing
SHH	Sonic hedgehog medulloblastoma
SMAC	Second mitochondria-derived activator of caspases
SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily A, member 4
SMARCB1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily B, member 1
SMO	Smoothed
SNCAIP	Synuclein alpha interacting protein
SNP	Single nucleotide polymorphism
SUFU	Suppressor of fused homologue (Drosophila)
SVM	Support vector machine
SWI/SNF	Switch/sucrose non-fermentable
t-sne	t-distributed stochastic neighbour embedding
TERT	Telomerase reverse transcriptase
TET2	Tet methylcytosine dioxygenase 2
TNF $\alpha$	Tumour necrosis factor alpha
TNFR1	Tumour necrosis factor receptor 1
TP53	Tumour protein p53

---

TSG	Tumour suppressor gene
UK	United Kingdom
UMAP	Uniform manifold approximation and projection
US	United States of America
UV	Ultraviolet
VEGF	Vascular endothelial growth factor
WGBS	Whole genome bisulfite sequencing
WHO	World Health Organisation
WNT	Wingless
XIAP	X-linked inhibitor of apoptosis protein
YAP1	Yes-associated protein 1
ZMYM3	Zinc finger MYM-Type containing 3

## List of figures

<b>Figure 1.1 – All cancers combined incidence rates in the UK for each gender by</b>	<b>21</b>
<b>Figure 1.2 – The stages of tumour progression</b>	<b>23</b>
<b>Figure 1.3 – The hallmarks of cancer</b>	<b>24</b>
<b>Figure 1.4 – The extrinsic and intrinsic pathways of apoptosis</b>	<b>26</b>
<b>Figure 1.5 – The invasion metastasis cascade</b>	<b>28</b>
<b>Figure 1.6 – Characteristics of cancer metabolism</b>	<b>31</b>
<b>Figure 1.7 – Models of sustained tumour growth</b>	<b>34</b>
<b>Figure 1.8 – Epigenetic mechanisms of expression control</b>	<b>37</b>
<b>Figure 1.9 – The conversion of cytosine to 5-methylcytosine</b>	<b>37</b>
<b>Figure 1.10 – Examples of epigenetic influence on tumorigenesis</b>	<b>42</b>
<b>Figure 1.11 – Distribution of CNS tumours in children by histological type</b>	<b>45</b>
<b>Figure 1.12 – Histological subtypes of medulloblastoma</b>	<b>49</b>
<b>Figure 1.13 – The molecular characteristics of the four-consensus primary molecular subgroups of medulloblastoma were first established in 2012</b>	<b>52</b>
<b>Figure 1.14 – How WNT canonical signalling is affected in medulloblastoma</b>	<b>54</b>
<b>Figure 1.15 – How canonical SHH signalling is affected in medulloblastoma</b>	<b>57</b>
<b>Figure 1.16 – Clinico-pathological features of the molecular subtypes of Groups 3 and 4 in medulloblastoma</b>	<b>60</b>
<b>Figure 1.17 – T1 weighted MRI of medulloblastoma</b>	<b>62</b>
<b>Figure 1.18 – ICH staining assay used to discern between WNT, SHH, Non-WNT/non-SHH medulloblastom.</b>	<b>63</b>
<b>Figure 1.19 – Difference between type I and type II probes on DNA methylation array chips</b>	<b>69</b>
<b>Figure 1.20 – Comparison between conventional WGBS and post-bisulfite adaptor tagging DNA library preparation approaches</b>	<b>71</b>
<b>Figure 2.1 - Bisulfite treatment of DNA produces up two different PCR products, resulting in the production of four different strands of DNA for each locus after amplification</b>	<b>92</b>
<b>Figure 2.2 - Bismark Alignment process</b>	<b>93</b>
<b>Figure 3.1 - Standard medulloblastoma subgroup &amp; patient risk stratification analytical pipeline.</b>	<b>104</b>
<b>Figure 3.2 – Graphical overview of all conducted analyses</b>	<b>104</b>

---

<b>Figure 3.3 - Effect of quality trimming on per-base sequence content distribution for NMB_318</b>	<b>110</b>
<b>Figure 3.4 - Observed methylation bias identified in read 2 of all samples at the final 2 bases of the 3' end.</b>	<b>112</b>
<b>Figure 3.5 – Overview of generated datasets</b>	<b>117</b>
<b>Figure 3.6 – Mean per base quality score distribution for all NMB samples</b>	<b>119</b>
<b>Figure 3.7 – Beta value density distribution plot for WGBS (ICGC = 30x / NMB = 10x) and Array methylation data</b>	<b>131</b>
<b>Figure 3.8 - Validation and test CpG set imputation model performance for all samples both high and low-pass WGBS data.</b>	<b>134</b>
<b>Figure 3.9 - Beta value distributions of non-missing and imputed CpG data for high and low-pass WGBS sample data.</b>	<b>136</b>
<b>Figure 3.10 – Mean test CpG set performance for downsampled high and low-pass WGBS data.</b>	<b>137</b>
<b>Figure 3.11 - Cross-platform Pearson correlation values for downsampled high and low-pass WGBS cohorts.</b>	<b>140</b>
<b>Figure 3.12 – Cross-platform correlation analysis for downsampled low-pass WGBS data.</b>	<b>142</b>
<b>Figure 4.1a – Demographic and molecular features of the current consensus molecular subgroups and subtypes of medulloblastoma</b>	<b>147</b>
<b>Figure 4.1b – Demographic and molecular features of the current consensus molecular subgroups and subtypes of medulloblastoma</b>	<b>147</b>
<b>Figure 4.2 - Current risk-adapted algorithm for medulloblastoma</b>	<b>150</b>
<b>Figure 4.3 – Graphical overview of all analyses conducted within</b>	<b>150</b>
<b>Figure 4.4 – Density plot of standard deviation value distribution of all features</b>	<b>162</b>
<b>Figure 4.5 – Chromosomal distribution of the top 10,000 most variable features derived from both WGBS and Array medulloblastoma cohorts</b>	<b>164</b>
<b>Figure 4.6 – Visualisation of the top 10,000 most variable features derived from our combined medulloblastoma DNA methylation microarray cohort</b>	<b>166</b>
<b>Figure 4.7 – Visualisation of the top 10,000 most variably methylated CpGs within the medulloblastoma WGBS cohort (n=77), subset to CpGs that map directly to probe sites located on the Illumina DNA Methylation 450k microarray.</b>	<b>167</b>
<b>Figure 4.8 – UMAP clustering of the top 10,000 most variable features as identified within our array (n =69), Liftover (n = 138) and WGBS (n = 69).</b>	<b>169</b>
<b>Figure 4.9 – UMAP clustering of the top 10,000 CpGs identified for the WGBS cohort, restricted to Group 3 and Group 4.</b>	<b>170</b>
<b>Figure 4.10 Identification of outlier sample in WGBS cohort</b>	<b>171</b>

---

---

<b>Figure 4.11 – Hierarchical clustering of the top 10,000 most variable features in our WGBS and Array medulloblastoma cohorts.</b>	<b>172</b>
<b>Figure 4.12 – t-sne: Overlaying WGBS samples onto MNP reference medulloblastoma samples (n = 426).</b>	<b>174</b>
<b>Figure 4.13. Mean elastic net classifier validation performance for Array, LiftOver and WGBS derived training sets</b>	<b>181</b>
<b>Figure 4.14. Mean elastic net classifier performance for array, liftover and each downsampled level of coverage</b>	<b>182</b>
<b>Figure 4.15. Mean random forest classifier validation performance for Array, LiftOver and WGBS derived training sets</b>	<b>184</b>
<b>Figure 4.16. Mean random forest classifier performance for array, liftover and each downsampled level of coverage.</b>	<b>185</b>
<b>Figure 4.17. Mean XGBoost (Linear booster) classifier validation performance for Array, LiftOver and WGBS derived training sets.</b>	<b>187</b>
<b>Figure 4.18. Mean XGBoost (Tree booster) classifier validation performance for Array, LiftOver and WGBS derived training sets.</b>	<b>188</b>
<b>Figure 4.19. Mean SVM (Linear kernel) classifier validation performance for Array, LiftOver and WGBS derived training sets</b>	<b>190</b>
<b>Figure 4.20. Mean XGBoost (Radial kernel) classifier validation performance for Array, LiftOver and WGBS derived training sets.</b>	<b>191</b>
<b>Figure 4.21. Mean logic regression classifier validation performance for Array, LiftOver and WGBS derived training sets</b>	<b>193</b>
<b>Figure 4.22. Mean logic regression classifier validation performance for Array, LiftOver and WGBS derived training sets</b>	<b>194</b>
<b>Figure 5.1 Graphical overview of all conducted analyses</b>	<b>209</b>
<b>Figure 5.2 Optimisation of CNVKit for low-pass WGBS</b>	<b>209</b>
<b>Figure 5.3 Presence/absence heatmap denoting the sensitivity of WGBS-derived whole/single arm aneuploidy calls in our low-pass NMB cohort (n = 36) compared to calls identified for matched array samples.</b>	<b>214</b>
<b>Figure 5.4 - Presence of a MYC amplification that is identified independent of the remaining gain in chromosome 8 by CNVKit</b>	<b>219</b>
<b>Figure 5.5 - Reference-free aneuploidy detection using CNVKit can be used to call focal events of aneuploidy that are missed when using methylation microarray</b>	<b>220</b>
<b>Figure 5.6 Depth frequency of passing variants identified by bs_call for NMB_169, NMB_181, NMB_436 and NMB_787.</b>	<b>222</b>
<b>Figure 5.7 Distribution of subgroup specific hyper/hypomethylated DMRs detected by DMRcate (Array) and metilene (WGBS) in WNT and SHH</b>	<b>228</b>
<b>Figure 5.8 Distribution of subgroup specific hyper/hypomethylated DMRs detected by DMRcate (Array) and metilene (WGBS) in Groups 3 and 4</b>	<b>229</b>

---

## List of tables

<b>Table 1.1 – Most common types of brain tumour</b>	<b>44</b>
<b>Table 2.1 - Clinical Demographics of NMB Cohort</b>	<b>81</b>
<b>Table 2.2 - Clinical Demographics of ICGC Cohort.</b>	<b>88</b>
<b>Table 3.1 – Coverage filters applied to each downsampled dataset</b>	<b>114</b>
<b>Table 3.2 – Sequencing statistics of NMB cohort</b>	<b>123</b>
<b>Table 3.3 – Methylation processing statistics for ICGC and NMB cohort.</b>	<b>125</b>
<b>Table 3.4 – Downsampled NMB cohort sequencing statistics</b>	<b>126</b>
<b>Table 3.5 – Methylation statistics for both ICGC and NMB cohorts</b>	<b>128</b>
<b>Table 3.7 – Methylation processing statistics for matched NMB methylation microarray cohort.</b>	<b>129</b>
<b>Table 4.1 – Demographics of combined medulloblastoma cohort</b>	<b>156</b>
<b>Table 4.2 – Adjusted t-sne and UMAP parameters used during visualization for each dataset</b>	<b>158</b>
<b>Table 4.3 – Chosen models for performance testing..</b>	<b>161</b>
<b>Table 4.4 – Number of top 10,000 most variable CpGs identified via WGBS that overlap with probes assessed by the Illumina methylation 450k and EPIC microarrays.</b>	<b>163</b>
<b>Table 4.5 – Classifier performance (4-group) for combined WGBS and array cohorts.</b>	<b>177</b>
<b>Table 4.6 – Classifier performance (7-group) for both WGBS and array cohorts.</b>	<b>177</b>
<b>Table 1. Gene validation data for NMB cohort used in aneuploidy analysis</b>	<b>204</b>
<b>Table 5.2 – NMF assignments of further subtypes present within subtype VII samples in our combined cohort</b>	<b>205</b>
<b>Table 5.3 – Summary of optimized CNVkit parameters for aneuploidy estimation of WGBS samples sequenced at low-pass</b>	<b>207</b>
<b>Table 5.4 – List of filter regions incorporated into CNVKit pipeline</b>	<b>208</b>
<b>Table 5.5 – List of samples in NMB cohort with known, validated MYC/MYCN amplifications</b>	<b>208</b>
<b>Table 5.6 – Number of segments called by WGBS and array CNV pipelines for matched NMB samples.</b>	<b>216</b>
<b>Table 5.7 – Total number of subgroup specific DMRs identified by DMRcate (Array)</b>	<b>216</b>
<b>Table 5.8 – Basic statistics of subgroup specific DMRs identified by DMRcate (Array) and metilene (WGBS)</b>	<b>224</b>

**Table 5.9 – Total number of subgroup specific DMRs identified by metilene (WGBS)** \_\_\_\_\_ **224**

**Table 5.10 – Total number of DMRs identified by metilene that overlap with at least one probe located on the Illumina Methylation450k or MethylationEPIC probe manifests.** \_\_\_\_\_ **224**

**Table 5.11 – Array subgroup-specific DMR distribution organized by annotation type** \_\_\_\_\_ **226**

**Table 5.12 – WGBS subgroup-specific DMR distribution organized by annotation type.** \_\_\_\_\_ **226**



# **Chapter 1 - Introduction**

## 1.1 – Cancer: Overview

Cancer is a term used to describe a collection of diseases that are typified by unregulated proliferation of cells and their spread from the site of origin to surrounding tissue or other sites of the body. There are over 200 different types and subtypes of cancer classified to date, where the tissue of origin dictates the cellular characteristics of the disease (National Health Service, 2020). Whilst the presence of a tumour (or tumours) is a key property of cancer, not all tumours are cancerous. Tumours can be benign, defined by an inability to spread to other locations (metastasis) from the site of origin and are generally not life-threatening, depending on their location. Cancerous tumours are classified as malignant, which show the capacity to spread into other tissues.

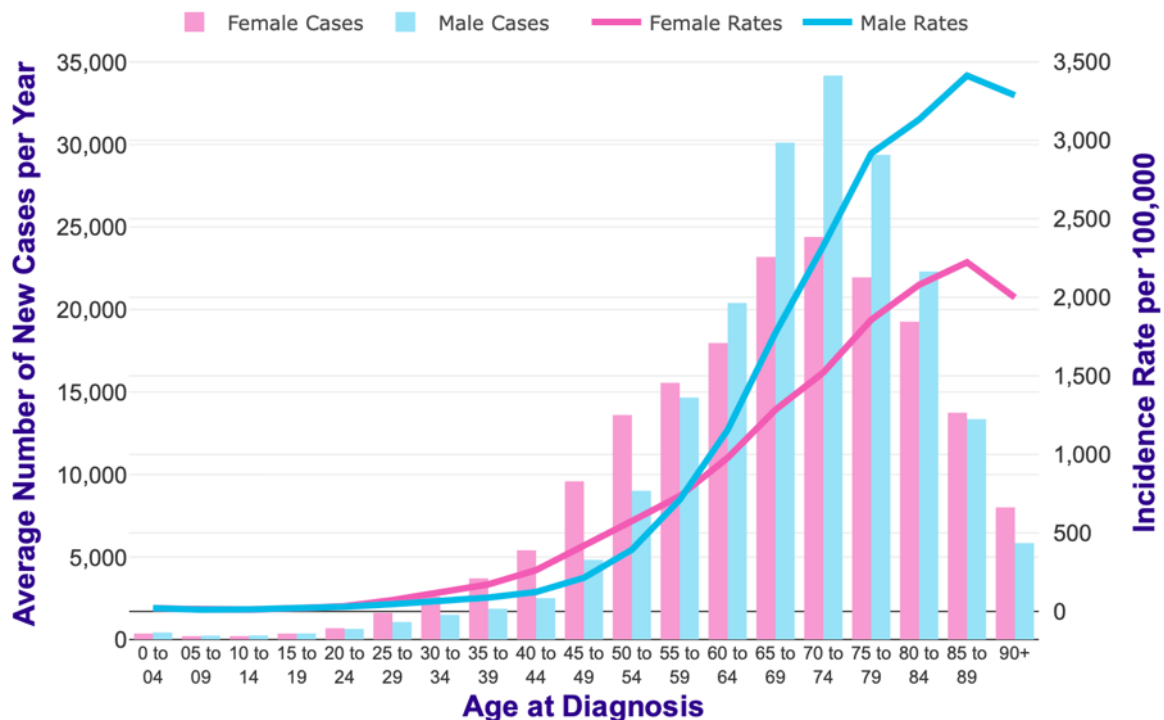
There are numerous different events that can trigger the beginning of cancerous growth in a cell. Cells can regulate their own rate of proliferation, and any cell which may become damaged or dysfunctional will apoptose under normal conditions. These processes are dictated by the normal expression of genes. Disruptions to the normal expression of these genes can result in the dysregulation of normal cell proliferation, causing it to become uncontrolled which subsequently results in tumour generation. Disruptions to gene expression is caused by mutation in the primary DNA sequence, leading to the classification of Cancer as primarily a genetic disease. The causes of these mutations can be both environmental and biological; typically, multiple mutations are acquired in tumours that can collectively contribute to tumorigenesis.

## 1.2 – Cancer: Epidemiology

Second only to heart disease, cancer is one of the world's leading causes of death globally. In 2020 it was estimated to be responsible for ~10 million deaths, of which 70% occur in middle-low-income countries (WHO, 2022). The most common cancers are of the lung and breast tissues, causing ~2.26 and 2.21 million cases in 2020 respectively. The deadliest cancer is also lung cancer, causing 1.8 million deaths globally (WHO, 2022). The economic burden of cancer is enormous, estimated to cost the US \$1.16 trillion in 2010, (Stewart & Wild, 2014), where both the prevalence of disease and cost of treatment has continued to increase as people age and are now less likely to die of other diseases.

In the UK between 1993 and 2018, the combined incidence rate of all cancers has increased across all age groups – 19% in 0-24, 22% in 25-49, 13% in 50-74 and 9% in 75+ age groups respectively (Cancer Research UK, 2022). There is no simple answer as to why this is occurring, but increased case numbers are largely due to the increasing size of the ageing population. Cancer incidence is strongly associated with age, with rates increasing significantly from ~55-59 years (Figure 1.1) (Cancer Research UK, 2022). Cancer is characterised by the accumulation of genetic mutations, which are more likely to occur over time, explaining the increased prevalence in these age groups. Whilst the

incidence of cancer is increasing, so too are survival rates. In the UK, cancer survival has doubled over the last 40 years, and half of all patients diagnosed will survive their disease for at least ten years (Cancer Research UK, 2021).



**Figure 1.1** – All cancers combined incidence rates in the UK for each gender (Male = blue, Female = Pink) by age group (Cancer Research UK, 2022).

### 1.3 – Drivers of tumorigenesis

Whilst cancer is a collection of similar diseases with significant amounts of individual genetic complexity, they can be grouped together based on several shared properties that arise as a direct consequence of mutations, resulting in tumorigenesis. The journey from a healthy cell to a cancerous one can be described histologically in four stages – hyperplasia, dysplasia, *in situ* cancer, invasive cancer (Figure 1.2). A mutated cell begins to excessively proliferate around the site of origin and is described as a region of hyperplasia. At some stage, the descendants of these cells will accumulate a further mutation that increases their proliferative capability even further. The collection of cells begins to look abnormal and less differentiated, being described as a region of dysplasia (pre-cancerous). Further mutations to these cells may result in even more dysplasia and loss of differentiation, at which stage it is described as cancerous *in situ*. If these cells begin to demonstrate a loss of adhesion and begin to spread, it is then described as a malignant tumour (invasive cancer).

In 2000, Hanahan & Weinberg published “The Hallmarks of Cancer”, which describes inherent cellular properties that all cancerous tumours share (Hanahan & Weinberg, 2000). As the mechanisms behind cancer have become better understood, the paper has been revisited in 2011 and 2022 to account for

advances in the field (Figure 1.3). These major properties are discussed in further detail in the following sections (Hanahan & Weinberg, 2011; Hanahan, 2022).

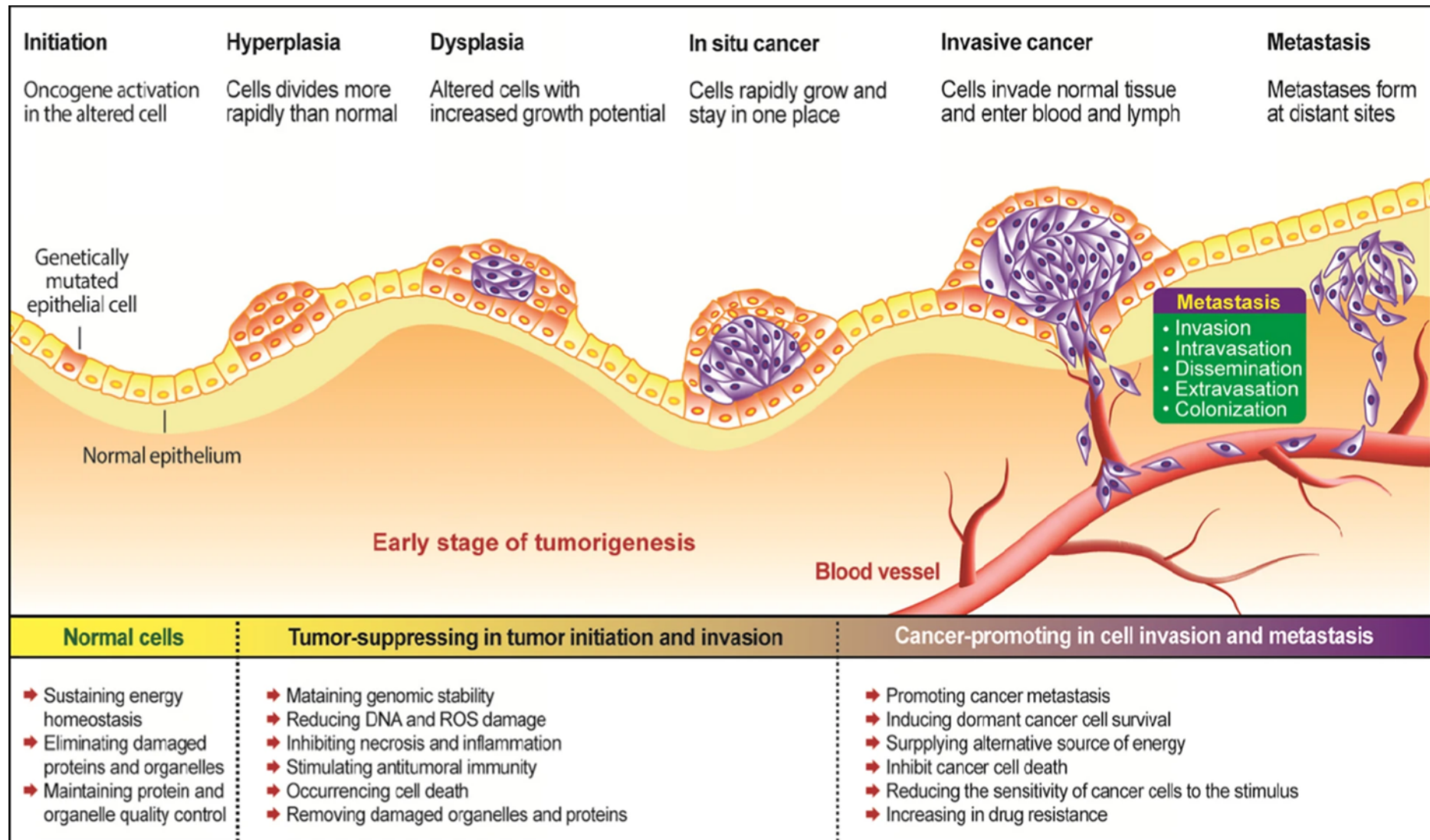
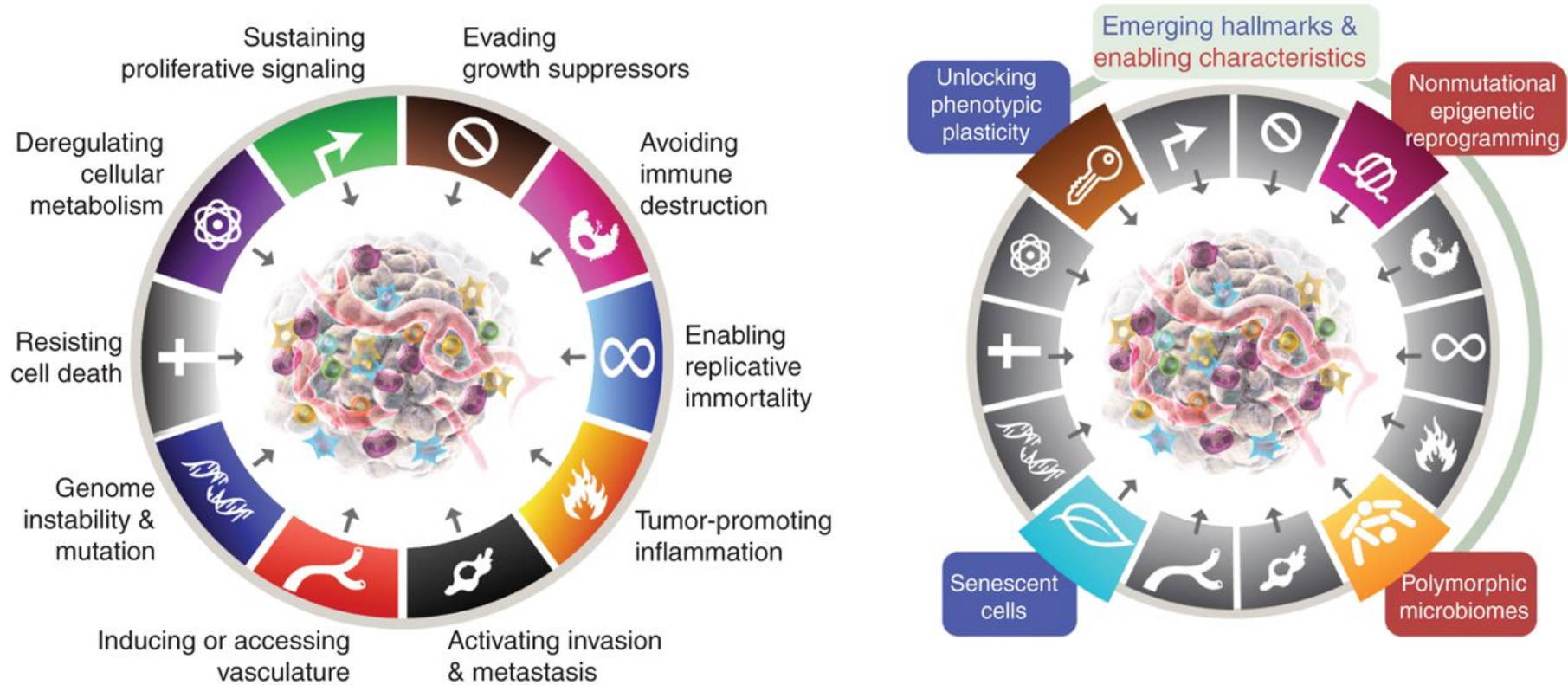


Figure 1.2 – The stages of tumour progression (Li et al., 2020)



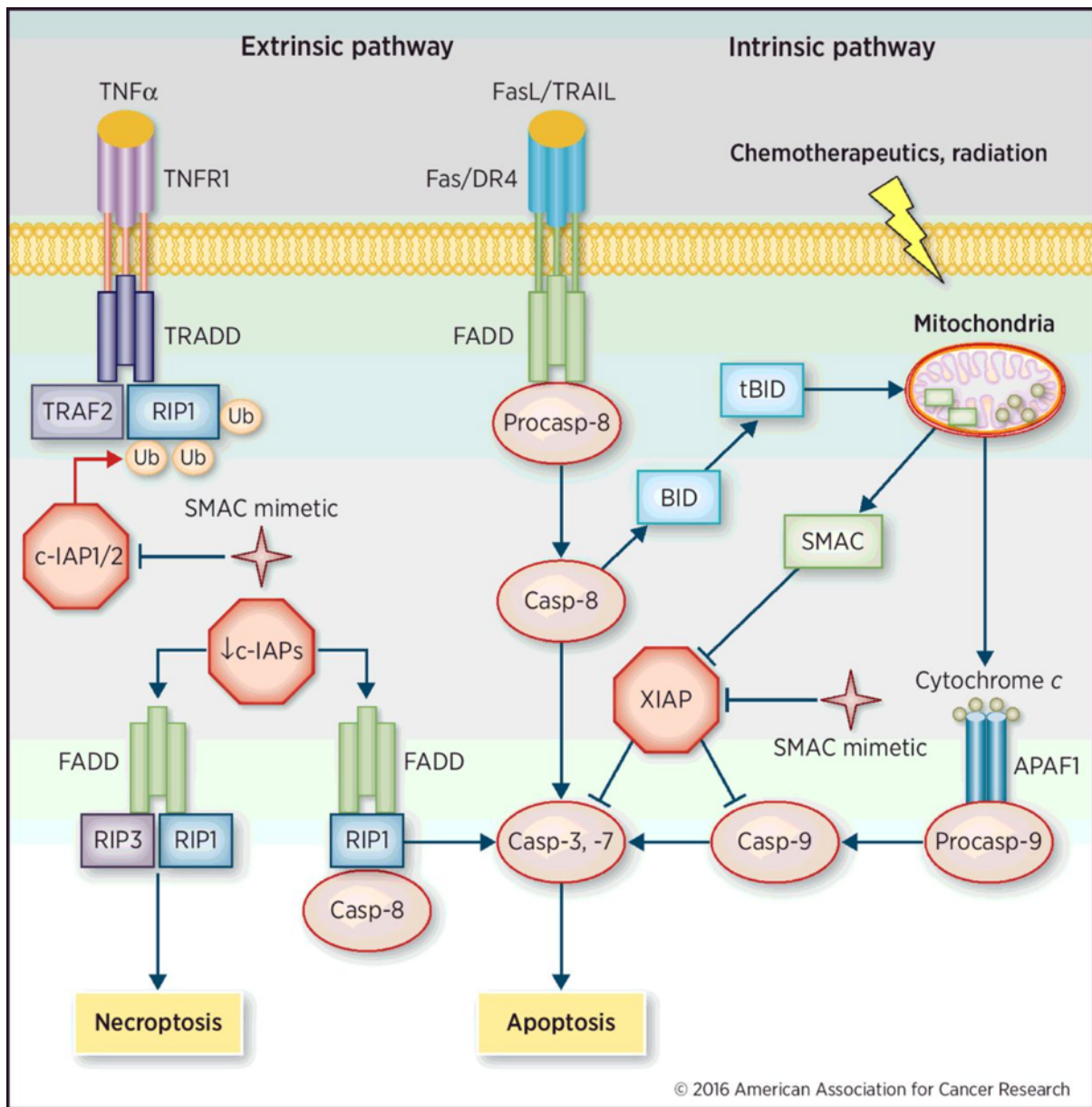
**Figure 1.3** – The hallmarks of cancer (Hanahan & Weinberg, 2022). Whilst the complexity underlying cancer is highly complex, the Hallmarks of Cancer provides a means to rationalise and organise both key characteristics and enabling traits of the disease. The traits are regularly updated, with the previous iteration (left) recently updated in 2022 (Hanahan & Weinberg) to account for four additional emerging hallmarks (right).

### 1.3.1– Resisting apoptosis

One of the major phenotypic traits tumour cells must acquire is the ability to evade the innate protective mechanism of apoptosis. Apoptosis is an intrinsic property of multicellular organisms that initiate programmed cell death in cells that begin to show signs of aberrant proliferation (Evan & Littlewood, 1998). The process acts to protect the whole organism over the survival of the individual cell by the introduction of various suppressive mechanisms in the cell replication process. Apoptosis is therefore a vital process that protects the cell from tumorigenesis. One example of apoptosis is the peeling of skin following sunburn to protect from the damaging effects of UV exposure. Apoptosis can be triggered from both the intrinsic and extrinsic environments of the cell, with both pathways converging at the activation of a caspase cascade that results in proteolytic cleavage of the cell (Figure 1.4). The cell effectively “digests” itself, and the residual material is then consumed by neighbouring cells and macrophages via phagocytosis (Hanahan & Weinberg, 2011).

For a cancer cell to gain the ability evade apoptosis, the acquisition of a mutation of key genes involved in mediating the pathway is necessary. Mutations to the intrinsic pathway, responsible for regulating internal proliferative signalling, is most widely implicated in enabling cancer tumorigenesis. A well-known component of apoptosis is the *p53* protein, which induces apoptosis in response to DNA damage. Mutations to the *p53* gene, *TP53* is one of the most common mutations observed in tumour cells that enables apoptosis evasion (Hanahan & Weinberg, 2011). Loss of *p53* enables the cell to accumulate further unchecked DNA damage, enabling cells to become further mutated, likely to develop further tumorigenic properties, and more likely to become cancerous.





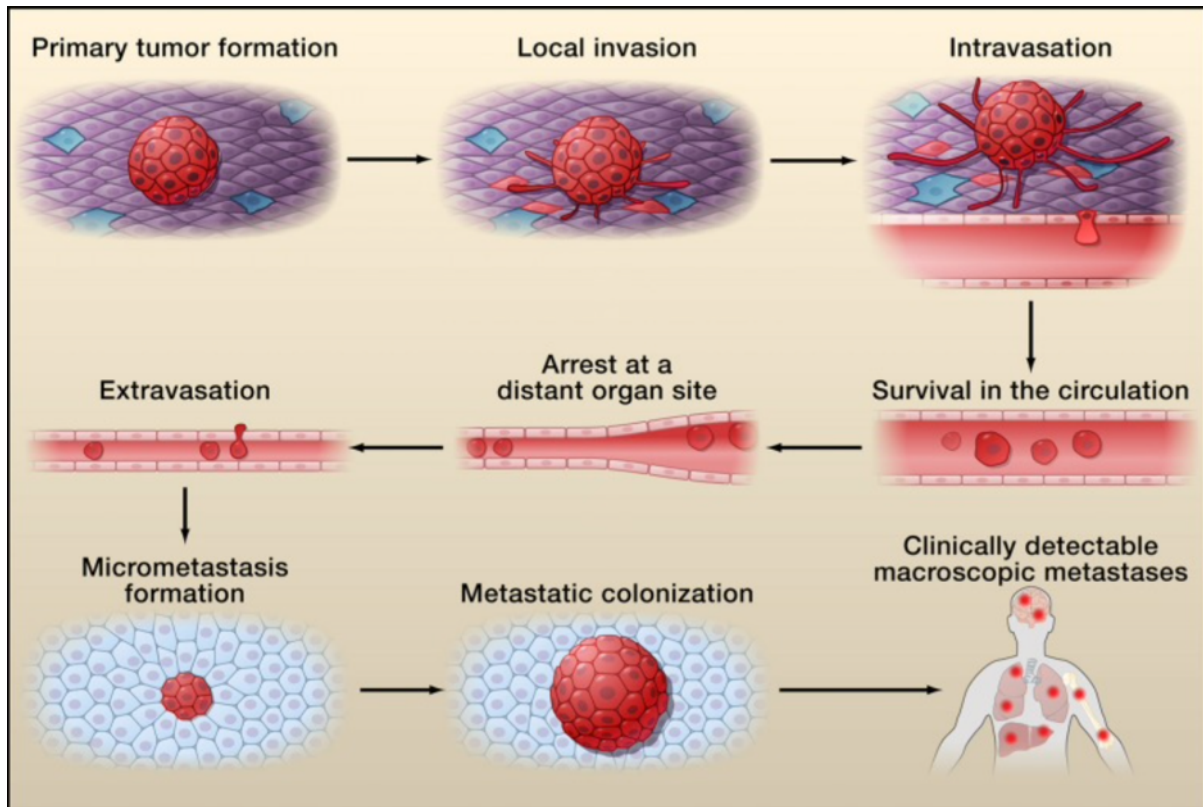
**Figure 1.4** – The extrinsic and intrinsic pathways of apoptosis (Derakhshan et al., 2017). Extrinsic activation of death receptors such as *TNFR1* by its associated ligand *TNFα* causes trimerisation of adaptor proteins and ultimately activation of caspase 8, ultimately resulting in apoptosis. Intrinsic pathways of apoptosis are initiated via cytogenetic insults like radiation or chemotherapy, causing the release of cytochrome c and *SMAC* into cytosol – resulting in the binding and degradation of inhibitors of apoptosis (IAPs) such as *XIAP*, causing apoptosis.



### 1.3.2 - Invasion & metastasis

Tumour cells that reach a higher state of malignancy will often gain the capability to disassociate from the original site of growth and invade nearby tissues or infiltrate the circulatory system and spread to distant sites around the body. This process of cell invasion is most often known as metastasis. Cancers that metastasise are more likely to cause death than non-metastatic tumours (Chaffer & Weinberg, 2011). Acquiring this trait is what separates benign and malignant tumour growths. The spread of cancer cells throughout the body presents a significant barrier to survival, by interfering with the healthy function of multiple organs separate from the primary tumour site. Whilst the precise genetic intricacies of metastasis are not fully defined, the process can be summarised in a sequence of steps termed the invasion-metastasis cascade (Valastyan & Weinberg, 2011) (Figure 1.5).

After primary tumour formation, a single cell may gain the capacity to invade the local surrounding extracellular matrix. Cells may then invade into the blood vessel lumen in a process termed intravasation followed by transportation into the circulatory system. If the cells have adapted significantly to survive in the circulation, they may then settle at a distant tissue site, where they can extravasate into a new tissue microenvironment and begin to form a new metastatic growth (colonisation). The number of genetic changes a tumour cell must undergo to achieve successful metastasis is significant. To migrate successfully, cells must be able to slow down their own proliferation, avoid apoptosis and simultaneously down-regulate cell attachment and up-regulate cell adhesion processes to aid movement, to name but a few.



**Figure 1.5** – The invasion metastasis cascade (Valastyan & Weinberg, 2011).

Individual tumour cells leave primary sites of growth, translocating systemically. If such cells gain the capability to survive in circulation and arrest at new organ sites, additional growths may arise in tissues distant from the primary tumour site (metastasis).

### 1.3.3 - Angiogenesis

Like healthy cells, cancer cells require nutrition to survive, as well as the capability to remove waste products. To facilitate this, the formation of new blood vessels in and around the tumour is required in a process called angiogenesis. Angiogenesis is common during embryogenesis but rarely occurs during adult life, where it is only activated transiently during wound healing and the female reproductive cycle (Hanahan & Weinberg, 2011).

Angiogenesis is regulated by various inducers and inhibitors, and the process is ‘switched on’ depending on the balance of these molecules in the cellular microenvironment. Angiogenic induction is largely governed by growth factors such as the vascular endothelial growth factor family of proteins (*VEGF*). Inhibition of angiogenesis typically prevents vascular generation indirectly by inhibiting cell migration and proliferation. For this reason, *p53* is also an indirect inhibitor of angiogenesis, as well as apoptosis (Teodoro, Evans & Green, 2007).

Whilst angiogenesis has long been appreciated as a hallmark of cancer development, recent evidence suggests that it is not always required for a tumour to thrive and spread. An increasing number of

cases have been reported where tumour tissues simply exploit existing vascular networks and grow alongside these networks in a process termed ‘vascular hijacking’ (Winkler, 2017). Nevertheless, there are numerous angiogenesis inhibitors available for treatment which show efficacy in treating some cancers (El-Kenawi & El-Remessy, 2013). For example, Axitinib (A *VEGF* receptor tyrosine kinase inhibitor) when used in combination with pembrolizumab resulted in significantly increased overall survival in patients with renal cell carcinoma (Rini *et al.*, 2019).

### **1.3.4 - Genome instability & mutation**

Acquiring the numerous phenotypic traits required to sustain cancerous growth requires several mutations. Individually, the rates of spontaneous mutations of specific genes are low and gaining all the mutations required to acquire all tumorigenic phenotypes even lower still. Yet, the fact that all cancers share these traits suggest that cancer cells can and do acquire tumorigenic phenotypes, despite these low probabilities.

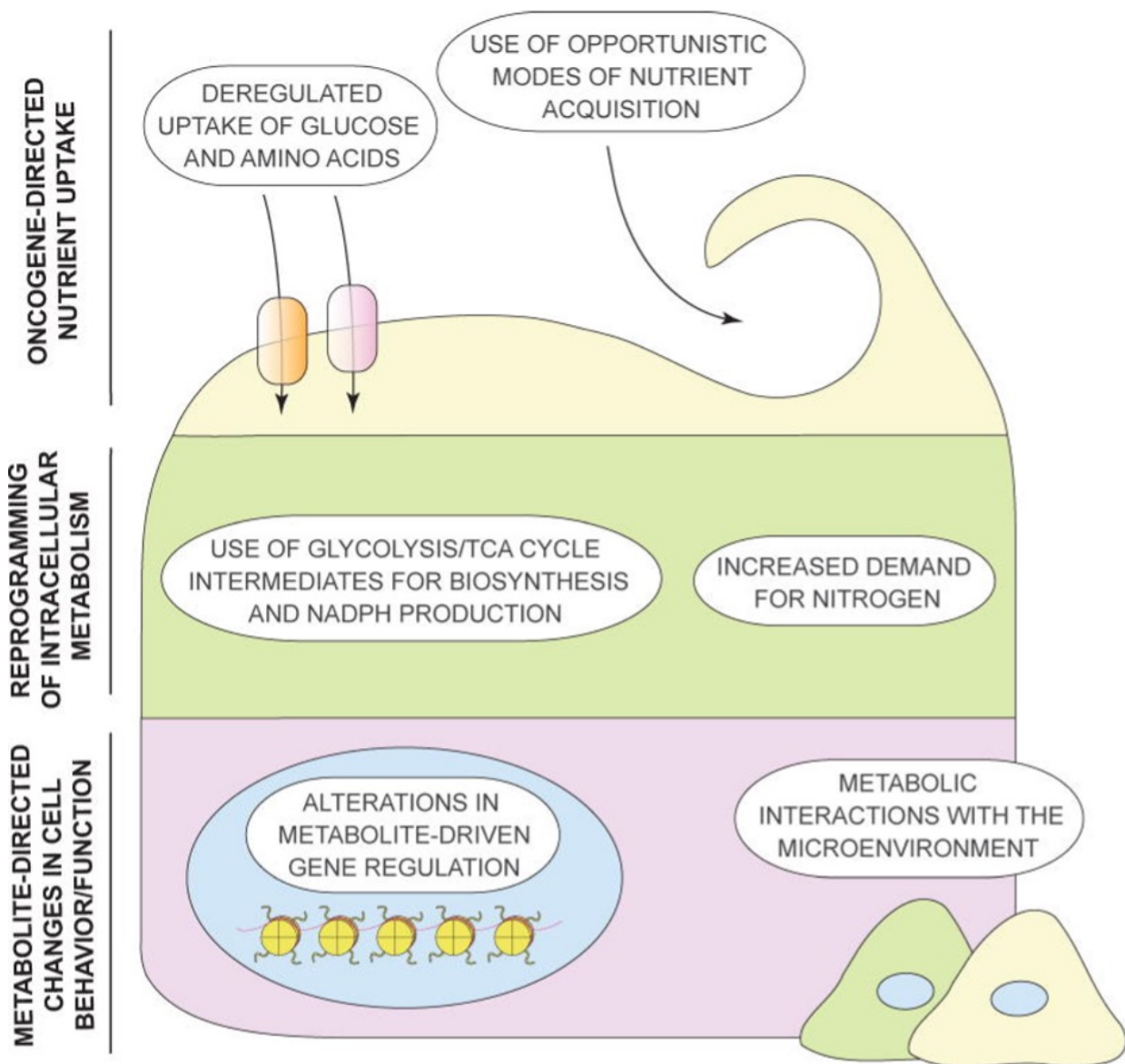
One of the main causes of early genomic instability is ‘Replication Stress’ (Zeman & Cimprich, 2014). DNA replication stress is caused by the loss of key pathways involved in maintaining normal DNA replication. DNA replication is initiated at thousands of origin sites in the genome, which form bi-directional replication “forks”. DNA replication from these sites is regulated by the cell, splitting the origin sites into ‘early-replicating’ and ‘late-replicating’ sites. There are numerous ‘back-up’ replication sites to ensure smooth DNA replication if other sites stall or become damaged (Zeman & Cimprich, 2014). Stalled or damaged forks place pressure upon DNA replication machinery at the sites that remain, increasing the rate of mutation during the replication process.

Cancerous cells display increased rates of mutation, causing significant genetic damage to cell machinery involved in maintaining genetic integrity. Ultimately this results in significant genetic instability, which is now considered as an enabling hallmark characteristic of cancer (Negrini, Gorgoulis & Halazonetis, 2010; Hanahan & Weinberg, 2011).

### 1.3.5 - Deregulation of metabolism

Alongside uncontrolled proliferation, another common trait of the cancer cell is an altered metabolism. Cell proliferation is an energy intensive process and as a result, tumour sites are typically nutrient-poor and there are often areas that are hypoxic. Cancer cells attempt to adapt to these environments by acquiring mutations that enable unconventional methods of metabolism (Pavlova & Thompson, 2016) (Figure 1.6). A long-observed metabolic alteration in cancer cell metabolism is that of glucose production. To meet the increased demands for glucose, cellular glucose transporters such as *GLUT1* are often mutated, causing increased uptake and subsequent utilisation (Hsu & Sabatini, 2008). Cancer cells have also been known to utilise glycolysis under aerobic conditions, in a process known as the Warburg effect (Warburg, Wind & Negelein, 1927). Switching to an anaerobic mechanism of respiration, even when in aerobic conditions seems contradictory. This alteration may be explained by its association with oncogenic activation and tumour suppressor de-activation (Jones & Thompson, 2009).

Analysis of some tumours discovered the presence of two subpopulations, varying in their metabolism. The aforementioned “Warburg-effect” population that rely on glycolysis is supplemented by an additional population that utilise the large numbers of lactate molecules produced as an energy source by utilising the citric acid cycle (Kennedy & Dewhirst, 2010). Whilst this phenomenon has not been generalised, it serves as an important example of how genomic mutations cause tumour cell populations to adjust their metabolic capabilities to thrive under highly variable environmental conditions.



**Figure 1.6** – Characteristics of cancer metabolism (Pavlova & Thompson, 2016). Cancer cells can undergo a multitude of metabolic changes that enable them to gain access to both unconventional sources of nutrients as well as increasing uptake of nutrients from conventional sources to help sustain the energy requirements for persistent cellular replication.

### 1.3.6 - Immune system evasion

Like apoptosis, the immune system acts as a protective mechanism for the body against tumorigenesis by surveying the body for abnormal cells and eliminating them. Whilst an effective method of preventing tumorigenesis, mutations gained by cancer cells can result in the successful evasion of immune surveillance. There are numerous known mechanisms by which cancer cells can escape the immune system. Cancer cells can cease the expression of tumour antigens and/or down-regulate the expression of antigen presenting molecules, rendering the T-cell response effectively useless (Monjazeb *et al.*, 2013). Other methods include the presentation of 'healthy' antigens, effectively tricking the immune system into thinking the tumour isn't a threat. A prominent example of immune suppression by cancer cells is through the expression of PD-L1, a ligand of PD-1 which is expressed on the T-cell surface. Binding of these molecules to one another results in the suppression of T-cell growth, which has a knock-on effect of slowing T-cell activation (Sun, Mezzadra & Schumacher, 2018).

Once a tumour cell gains the ability to evade the immune system, this sub-population is effectively favoured via natural selection. The immune system is capable of destroying normal tumour cells, but the resistant sub-population survives, resulting in subsequent tumour growths conferring this resistance. The heterogeneity of cancers only increases the likelihood of such sub-populations arising and makes it inherently difficult to develop therapeutic measures to counteract this (Vinay *et al.* 2015).

### 1.3.7 - Inflammation

Aside from successful evasion of the immune system, some tumours are successfully infiltrated by immune cells, beginning the immune response against cancer cells within. Basic understanding of the immune response suggests that the body making efforts to eradicate the tumour through intrinsic means is positive. It has been shown that an innate immune cell presence of at least some degree is detectable in most tumours, bringing an associated inflammatory response along with it (Pagès *et al.*, 2010). Chronic inflammation can also play a role in initiating tumour development in tumorigenesis. Inflammation of the oesophagus caused by acid reflux (GERD) can lead to Barrett's oesophagus and ultimately oesophageal adenocarcinoma in some patients (Bhat *et al.*, 2011).

The inflammatory environment has been shown over the past decades to enable the development of many tumorigenic traits that promote tumour growth and progression. The sustained production of growth factors and reactive oxygen species can signal tumour cells to proliferate further as well as induce further DNA mutations (Todoric, Antonucci & Karin, 2016). It has also been shown to stimulate angiogenesis and metastatic progression (Grivennikov, Greten & Karin, 2010). Consequentially, a sustained inflammatory response can induce an evolutionary effect on tumour development, promoting the proliferation of mutated cancer cells that are able to thrive in the microenvironments that are intended to destroy them.

### 1.3.8 - Sustained proliferation

The most infamous characteristic of a tumour is the capability of the cells therein to sustain proliferation. Under normal conditions, proliferation is a tightly controlled process. The cell cycle is regulated by an array of growth promoters and inhibitors at each stage. This controls the number of cells that undergo proliferation at any one time, ensuring that healthy tissue function and structure is maintained. Cancer cells overcome these mechanisms of control through the altered expression of these molecules. Methods include the over-expression of cyclin dependent kinases (CDKs), molecules directly involved in cell cycle progression. Alternatively, cancer cells may instead signal surrounding healthy cells to secrete additional growth factors which are detected by cell surface receptors to initiate the cell cycle (Hanahan & Weinberg, 2011).

Increasing a cell's susceptibility to promotion of proliferation must also be supplemented by mutation or inactivation of genes that encode for molecules that negatively regulate the process. Tumour suppressor genes (TSGs) act as gatekeepers at various stages of the cell cycle. Prominent examples include *TP53*, and retinoblastoma associated (*Rb*) proteins (Section 1.5.1) (Hanahan & Weinberg, 2011).

### 1.3.9 - Cell immortality

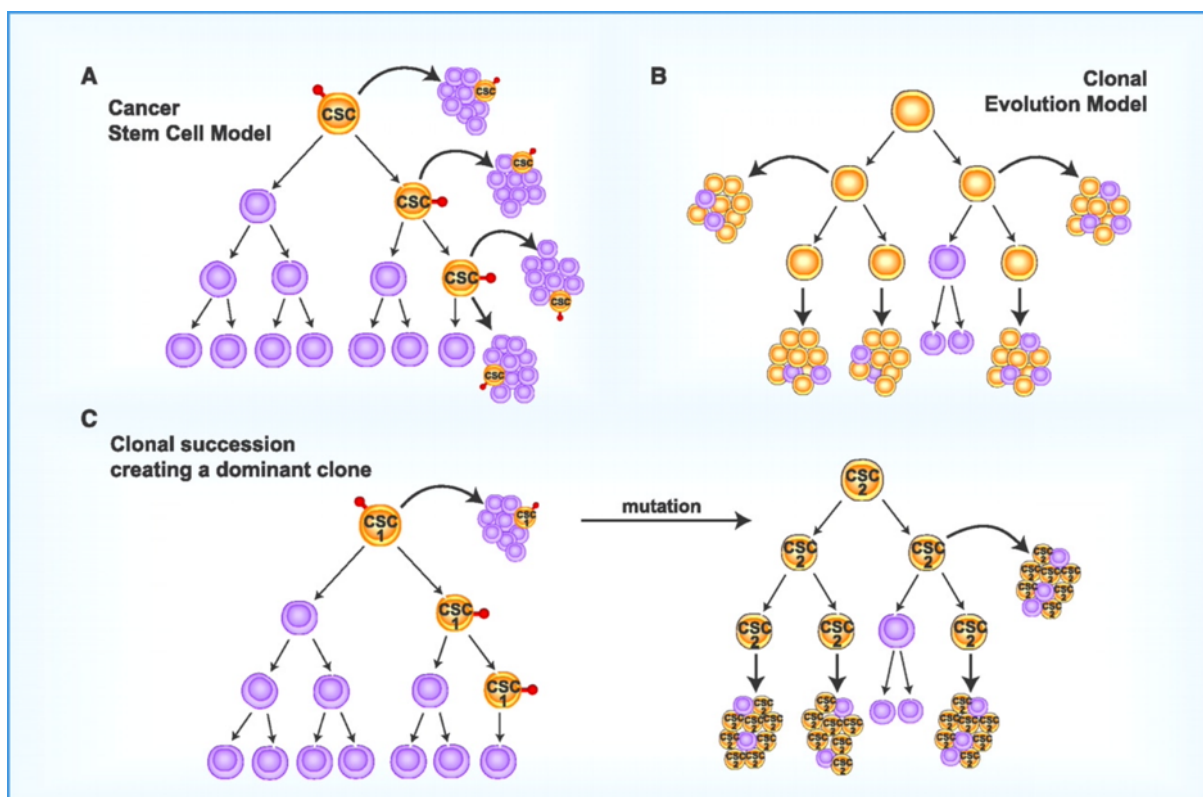
Avoiding apoptosis and maintaining proliferation are a potent tumorigenic combination but are limited by cellular senescence. Senescence describes the cellular ageing process. Healthy cells are limited by the number of times that they can successfully divide. Eventually, normal cells enter a state where they are unable to proliferate yet remain viable (senescent). Cells that continue to divide after becoming senescent enter crisis, a stage of mass cell death. Both crisis and senescence act as important barriers to cancerous growth in cells, at the price of inevitable death due to ageing (Hanahan & Weinberg, 2011).

In rare cases, cells evade crisis and emerge immortalised, enabling them to replicate indefinitely. These cells achieve this by maintaining the length of their telomeres, a repetitive nucleotide sequence located at the ends of each chromosome. Under normal cellular conditions, telomeres progressively degrade over successive cell divisions, initiating senescence once fully eroded. Cancer cells can activate a normally silent gene, *hTERT*, that encodes for the catalytic component of the telomerase complex. Telomerase is capable of lengthening telomeres, essentially bypassing senescence, and its expression is detected in ~90% of cancers (Jafri *et al.*, 2016). Without becoming immortalised, it is difficult for tumours to maintain sustained growth, making this trait an important barrier that must be overcome by cancer cells to fully harness their replicative potential (Hanahan & Weinberg, 2011).



### 1.3.10 – Cancer stem cells

Stem cells are capable of continued self-renewal as well as differentiation into specialised cell types. Evidence of a sub-population of cancerous stem cells in tumours was first reported in the 1990s in acute myeloid leukaemia (Lapidot *et al.*, 1994). Since then, stem-cell sub-populations have been identified in numerous solid tumours, including glioblastoma and breast cancers (Singh *et al.*, 2003, Al-Hajj *et al.*, 2003). It is hypothesised that for some tumours, cancer stem cells (CSC's) are the cell of origin and are responsible for resistance to chemotherapeutic agents and metastases distant from the primary site (Dawood, Austin & Cristofanilli, 2014) (Figure 1.7). CSCs diversify the cell-type population through asymmetric cell division. This ultimately results in increased tumour heterogeneity and can result in tumours becoming resistant to treatment and present higher risk of metastasis and/or relapse.



**Figure 1.7** – Models of sustained tumour growth (Adams & Strasser, 2008)

**A)**- CSCs (gold) are often relied upon for tumour growth. In this model, tumour growth and differentiation is linked directly to that of changes within the CSC.

**B)** – Clonal evolution models assume that most tumour cells contribute to tumour self-renewal, Differentiation within sub-populations of cells contribute to significant tumour heterogeneity, where different epigenetic, cytogenetic and mutation patterns may be observed throughout the tumour.

**C)** - A mixture of both A & B, a tumour is initially driven by cells with stem like properties. Mutations in one cell produces a dominant phenotype, resulting in enhanced self-renewal and differentiation throughout the tumour



### 1.3.11 – Enabling Characteristics

In 2022, Hanahan & Weinberg presented several ‘enabling characteristics’ that represent our continued progress in understanding cellular traits that contribute to tumorigenesis – phenotypic plasticity, epigenetic reprogramming, polymorphic microbiomes and senescent cells (Figure 1.3).

Phenotypic plasticity refers to the capacity of cancerous cells to be able to reverse or evade the process of differentiation, which typically results in cells that no longer proliferate. Evading this process is now thought to be a key trait of tumorigenesis, where cancerous cells may fully de-differentiate back to a stem-like state or transdifferentiate from one differentiation pathway to another, resulting in tumour cells acquiring wholly new traits (Yuan *et al.*, 2019). Phenotypic plasticity is most easily demonstrated in colon cancers, where it is necessary that cancer cells escape the process of cell exfoliation and permanent renewal. Transcription factors such as *HoxA5* are expressed during differentiation of epithelial cells in the colon, and are often lost in colonic cancer cells, resulting in cells reverting to a more dedifferentiated state (Ordonez-Moran *et al.*, 2015).

Non-mutational epigenetic reprogramming recognises that epigenetic modifications (changes to gene expression independent of changes to the underlying genetic sequence) plays a significant role in contributing to tumorigenesis, independent of mutation or genomic instability. Epigenetic regulation of gene expression is a long-recognised mechanism underlying cell differentiation and development; and induced changes to the epigenetic landscape throughout the genome can contribute to aberrant expression and ultimately the acquisition of other hallmark traits of cancer (Hanahan & Weinberg, 2022). Epigenetic reprogramming may occur in numerous ways. Microenvironmental changes such as increased hypoxia can cause substantive hypermethylation, resulting in expression loss of key tumour suppressor genes (Thienpont *et al.*, 2016). Epigenetic reprogramming can be heterogenous, altering expression profiles of cells differently, causing increased phenotypic diversity throughout a tumour and increasing the chance of cancer cells populations acquiring other hallmark traits (Lu *et al.*, 2020).

The microbiota – a term used to describe the microorganisms that reside within bodily tissues, is increasingly being recognised to have a significant effect on human health (Thomas *et al.*, 2017). The microbiota varies substantially between individuals, and its characteristics can subsequently impact phenotypic traits of cancer (Dsutzev *et al.*, 2017). A prominent example is that of the gut microbiome, where *E. coli* cells that carry a *PKS* locus have been shown to have mutagenic effects on the human DNA and is linked to causing “hallmark-enabling mutations” (Pleguezuelos-Manzano *et al.*, 2020). Faecal transplants – intended to revert the microbiome back to a state associated with healthy individuals – have been shown to restore patient response to immunotherapy in melanoma (Baruch *et al.*, 2021). There is a growing appreciation of the role that the microbiome plays in enabling and inhibiting tumorigenesis and although its exact mechanisms are only just beginning to be understood, it offers significant potential to increase patient survival.

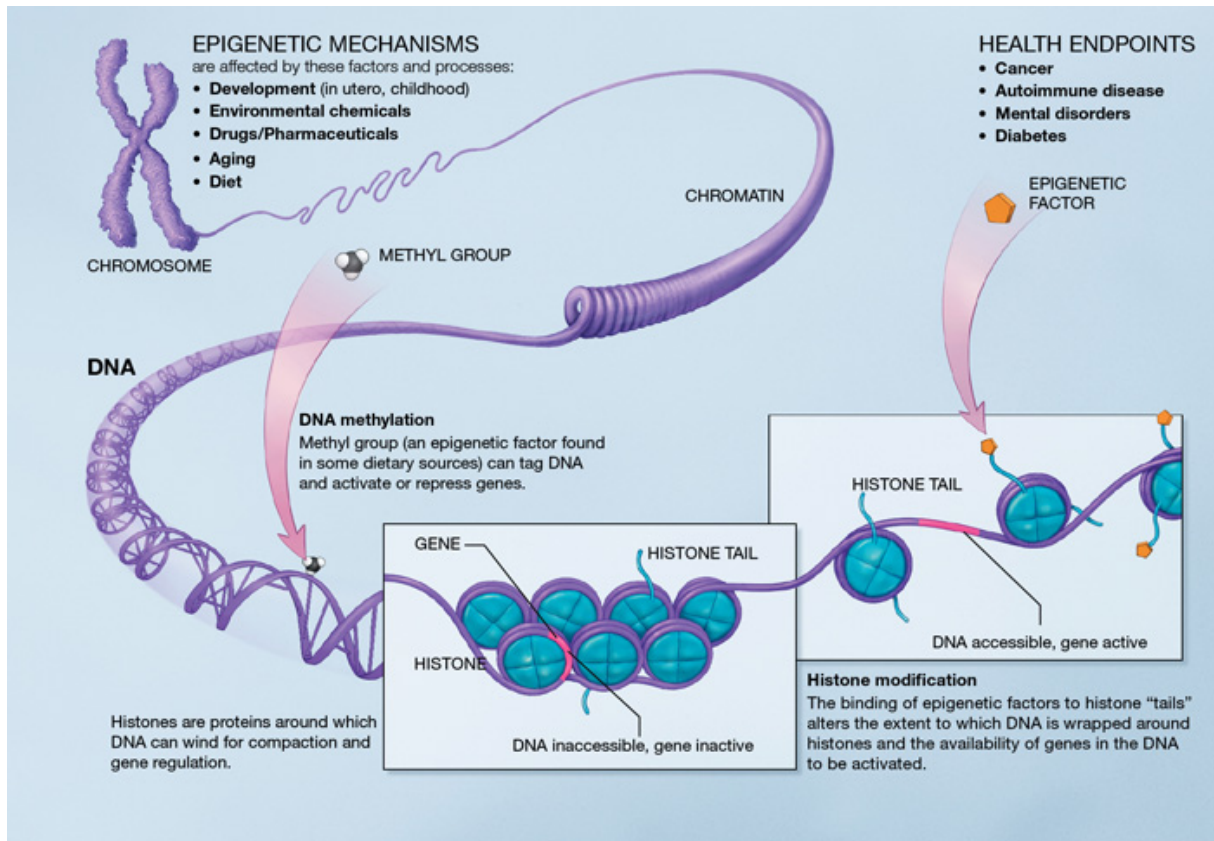
## 1.4 – Oncogenes & tumour suppressor genes

Any gene that has the potential to cause tumorigenic traits if mutated is termed as an oncogene. Expression of oncogenes are essential for the initiation and maintenance of tumorigenesis. Wild-type genes that become oncogenic once mutated are termed proto-oncogenes and typically consist of genes directly involved in cell proliferation. Oncogenes contribute to tumorigenesis in cancer and can be activated by several different mechanisms. Oncogenes can be structurally altered via chromosomal rearrangements, amplifications or gene fusions or can undergo epigenetic modification that results in aberrant expression (Botezatu *et al.*, 2016).

Conversely, any gene that prevents tumorigenesis in the cell is named a tumour suppressor gene (TSG). Whereas oncogenes need to be activated or over-expressed, TSGs must be inactivated or under-expressed in the cancer cell to encourage tumorigenesis. TSGs are recessive in nature and require mutations to both genes on the chromosome to be inactivated. This ‘two-hit hypothesis’ was first described back in the 1970’s with the identification of one of the first TSG, *Rb1* in retinoblastoma (Knudson Jr., 1971).

## 1.5 – Epigenetics

The word epigenetics means ‘upon’ or ‘above’ genetics and is scientifically defined as any heritable alteration to gene expression that is not caused by direct changes to the primary DNA sequence. Epigenetic mechanisms cause changes to the structure and conformation of chromatin, therefore having a regulatory effect on gene transcription. Stable epigenetic modifications in cells are crucial to development and maintaining the desired patterns of gene expression. For example, muscle cells remain as such due to epigenetic silencing of genes that do not promote a ‘muscle cell-like’ phenotype. Dysregulated epigenetic mechanisms can result in aberrant gene expression and subsequently lead to or aid in the development of cancer tumorigenesis (Sharma, Kelly & Jones, 2010). The types of epigenetic mechanisms can be grouped into four classes – DNA modification, histone modification, chromatin remodelling, and non-coding RNA mechanisms – and are discussed in detail below (Figure 1.8).



**Figure 1.8** – Epigenetic mechanisms of expression control (National Institute of Health, 2021). Epigenetic control of expression involves multiple different chemical compounds that bind to DNA or proteins that increase or decrease the accessibility of that region of DNA to enzymes involved in DNA transcription and ultimately expression. Epigenetics play a crucial role in cellular development, where epigenetic signatures help guide stem cell differentiation.

### 1.5.1 - DNA methylation

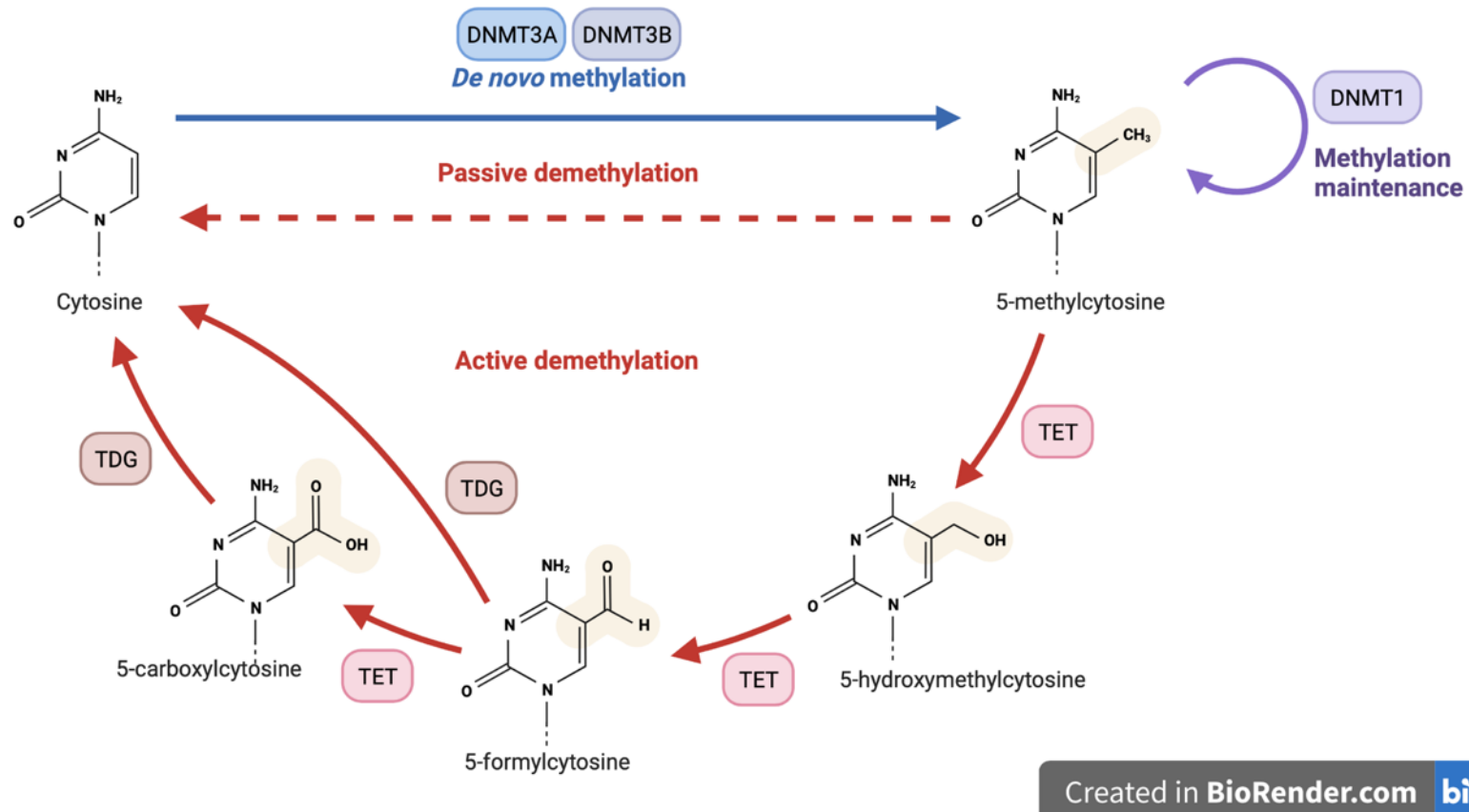
The most widely studied epigenetic modification is that of DNA methylation. It is a reversible reaction involving the covalent addition of a methyl group ( $\text{CH}_3$ ) to a cytosine nucleotide at the C-5 position in a DNA sequence by DNA methyltransferases (DNMTs). The newly modified base, 5-methylcytosine (5mC), is associated with silencing the transcriptional activity of genes on and around the area of methylation (Li & Zhang, 2014) (see Figure 1.9). The change is heritable, with *DNMT1* responsible for conferring DNA methylation onto the newly synthesised DNA strand during the replication process. Over 98% of DNA methylation occurs at a cytosine residue that is immediately followed by a guanine residue, termed CpG sites. They are unevenly distributed across the genome, with areas of unmethylated CpGs associated with increased rates of transcription. DNA methylation therefore plays a critical role in ensuring the correct transcription of genes during development and differentiation (Smith & Meissner, 2013).

Methylated DNA is present among most transposable and viral elements that comprise ~45% of the mammalian genome. This demonstrates that DNA plays a protective role in these intergenic regions,

preventing the expression of genetic elements that may be detrimental to the organism (Walsh, Chaillet & Bestor, 1998). The areas of DNA that contain CpGs that are unmethylated most often occurring in densely populated regions are termed CpG islands. Many are found at promoter sites or at genes involved in developmental regulation, enabling their transcription. It is hypothesised that the unmethylated state of CpG sites is maintained through the concentrations of transcription factors present that bind to these areas of sequence, exhibiting an inhibitory effect (Smith & Meissner, 2013). The locations of these sites are conserved among mammals, emphasising the functional importance of these sites (Illingworth *et al.*, 2010).

There is significant crosstalk between DNA methylation and other epigenetic mechanisms. *DNMT* enzymes are known to interact with methylation and acetylation enzymes involved in histone modification (Section 1.6.2) with evidence suggesting that they co-operate with each other to transcriptionally regulate a region of DNA (Fuks *et al.*, 2000, Fuks *et al.*, 2003).

## DNA Methylation



**Figure 1.9** – Diagram depicting the conversion of cytosine to 5-methylcytosine via the addition of a methyl group, facilitated by DNA methyltransferase. Methylation status is typically maintained throughout life by the enzyme DNMT1 but can be reversed via ten-eleven-translocation methylcytosine dioxygenases (TET).

## 1.5.2 - Histone modifications

To facilitate the condensing of DNA into chromosomes, DNA is first packed into nucleosomes, regions of DNA approximately 150bp in length encircling a histone octamer (Luger, Dechassa & Tremethick, 2012). In regions of DNA that are densely populated with nucleosomes, the positive charge of the nucleosome interferes with the negative charge of DNA, preventing the binding of transcription factors or polymerases. Analogous to DNA methylation, histones are also modified. This modification can occur in a variety of ways and serve as an important epigenetic modification that regulates gene expression (Bannister & Kouzarides, 2011).

Each histone molecule contains N-terminal 'tails' that protrude from an organised central structure. Multiple chemical modifications can occur at amino acid residues located on these tails: for example, methylation, acetylation, ubiquitination, and phosphorylation (Kouzarides, 2007). Histone methylation occurs at lysine residues and can occur at three levels depending on the number of added methyl groups: monomethylation, dimethylation and trimethylation. The number of added methyl groups to lysine 4 of histone molecule *H3* (*H3K4*) within the nucleosome has been shown to be correlated to gene expression levels, with high levels of trimethylated *H3K4* associated with 5' regions of nearly all active genes (Ruthenburg, Allis & Wysocka, 2007). Conversely, unmethylated *H3K4* is a marker for transcriptionally silent chromatin. Acetylation of histone residues has been identified as a residue that can decondense chromatin. The acetyl residue neutralises the basic charge of the lysine, resulting in nucleosomes becoming unpacked and increasing the rate of transcription at these sites (Lawrence, Daujat & Schneider, 2016).

Histone modification serves two functional purposes by enabling the structural remodelling of chromatin. It acts as a storage mechanism for DNA, by establishing densely packed chromatin, encouraging nucleosomes to be transcriptionally inaccessible (heterochromatin), or to enable nucleosomes to open the structure of DNA and be transcriptionally accessible (euchromatin). Histone modifications serve as an important epigenetic regulator through this function, by controlling which sections of the genome can be expressed and by maintaining the structural integrity of chromatin, ensuring that the genome is able to be efficiently stored inside the confines of the cell nucleus.

## 1.5.3 - Chromatin remodelling

The structural remodelling of chromatin enabled by the modification of histones significantly effects transcriptional profiles of the associated cell (section 1.5.2). In addition to histone modification enabled remodelling, various large protein complexes, termed ATP-dependent chromatin remodelling enzymes, provide the means to alter chromatin structure, utilising ATP hydrolysis to both 'open' and/or condense chromatin structure by displacing or translocating histone octamers from the underlying

DNA sequence. Chromatin remodelling enzymes can be categorised into four major families based upon their ATPase and associated flanking domains – *SWI/SNF*, imitation *SWI (ISWI)*, chromo-domain, helicase, DNA-binding (*CHD*), and inositol requiring 80 (*INO80*).

The most understood family of chromatin remodelling complexes are that of *SWI/SNF* – which are implicated in differentiation and DNA repair and mutations to *SWI/SNF* complexes are associated with cancer. For example, *SMARCA4* and *SMARCB1* are associated with medulloblastoma and atypical teratoid/rhabdoid tumours respectively (Doussouki, Gajjar & Chamdine, 2019; Holdhof *et al.*, 2021). Cells express a range of remodelling complexes that act upon chromatin in unison. It is theorised that the composition/distribution of chromatin remodelling complexes that exist within the cell are what establishes chromatin architecture globally, and the degree to which each complex is present helps to define cell identity (Langst & Manelyte, 2015).

## 1.5.4 - Methylation & cancer

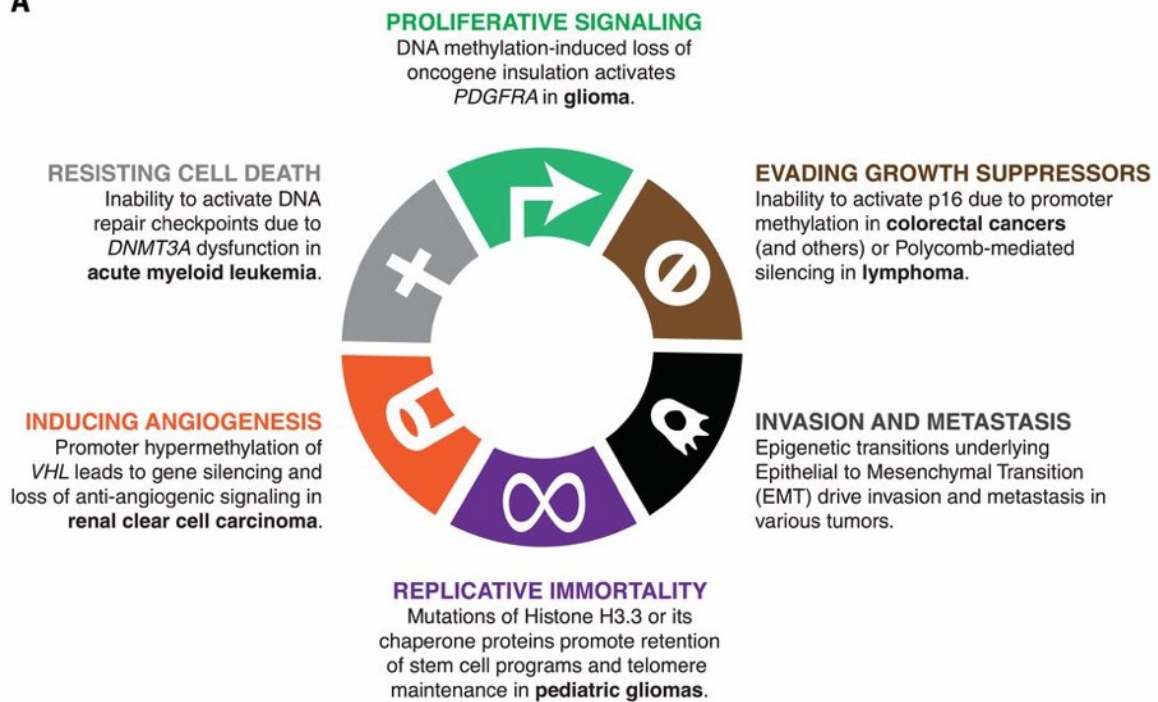
The critical role of epigenetic mechanisms in tumorigenesis has long been recognised. DNA methylation is thought to play the most significant role, by facilitating the transition from normal gene expression to a pro-tumorigenic state at specific loci. Epigenetic alterations have been detected in several cancers and are now recognised as a hallmark property of tumorigenesis (Shen & Laird, 2013). In the paediatric brain tumour medulloblastoma, several regions of hypomethylation (abnormally low levels of methylation) associated with increased gene expression were detected including several partially methylated domains affecting nearly a third of the whole genome (Hovestadt *et al.*, 2014).

Epigenetic silencing events have been associated with causing several recurrent ‘driver mutations’ of TSGs and oncogenes in different tumour types (Choi & Lee, 2013). Many of these genes are associated with many other cancer hallmarks, such as angiogenesis and cell cycle regulation (Figure 1.10). Whilst these regions of hypermethylation have been identified, cancer cells are broadly associated with lower levels of global methylation. Hypomethylation of intergenic *Alu* and *LINE-1* elements has been strongly linked to tumorigenesis, as their expression results in subsequent genomic rearrangements and general instability (Virani *et al.*, 2012).

Analysis of the entire DNA methylation landscape of a tumour is referred to as the cancer methylome and has evolved into a significant field of study over recent decades. The number of DNA methylation studies involving cancer is large and has led to the identification prognostically relevant biomarkers in numerous cancers such as lung and colorectal cancer (Paska & Hudler, 2015). There are several methylation biomarkers of cancer for which tests are now commercially available. A prominent example is the Cologuard test kit, which is a test for the methylation status of the *VIM* gene (Li *et al.*, 2014).



A



**Figure 1.10** – Examples of epigenetic influence on tumorigenesis (Flavahan, Gaskell & Bernstein, 2017)



## 1.6 – Tumours of the central nervous system

After leukaemia's, CNS tumours are the second most frequent type of tumour in children (26% of tumours in children aged 0-14) and are the most common solid tumour overall (Cancer Research UK, 2021). The central nervous system comprises the brain and spinal cord, and both are responsible for controlling bodily functions that are essential to survival. The spinal cord is responsible for relaying signals from the brains, as well as enabling sensory nervous signalling (sensation) and muscular movement. This fundamental role in ensuring survival and maintaining quality of life results in tumours of the CNS being a serious obstacle to survival, as any treatment (surgery, therapy) must take into consideration the surrounding healthy tissue that may also be damaged/affected by treatment.

Epidemiologically, patients with CNS tumours in the UK have poor outcomes and new treatments are urgently needed. In the years 2016-2018, over 12,288 new patients were diagnosed with a tumour of the CNS (brain, spinal cord, intracranial & other CNS tissue), with an age standardised incidence rate of 18.6 per 100,000 in the population (Cancer Research UK, 2021). Collectively, 12% of patients survive their tumour for 5 years, and the number of deaths from CNS tumours in the years 2016-2018 was 5,380 on average (Cancer Research UK, 2021). The overall change in incidence of CNS tumours is also increasing, with a recorded increase of 34% since the 1990s (Cancer Research UK, 2021). These concerning statistics are compounded by our relative lack of knowledge about CNS tumours compared to that of other cancers. This is due to their rarity and large number of histological types. The World Health Organisation (WHO) currently identify more than 120 types of CNS tumour and have attempted to standardise the collection and classification of data using molecular information (Louis *et al.*, 2021).

### 1.6.1 - Types of brain tumour

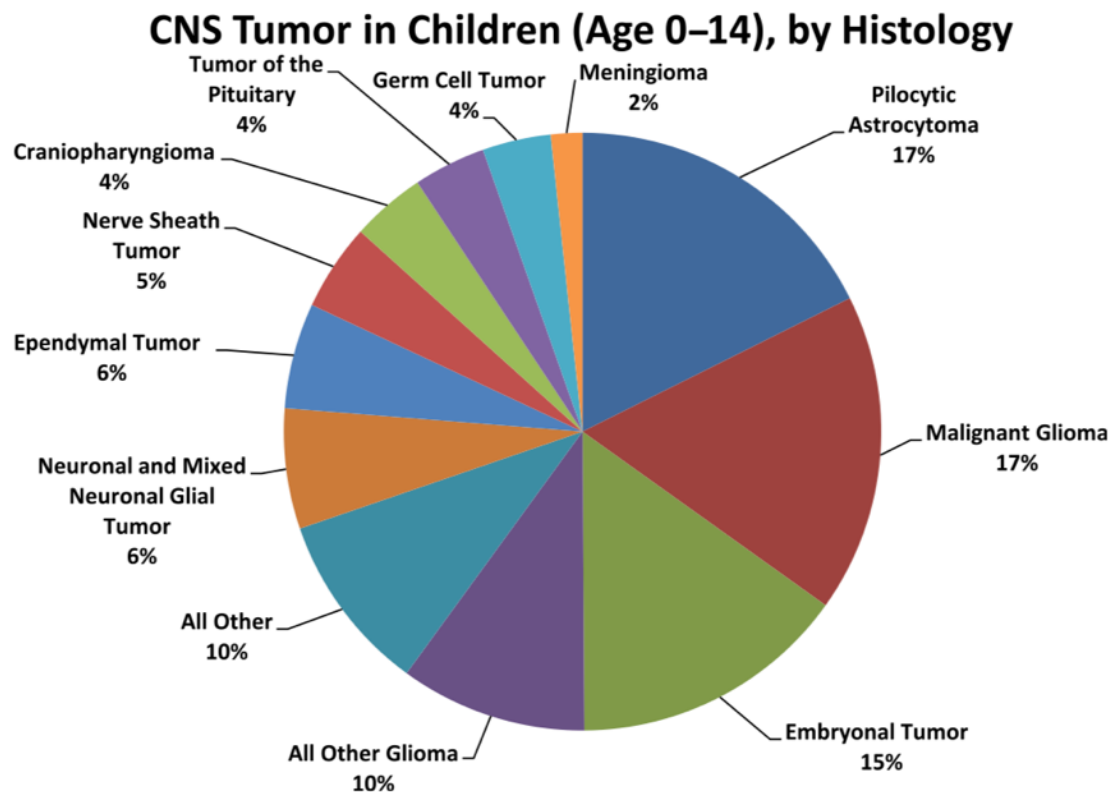
Most tumours of the CNS affect the brain. In 2021, the WHO updated their previous 2021 classification of 'Tumour Types of the Central Nervous System' to present a more accurate method of grouping different brain tumours (Louis *et al.*, 2021). This was to incorporate molecular information into the previous histological classification, due to advances in the research field enabled by genomics research. Across all age groups in the UK, the most common brain tumour was categorised as astrocytoma (34%), a type of brain cancer that originates in the astrocytes of the cerebrum. This is followed by meningiomas, that occur in the meninges (brain membrane tissue) that comprise 21% of cases, followed by tumours of unknown type (14%) and numerous others that occur in less than 10% of cases (Table 1.1).

Morphological Group	% of all Brain, other CNS and intracranial tumour cases	% of these more aggressive	% of these less aggressive
Astrocytomas	34%	95%	5%
Meningiomas	21%	8%	92%
Pituitary	8%	1-2%	98-99%
Gliomas unspecified	6%	*	*
Cranial and paraspinal nerve tumours	6%	5%	95%
Oligodendrogliomas	3%	*	*
Ependymomas	2%	75%	25%
Embryonal tumours	2%	100%	0%
Other tumour types	5%	*	*
Unknown or unspecified type	14%	*	*

**Table 1.1** – Most common types of brain tumour across all age groups (Cancer Research UK, 2019). Medulloblastoma, an embryonal tumour, occupies a small percentage of brain tumour incidence when considering both children and adults.

## 1.6.2 - Paediatric brain tumours

For children, brain tumours are the second most common type of tumour (Cancer Research UK, 2021). The reason for this increase in prevalence is largely due to a significant decrease in prevalence of other cancers that are more significantly influenced by environmental factors. Younger patients are at a lower risk of being influenced by environmental factors due to their age, as environmental factors become more influential the longer they are present. The distribution of brain tumours varies significantly when considering the age group of the patient. For paediatric patients, the distribution of brain tumour types is more evenly spread, with embryonal tumours appearing in over 15% of patients alongside pilocytic astrocytoma and glioma with 17% each respectively (Figure 1.11). This significant increase in the prevalence of embryonal tumours in children is due to the nature of the cells from which they develop. Embryonal tumours arise from cells that remain from foetal development in the womb, and are broadly classified into 7 types; Medulloblastoma, CNS neuroblastoma, *FOXR2*-activated, CNS tumour with *BCOR* internal tandem duplication, CNS embryonal tumour, Embryonal tumour with multi-layered rosettes, Cribriform neuroepithelial tumour and Atypical teratoid/rhabdoid tumour (Louis *et al.*, 2021).



**Figure 1.11** – Distribution of CNS tumours in children by histological type (McNeill, 2016). When focusing on children, the distribution of brain tumours differs significantly, with embryonal tumours occupying a significantly increased proportion of all tumours that occur in this age group.

## 1.7 – Medulloblastoma

Medulloblastoma (MB) comprises a subset of embryonal, high grade, malignant brain tumours of the posterior fossa or cerebellum, located at the lower rear section of the brain. It predominantly occurs in children and males but is not restricted to any demographic. Of all paediatric brain tumours, it is the most common, representing around a fifth of diagnosed brain tumours in children (Khanna *et al.*, 2017). The malignant nature of medulloblastoma is recognised by its classification as a grade IV tumour by the WHO (Louis *et al.*, 2016). It is predisposed to metastasis in the neuroaxis, accommodated by its proximity to the fourth ventricle of the cerebrospinal fluid (CSF) system and embryonal origin (Valtz *et al.*, 1991). The majority of medulloblastomas originate in the vermis, a section of tissue that bridges across both sides of the cerebellum. In adults, disease is more likely to arise around the hemisphere of the cerebellum.

Since its naming and distinction from glial tumours in the 1920s (Cushing & Bailey, 1925), medical descriptions of medulloblastoma now recognise a wide range of histological and molecular subgroups (Section 1.7.4). This culminated in the 2016 WHO reclassification of the disease by integrating genetic

profile and histology to enable a refined diagnosis associated with more accurate prognostic outcome (Louis *et al.*, 2016). Most medulloblastomas are sporadic (~95%) and no root cause has been determined. Environmental factors are not deemed a determining factor of tumour development due to its frequent presentation in children, where environmental influences do not have time to become a risk (de Bont *et al.*, 2008). Multiple familial conditions are associated with increased risk of developing medulloblastoma, with Gorlin, Turcot and Li-Fraumeni syndromes associated with increased risk of tumour development. The disease is discussed further in detail in the following sections.

### 1.7.1 - Epidemiology

Medulloblastoma is the most commonly diagnosed brain tumour of the primary CNS in the paediatric population (Ward *et al.*, 2014). The most recent epidemiological data is sourced from the Central Brain Tumour Registry of the United States (CBTRUS), which reports that approximately 20% of all paediatric tumours are MB, with an annual average of 297 cases (0.49 per 100,000) diagnosed per year (Ostrom *et al.*, 2021). Out of all embryonal tumours diagnosed in children, MB consists of 68.9% of cases in 0–19-year-olds (Figure 1.11). Incidences have been recorded in older age groups, but the prevalence of the disease decreases with age in agreement with its embryonal origin. Further subdivision of this demographic reveals a bimodal peak in incidence of MB among ages 3–4 and 8–10, at a rate of 0.53 and 0.58 per 100,000 respectively. MB is also disproportionately represented by more males than females, occurring at a ratio of 1.7:1 in the US (Ostrom *et al.*, 2021). No ethnic predisposition is currently apparent, but data collection remains in its infancy in this regard, and it will take some years before any trends relating to this aspect will be revealed (Smoll & Drummond, 2012). Medulloblastoma presentation associated with familial cancer predisposition syndromes (e.g., familial adenomatous polyposis, nevoid basal cell carcinoma syndrome) accounts for ~5% of diagnosed cases (Smoll & Drummond, 2012, Waszak *et al.*, 2018).

Survival has seen a radical improvement over the past 70 years, rising from a 5-year survival rate of 22% in the 1950s to 49% in the 1970s thanks to the addition of chemotherapy to standard treatment protocols (Farwell, Dohrmann & Flannery, 1984, Packer *et al.*, 1991). Further improvements in other areas such as surgical technique, appropriate therapy escalation, and improved palliative care for patients has pushed MB survival rates even higher, reaching up to 89% in the present day in favourable risk-group patients (Ramaswamy *et al.*, 2016). Survival data for patients in the US is relatively high compared to other embryonal brain tumours at 1, 2, 5 and 10 years, with a survival percentage of 89.3%, 83%, 73.2% and 64.9% respectively across all age groups (Ostrom *et al.*, 2021). Survival rates are highest in adolescents and young adults, whereas rates in adults over 40 years are lowest (Ostrom *et al.*, 2021). This is likely due to the increased amount of research targeted at childhood MB, for which targeted therapies and clinical trials have been focused, whereas adult MB has not received such attention, likely due to low prevalence causing difficulties in funding (Lassaletta & Ramaswamy, 2016).

Lowest survival rates are observed in patients that suffer relapse. MB relapse is usually metastatic (~85% of patients) and is associated with a 3-year survival rate of 18%, far lower than any other patient group (Koschmann *et al.*, 2016).

## 1.7.2 - Clinical presentation/symptoms

Most symptoms associated with medulloblastoma are because of its location in the brain. The cerebellum is crucial for healthy motor control and sensory regulation. It receives sensory information from the spinal cord and regulates the resulting motor response. It is responsible for voluntary actions such as balance and coordination, and controls muscular activity associated with movement.

The presence of a growing cellular malignancy in the cerebellum can present several symptoms associated with aberrant cerebellar function. Common symptoms that have been recorded include headaches, nausea, tiredness, poor balance, and coordination. Headaches are the most commonly presented symptom, present in 65% of patients (Coleman *et al.*, 2012), and are typically most intense in the morning before improving throughout the day as the patient is in an upright position.

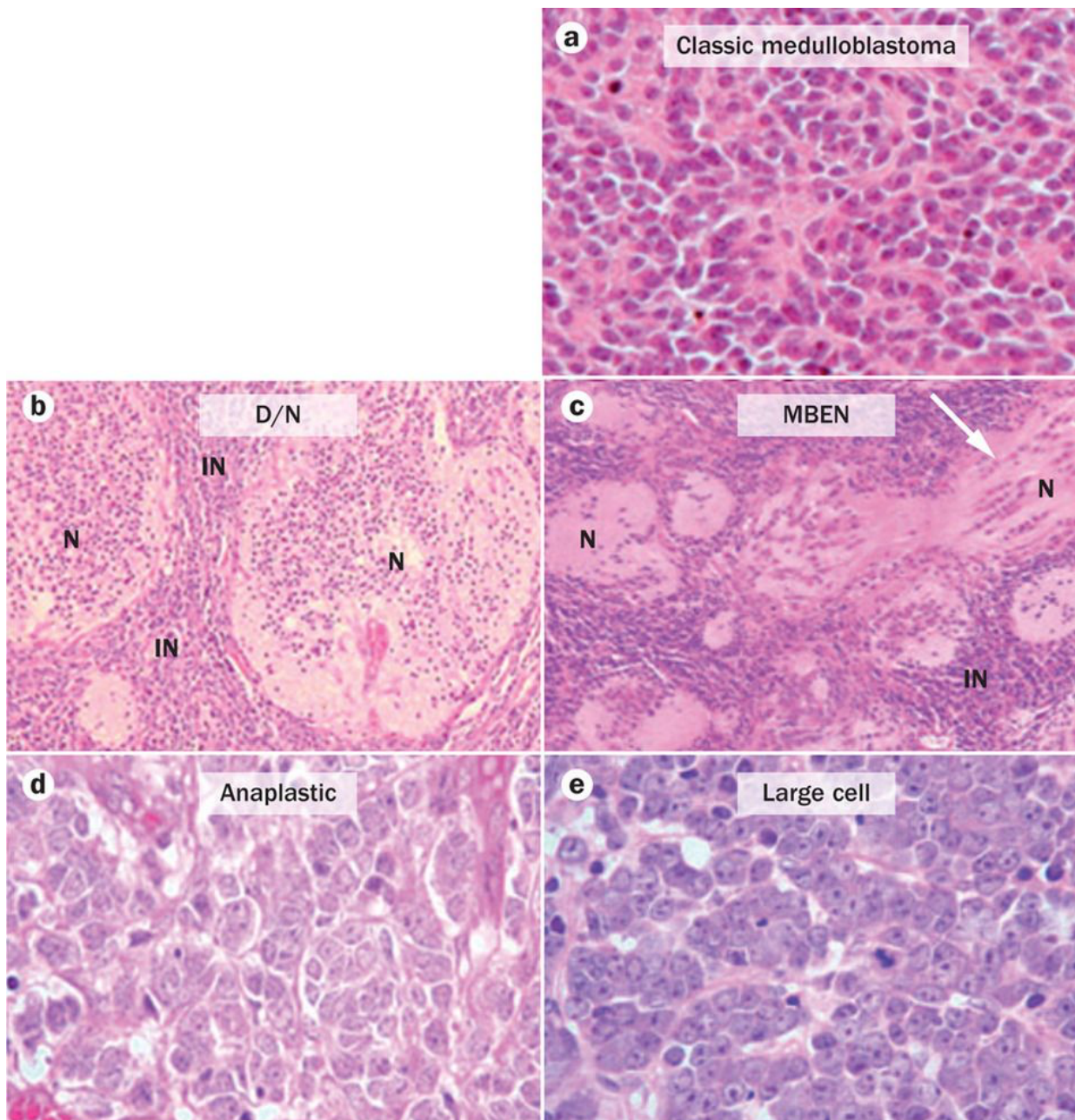
These symptoms typically arise due to increased intracranial pressure caused by a growing cell mass in the brain. Medulloblastomas frequently infiltrate the fourth ventricle of the brain, preventing adequate circulation of cerebrospinal fluid (hydrocephalus). This can amplify symptoms, as well as cause other conditions such as optic nerve swelling (papilledema), and ataxia in 76% and 62% of patients respectively (Coleman *et al.*, 2012).

The presentation of medulloblastoma varies from patient to patient. This is dependent on several factors but is largely dependent on the precise location of the tumour in the cerebellum and degree of metastasis. Symptoms, when presented individually, are not unique to medulloblastoma. This non-specificity can cause problems for patients and clinicians, often leading to misdiagnosis in the early onset stages until symptoms are more acute (Brasme *et al.*, 2012).

### 1.7.3 - Histology

Medulloblastoma is a small round cell tumour, characterised by its high cellularity, pleomorphism and excessive nuclei: cytoplasm ratio (Borowska & Jóźwiak, 2016). Homer-Wright rosettes (Rosettes with neuropil at centre) are also typically present in samples. Within the general histology of medulloblastoma, there are four unique histological variants; Classic, Desmoplastic/Nodular (DN), Medulloblastoma with Extensive Nodularity (MBEN) and Large Cell/Anaplastic (LCA) (Figure 1.12). Regardless of variant, all histological subgroups are classified as grade IV, owing to the primitive, embryonic nature of its small cell tumour classification (Louis *et al.*, 2016). Histology of medulloblastoma tumours is currently optimally determined in two stages. Local histological diagnosis is carried out by a neuropathologist after surgical excision, followed by referral to a specialist diagnostic institution for central pathological review (CPR) (Teot *et al.*, 2007). Discrepancies between local and central review are frequent, due to its rarity and similarity to other small cell tumours of the brain and, consequently, patients can be faced with misdiagnosis of their disease and/or its histology.





**Figure 1.12** – Histological subtypes of medulloblastoma (Gajjar & Robinson, 2014).

**A)** Classic medulloblastoma histology, characterised by the appearance of large numbers of small cells with high nuclear-to-cytoplasmic ratio. **B)** Desmoplastic nodular medulloblastoma. This group is characterised by the appearance of nodules (N) and internodular (IN) regions. **C)** Medulloblastoma with extensive nodularity (MBEN) is similar to D/N, but is distinguished by nodular regions occurring more abundantly. **D)** Anaplastic histology is identified by the appearance of numerous apoptotic bodies and nuclear pleomorphism **E)** aptly named, large cell histology displays prominent nuclei and often occurs in tandem with anaplasia. Both D and E histology's are often grouped together into a large cell/anaplastic category (LCA).

### 1.7.3.1 – Classic histology

The classic histology variant of medulloblastoma (CMB) is characterised in much the same way as the disease is described; highly cellular, pleiomorphic, small, round cells with Homer-Wright rosettes present in some cases (Figure 1.12a). It is the most common histological variant, found in approximately 70% of cases (Ellison, 2010). CMB has been found in all four molecular subgroups of medulloblastoma (Taylor *et al.*, 2012) and consist of over >90% WNT subgroup cases (Ellison *et al.*, 2011). Prognosis in CMB's is good, but poorer than desmoplastic variants in SHH tumours (Northcott *et al.*, 2011).

### 1.7.3.2 – Desmoplastic/nodular

Desmoplastic Medulloblastoma (DN MB) is differentiated from CMB by the presence of nodules between regions of desmoplasia (growth of fibrous tissue) (Figure 1.12b). Nodules comprise of differentiated neural cells whereas cells of the internodular spaces are undifferentiated. Regions of nodularity are sparse in DN MB (Ellison, 2010). DN MB are subdivided into two histology's; conventional and paucinodular, found in ~67% and 33% of DN MB cases respectively (Schroeder *et al.*, 2014). DN MB is typically found in young children, representing over half of cases found in infants. DN MB cases are found in ~5% of all MB cases and desmoplastic variants (including MBEN) comprise the majority of SHH molecular subtype tumours (Taylor *et al.*, 2012). Prognostic outcome is better than both classic and large cell histology's and is therefore currently classified as 'low-risk' (Borowska & Józwiak, 2016).

### 1.7.3.3 – Medulloblastoma with extensive nodularity (MBEN)

DN MB's presenting with increased nodularity of irregular shapes are described as a separate MBEN histological variant (Figure 1.12c). They comprise of ~20% of desmoplastic MB cases and is the rarest histological variant overall, comprising of just 1% of cases (Korshunov *et al.*, 2018). MBEN histology tumours are present only in infants and young children and are prognostically the most favourable histological variant (Giangaspero *et al.*, 1999). All MBEN histology tumours belong to the SHH subgroup and has been hypothesised as an "end-point" of DN differentiation (Taylor *et al.*, 2012, Borowska & Józwiak, 2016).



### 1.7.3.4 – Large cell/anaplastic

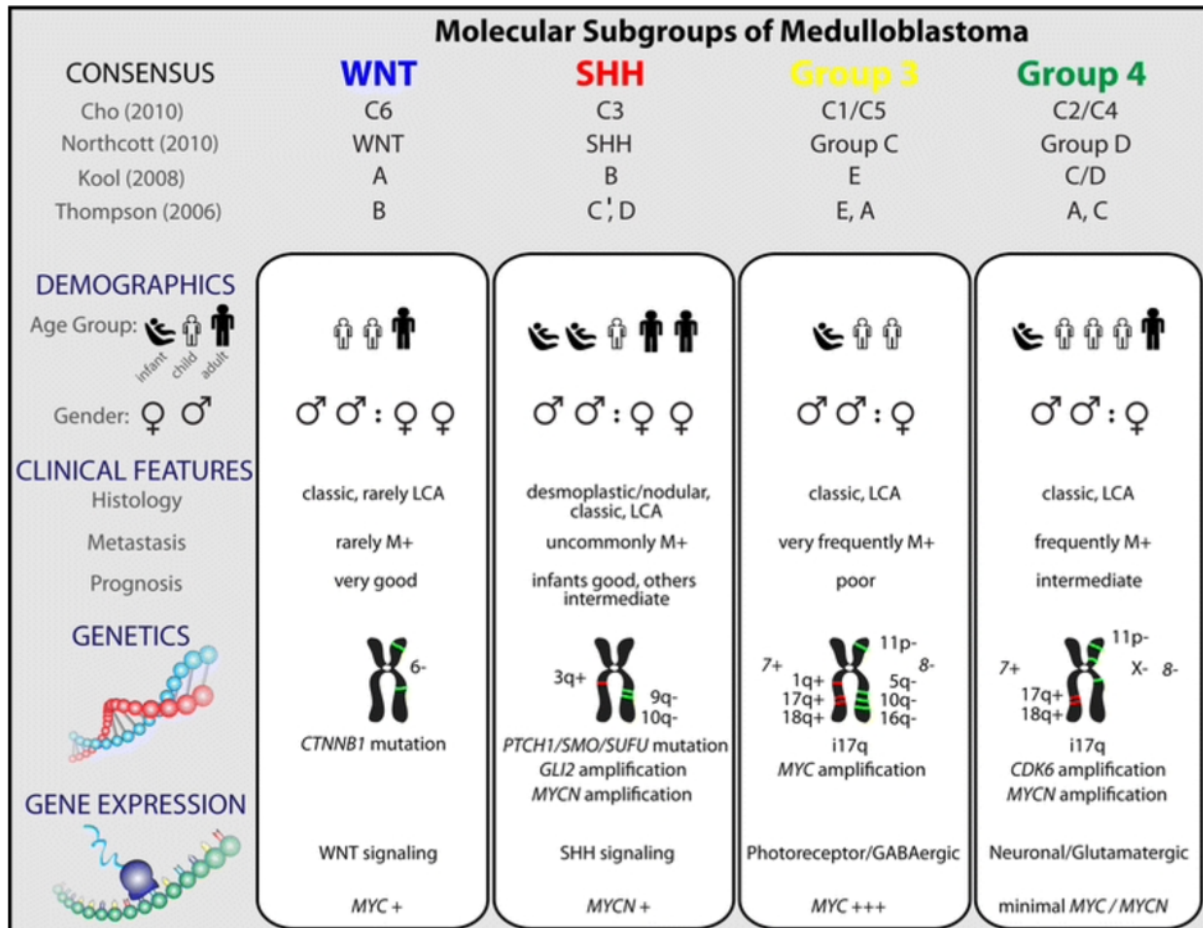
Large cell and anaplastic (LCA) histological variants account for ~20% of all MB tumours (Schroeder *et al.*, 2014) (Figures 1.12d & 1.12e). Whilst the two histology's are separate, they are typically found co-existing in samples. Large cell MB is characterised by the presence of significantly larger cells, with large nuclei containing open chromatin. Anaplastic histology's are poorly differentiated and contain significant pleomorphic nuclei (Borowska & Józwiak, 2016). LCA tumours are aggressive, frequently metastatic, and consequently are associated with a poor prognostic outcome. LCA variant tumours have been reported in all molecular subgroups of medulloblastoma, but are most prominent in the group 3 subtype, which is also associated with poor outcome (Northcott *et al.*, 2012).

## 1.7.4 - Medulloblastoma is comprised of distinct molecular entities

Molecular analysis of medulloblastoma revealed underlying heterogeneity within the disease. Transcriptional and methylation profiling research in the late 2000s-early 2010s identified four distinct molecular subgroups of medulloblastoma; Wnt/wingless (WNT), Sonic Hedgehog (SHH), Group 3 & Group 4 (Thompson *et al.*, 2006; Kool *et al.*, 2008; Cho *et al.*, 2011; Northcott *et al.*, 2011; Schwalbe *et al.*, 2013). Critically, each molecular subgroup is also associated with clinical features such as patient demographic and prognostic outcome (Cho *et al.*, 2011; Northcott *et al.*, 2011; Taylor *et al.*, 2012). The clinical utility of molecular subgrouping of tumours is such that they predict disease outcome and behaviour more accurately than histological diagnosis (Taylor *et al.*, 2012). Molecular subgroups are independent of histological subtype, but there are areas of overlap between the two (Section 1.7.3).

The molecular subgrouping consensus reached in 2012 has since been integrated into diagnostic protocols by the WHO (Louis *et al.*, 2016; Louis *et al.*, 2021). This has placed medulloblastoma at the forefront of precision oncology, providing a platform for targeted therapeutics and subgroup directed clinical management to enter clinical trials. Subgroup directed clinical trials of medulloblastoma have just recently begun being implemented into clinical trials design. WNT subgroup patients are currently being trialled for therapy de-escalation (CTI: NCT02212574), whereas new SHH subgroup directed therapeutics are being trialled in relapsed MB patients (CTI: NCT01708174). Current treatments for MB are unsatisfactory in terms of the damaging effects it has on the developing brains of children who survive, as well as the poor survival chances of adults and patients that suffer relapse (Section 1.7.8). Taking molecular subgroup into account when designing new clinical trials ensures that robust, personalised treatment strategies can be adopted for future patients.

Since then, the molecular subgroups of medulloblastoma have been further defined, with various molecular subtypes identified within SHH, Group 3 and Group 4 (Cavalli *et al.*, 2017; Schwalbe *et al.*, 2017; Sharma *et al.*, 2019; Northcott *et al.*, 2019), which are discussed further in the following sections (Figure 1.13).



**Figure 1.13** – The molecular characteristics of the four-consensus primary molecular subgroups of medulloblastoma were first established in 2012 (Taylor *et al.*, 2012), and are each defined by a distinct set of demographic, clinical and genetic features.

### 1.7.4.1 – WNT medulloblastoma

Accounting for 10% of all medulloblastoma diagnoses, WNT medulloblastoma is the rarest subgroup. 5 year-survival rates for WNT patients are very high (~90%), with the primary cause of death due to treatment complications and not the tumour itself (Ellison *et al.*, 2011). The WNT subgroup is named so by tumours having characteristic somatic mutations of genes involved in the WNT/beta-catenin signalling pathway (Section 1.7.4). Individuals with Turcot syndrome are predisposed to WNT medulloblastoma because of germline mutation of the *APC* gene, a WNT pathway inhibitor.

The large majority of WNT medulloblastomas have a classic histology, with a small proportion of LCA phenotypes being reported (Louis *et al.*, 2007). Regardless of histology, the prognosis in WNT patients

does not appear to be affected by histology or *TP53* mutation (Ellison *et al.*, 2011). Whilst medulloblastoma is more common in males overall, the gender ratio of patients with WNT tumours is approximately 1:1 (Northcott *et al.*, 2019). WNT tumours can occur at any age, but they are most common in children >3 years and adults. Chromosomal alterations are limited in WNT tumours, but 90% of WNT patient tumours have monosomy of chromosome 6 (Clifford *et al.*, 2006).

It is theorised that the excellent prognosis associated with WNT medulloblastomas exhibiting 'leaky' vasculature caused by the high expression of antagonists of WNT signalling. WNT tumours are often haemorrhagic during resection, which is suggested to render them more susceptible to chemotherapies than medulloblastomas of other subgroups (Northcott *et al.*, 2019). The favourable prognosis of patients with WNT tumours has led to the suggestion that they are being 'over-treated'. Given that medulloblastoma treatment is associated with significant morbidity in the long term, it has recently been proposed that clinical trials explore de-escalation of therapy whilst maintaining survival rates in the WNT population. Preliminary research comparing the intellectual abilities of children treated with varying levels of therapy appear to support this proposal in WNT patients (Moxon-Emre *et al.*, 2016). Indeed, exploratory trials of this nature have now begun to take place, with a phase 2 study having started in 2017 (CTI: NCT02724579) and the PNET5 trial, which aims to introduce a 'low-risk' treatment arm for WNT patients (CTI: NCT02066220).

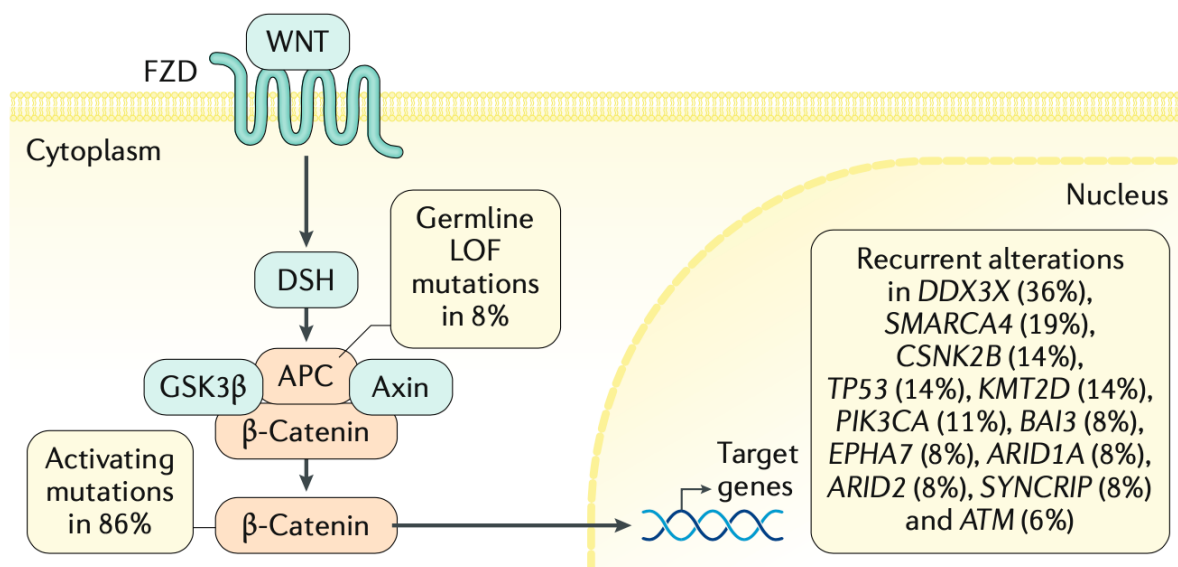
### 1.7.4.2 – WNT signalling & implicated genes

The WNT signalling pathway is highly conserved in the animal kingdom. It is comprised of approximately 19 secreted glycoproteins, and is a fundamental component of cell proliferation, tissue homeostasis and control of embryonic development (MacDonald, Tamai & He, 2009). Mutations in this pathway are associated with numerous birth defects and contribute to cancer pathogenesis (Clevers, 2006). The most critical pathway of WNT is 'canonical WNT signalling'. Canonical WNT signalling functions by regulating the concentration of beta-catenin, a co-activator of the *TCF/TLE* transcription factor (Valenta, Hausmann & Basler, 2012). Aside from monosomy 6, a key characteristic of WNT MB tumours is the nuclear accumulation of  $\beta$ -catenin. Such accumulation is acquired due to mutations of genes involved in WNT signalling, leading to oncogenic activity and transcriptional upregulation.

The most prevalent somatic mutation in WNT tumours, found in nearly all WNT tumours is that of *CTNNB1*, encoding for beta-catenin, occurring in approximately 90% of patients (Taylor *et al.*, 2012). The majority of *CTNNB1* mutations are stabilising (~70%), preventing its usual breakdown and increasing its ability as an effector (Robinson *et al.*, 2012). Missense mutations of DEAD-Box RNA Helicase (*DDX3X*) are also enriched in WNT subtype Medulloblastoma, occurring in 36% of patients (Oh *et al.*, 2016). *DDX3X* is implicated in chromosomal segregation (Pek & Kai, 2011), as well as regulation of gene expression (Schröder, 2006). *DDX3X* is located on the X chromosome, and

subsequently displays a strong bias for mutation in males (Dunford *et al.*, 2017). Familial mutation of the *APC* gene (Turcot syndrome) predisposes an individual to WNT medulloblastoma, as well as colorectal adenomas (Hamilton *et al.*, 1995). Less commonly mutated genes include *AXIN1* & *AXIN2*, which are also components of canonical WNT signalling, as well as *TP53* (Taylor *et al.*, 2012). However, whilst the presence of *TP53* is indicative of poor prognosis in SHH MB, it does not appear to impact survival in WNT tumours (Zhukova *et al.*, 2013).

Targeting of the WNT pathway using therapeutics is complicated, due to its role in several critical somatic processes independent of its potential to cause tumorigenesis. WNT signalling has important roles, such as skin regeneration, liver repair, and neurogenesis (Kahn, 2014), and any targeted therapies must be carefully considered as to not interfere with signalling in healthy tissue.



**Figure 1.14** – How WNT canonical signalling is affected in medulloblastoma (Northcott *et al.*, 2019). The exposure of extracellular WNT signalling results in the sustained activation and accumulation of the transcription factor beta-catenin, resulting in the aberrant expression of target genes. Percentage figures next to genes denote the frequency at which mutations are present within the WNT-subgroup population.

### 1.7.4.3 – SHH medulloblastoma

Like the WNT subgroup, the SHH subgroup is named so due to the Sonic Hedgehog signalling pathway that is prominently affected in tumours of this type (Section 1.7.4.4). SHH tumours share some commonality in terms of chromosomal alteration compared to that of other molecular subtype tumours, but loss of chromosome 9q is limited to the SHH subgroup (Cho *et al.*, 2011). The familial condition Gorlin syndrome (*PTCH1* mutation) and Li-Fraumeni syndrome (*TP53* mutation) are known to leave affected individuals predisposed to developing SHH subgroup MB.

~30% of MB tumours belong to the SHH subgroup, and incidence rates are binomially distributed, occurring frequently in infants <3 years and adults >16years old, but less frequently in-between (Northcott *et al.*, 2011). The two are typically found in different areas of the cerebellum, with infant SHH MBs found in the midline and adult SHH found in the hemispheric areas. There is no gender imbalance for SHH subgroup tumours, with most cohort studies observing a gender ratio of 1:1, like the WNT subgroup. Interestingly, all MBEN histology group tumours belong to the SHH subtype and many tumours with DN histology are also SHH (Kijima & Kanemura, 2016). However, not all SHH tumours belong to these histological groups, with classic and LCA histology's also being recorded in cohort studies (Taylor *et al.*, 2012). Prognosis in SHH patients is intermediate (~75% 5-year-survival), falling in-between WNT and Group 3 MB, which have good and poor prognoses respectively. However, there is significant prognostic variability present in the SHH tumour subgroup. SHH tumours with *TP53* mutation have very poor outcomes, with an overall 5-year survival of ~40% compared to that of ~80% in *TP53*-Wildtype SHH MB (Zhukova *et al.*, 2013). This has led to the sub-classification of SHH MB into '*TP53*-mutant' and '*TP53*-Wildtype' by the WHO (Taylor *et al.*, 2012).

More recent transcriptional and methylation profiling studies involving larger cohorts of SHH subgroup tumours revealed that SHH can be characterised into subtypes – alpha, beta, gamma, and delta, based upon distinct cytogenetic profiles and patient demographics (Cavalli *et al.*, 2018). Other splits of SHH medulloblastoma have shown that infant medulloblastoma (iSHH – largely corresponding to the beta and gamma subtypes identified by Cavalli *et al.*) can be further split into two groups – iSHH-I, enriched for *SUFU* and gain of chromosome 2, and iSHH-II, enriched for *KMT2D* and *BCOR* mutations respectively.

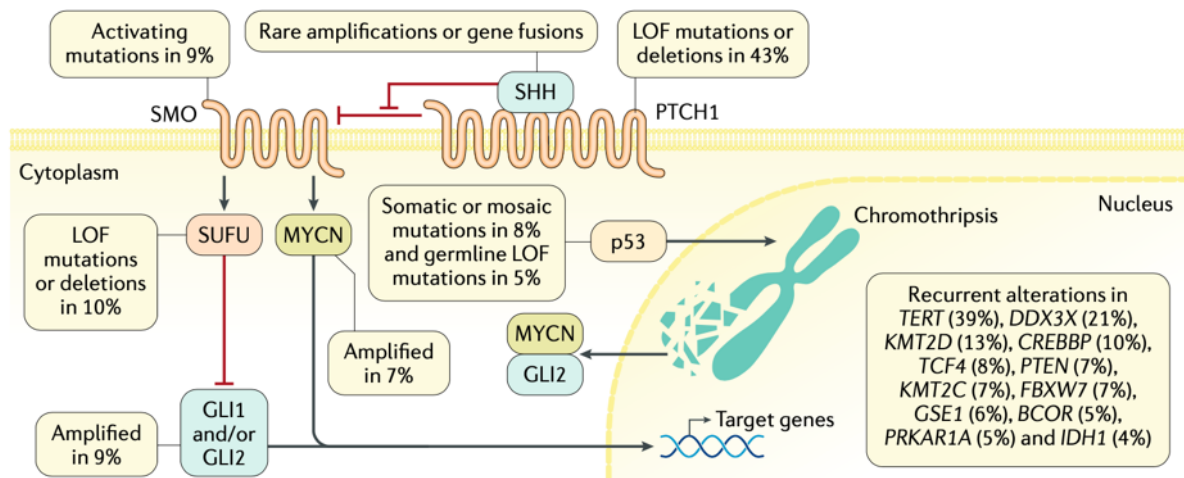
### 1.7.4.4– SHH signalling & implicated genes

The Sonic Hedgehog (SHH) signalling pathway is crucial to regulation and maintenance of developmental processes of the cell. It is involved in embryonic patterning and cellular differentiation, important processes that lead to the formation of organ tissues in multicellular organisms (Simpson, Kerr & Wicking, 2009). Aberrant mutations in this pathway have been linked to tumorigenesis and birth defects (Choudhry *et al.*, 2014). ‘SHH protein’, a hedgehog signalling ligand has been shown affect cell differentiation differently depending on its concentration. For example, in SHH naïve neural cells, different SHH concentrations caused cells to differentiate into different cells (Ericson *et al.*, 1997). Individuals with Gorlin syndrome have a germline mutation of *PTCH1*, the *ptc* receptor involved in Shh signalling (Lo Muzio, 2008). Other common mutations found in sporadic SHH MB also include *PTCH1* as well as *SUFU* (Brugieres *et al.*, 2010), *SMO*, *GLI1* and *GLI2* amplification (Northcott *et al.*, 2011). *GLI1*, a transcription factor, has been shown to interact with *E-CAD*, a component of the canonical WNT pathway, suggesting that there is some crossover between both the WNT and SHH pathways (Li *et al.*, 2007). This suggests there may be shared therapeutic targets among the two tumour subgroups.

As mentioned in 1.7.4.3, SHH subgroup tumours appear to have distinct transcriptomic profiles separated by age. Infant subtype SHH frequently present with *PTCH1* and *SUFU* germline mutation, whereas adult SHH is characterised by *PTCH1* and *SMO* mutations. Childhood SHH is more heterogenous, with the presence of *Gli2* and/or *MYCN* amplification characteristic of this subgroup as well as mutations of *TP53* (Archer, Mahoney & Pomeroy, 2017). Chromosomal changes in SHH tumours are also characterised by age. Infant and adult SHH frequently present with loss of chromosome 9q, the location of *PTCH* and *SMO* genes in the SHH pathway. Childhood SHH MB is infrequently associated with a wide range of chromosomal copy number alterations and changes due to mass chromosomal rearrangement (chromothripsis) (Rausch *et al.*, 2012).

Clinical treatments for SHH MB vary depending on subgroup. *TP53*-Mutant SHH tumours are classed as ‘high-risk’ and treatment is escalated appropriately. *TP53*-Wildtype tumours associated with a good prognosis and are associated with deescalated treatment strategies. Clinical trials for treatment of SHH tumours are currently ongoing using therapeutics targeting the SHH pathway. *SmO* inhibitors such as vismodegib have shown promising results in adult patients (Robinson *et al.*, 2015), where >80% of adult patients present with mutations in *SMO* or *PTCH1* (Kool *et al.*, 2014). Such a therapy is not suitable for children however, due to the important role SHH signalling plays in development (Frappaz *et al.*, 2021). It is therefore likely that this subgroup will be the first of the four to benefit from subgroup-directed therapeutics.





**Figure 1.15** – How canonical SHH signalling is affected in medulloblastoma (Northcott *et al.*, 2019). The SHH pathway is primarily regulated by the uptake of SHH via the binding to smoothed-Ptc complex, encoded by *SMO* and *PTCH1* respectively. Percentage values correspond to the proportion of SHH MB patients that exhibit mutations in the associated gene.

### 1.7.4.5 – Group 3 and Group 4 medulloblastoma

In contrast to WNT and SHH medulloblastoma, Group 3 (and Group 4) tumours are not characterised by aberrations of one specific signalling pathway. The two groups share many molecular similarities, and patient demographics exhibit significant overlap. However, the two groups differ when concerning prognosis. Group 3 medulloblastoma comprises ~25-30% of cases, occurring very rarely in adults, displaying a peak in incidence at 3-5 years of age (Northcott *et al.*, 2017). Group 3 medulloblastoma occurs nearly twice as often in males than females and is associated with the poorest survival out of all subgroups – with a 58% five-year overall survival reported in children and a 45% overall survival reported in infants (Northcott *et al.*, 2017). Many high-risk features are present in group 3 medulloblastoma, such as large/cell anaplasia, increased rates of metastases at presentation and high level *MYC* amplification, occurring in ~17% of patients in this group (Northcott *et al.*, 2012). Group 3 medulloblastoma often displays complex cytogenetic profiles, with isochromosome 17q, 8 loss, gain of 1q and 7 identified as prominent features in this group. Genetic markers of disease are not common, where *SMARCA4*, *KMT2D*, and *CTDNEP1* the only genes enriched in patients (Northcott *et al.*, 2019). The *SMARCA4* and *KMT2D* genes are known to have a functional role in chromatin remodelling (Kijima & Kanemura, 2016). This suggests that these mutations may in part explain for the large variation in chromosomal aberrancies that have been detected in Group 3 tumours. Group 3 medulloblastoma often displays complex cytogenetic profiles, with isochromosome 17q, 8 loss, gain of 1q and 7 identified as prominent features in this group.



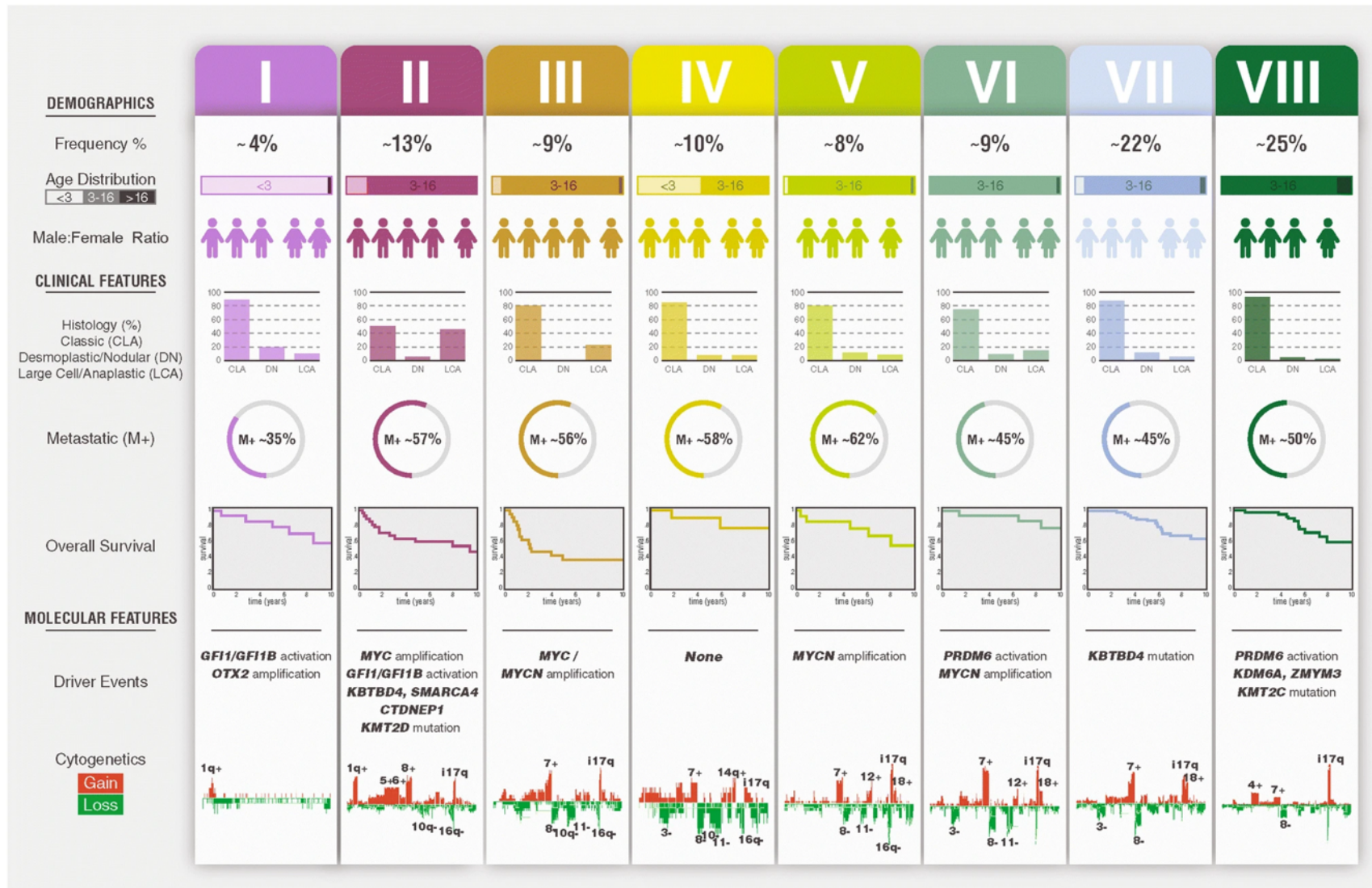
Amplifications of the *MYC* oncogene are most attributable to group 3 tumours. Whilst *MYC* mutations are present in all other subgroups, amplifications of *MYC* occur most frequently within this subgroup (~17%) and crucially, are linked with poorer prognostic outcome in group 3 patients than those that do not harbour the amplification. As well as amplification of *MYC*, increased expression can be driven by translocation with that of *PVT1* (Northcott *et al.*, 2012). *MYC* drives the expression of many metabolic processes linked to tumorigenesis in many different cancers, including medulloblastoma and is therefore a natural target for therapeutic development (Miller *et al.*, 2012). However, there are significant obstacles standing in the way. As a transcription factor, it has no distinct active site for inhibition and its location in the nucleus renders the use of antibodies to be impractical (Chen, Liu & Qing, 2018). This has resulted in more innovative, indirect approaches to inhibition being tested in clinical trials - an example of which would be the targeting of the cyclin-dependent kinase 4 and 6; their inhibition attenuates the transcription of *MYC* (Chipumuro *et al.*, 2014). An ongoing phase 1 trial looking into the effects of Palbociclib (*CDK4/6* inhibitor) is currently underway in paediatric patients with medulloblastoma and other tumours of the central nervous system (CTI: NCT02255461).

Group 4 is the final clinically recognised molecular subgroup of medulloblastoma. Like group 3 it is not as well characterised compared to WNT and SHH subgroups, where a detailed description of underlying pathogenesis driving the subgroup is lacking. Group 4 is the most common subgroup of medulloblastoma, comprising ~35% of cases (Kijima & Kanemura, 2016). They are the most common subgroup of medulloblastoma, comprising ~35% of cases (Kijima & Kanemura, 2016), and the most prominent cytogenetic change is isochromosome 17q. Notably in females, it is common to observe a loss of X chromosome, occurring in 80% of female tumours (Taylor *et al.*, 2012). Metastasis is common in this subgroup (~35%) but is not as high as that of Group 3 tumours (Kool *et al.*, 2012). Group 4 tumours have an intermediate prognosis, like that of SHH patients, with a 5-year survival rate of 75% (Khatua, 2016). They are histologically like WNT tumours, being predominantly classic and rarely presenting with the large cell/anaplastic variant. Group 4 tumours are most common in children over 3 years of age and are rare in both infants and adults. It has previously been described as “the prototypical medulloblastoma: a 7-year-old boy with a classic histology medulloblastoma that has an isochromosome 17q” (Taylor *et al.*, 2012). The most prevalent driver event is over-expression of *PRDM6*, which is tightly linked to tandem duplication of *SNCAIP*, occurring in 17% of patients (ref). Like group 3, mutations are rare in group 4, with no single gene occurring in over 10% of patients. Common mutations include *ZMYM3*, *KMT2C*, *KBTBD4* and *MYCN* amplification, all occurring in 6% of patients. Unfortunately, it remains as the only molecular subgroup of medulloblastoma without a murine model, rendering detailed genetic study *in-vivo* difficult. This lack of genetic characterisation has also hindered any development into Group 4-directed therapeutics. With group 4 being the most common molecular subgroup of medulloblastoma, it is imperative that the underlying biology behind these malignancies is understood to advance therapeutic development as quickly as possible.

### 1.7.4.6 - Identification of subtypes within Group 3 and Group 4 medulloblastoma

A lack of molecular markers distinct to each subgroup originally meant that identifying tumours of Group 3 and Group 4 difficult in a diagnostic setting. Whilst immunohistochemistry (IHC) is recommended for the WNT subgroup (where *CTNNB1* is a clear marker), no such markers exist for these two groups. Identification is instead based upon methylation profiling, where tumours of each subgroup cluster amongst its counterparts, allowing for tumours to be grouped effectively (Sharma *et al.*, 2019). As methylation datasets have increased, methylation profiling has demonstrated success in identifying further subtypes present within groups 3 and 4, helping to unpick the underlying heterogeneity present among both groups (Northcott *et al.*, 2017). Multiple subtype models have been proposed – including the Schwalbe *et al.*, classification, which split group 3 and group 4 into low and high-risk groups (Schwalbe *et al.*, 2017) and the Cavalli *et al.*, classification which identified alpha, beta, and gamma subsets within each molecular subgroup respectively (Cavalli *et al.*, 2017).

In 2019, these different definitions were reconciled via a combined analysis of over 1500 medulloblastoma samples via DNA methylation array in conjunction with a large portion of matched sample transcription data profiles (Sharma *et al.*, 2019). Over eight subtypes (I-VIII) were identified within these groups, each enriched with distinct molecular and clinic-pathological profiles (Figure 1.16). The eight consensus subtypes include both mixed groups and groups that comprise patients that solely belong to either subgroup – methylation profiling has been shown to discern between the groups effectively. These subtypes have since been recognised in the latest WHO classification of central nervous system tumours (Louis *et al.*, 2021) due to their associated clinical utility, and their discovery has subsequently enabled researchers to assess the heterogeneity within Group 3 and Group 4 in a more informed manner and will aid efforts to define the groups biologically in the years to come.



**Figure 1.16** – Clinico-pathological features of the molecular subtypes of Groups 3 and 4 in medulloblastoma (Sharma *et al.*, 2019). Like previous studies that identified the four classic molecular subgroups, further heterogeneity has been identified within subgroups, with 8 subtypes currently recognised within group 3 and 4.

---

## 1.7.5– Diagnosis of medulloblastoma

Patients presenting with symptoms (Section 1.7.2) that are consistent with a brain tumour, particularly symptoms that are progressive, are referred for a complete physical examination, neurological examination, and comprehensive review of their medical history. Physical examination of a patient with medulloblastoma will confirm the presence of commonly associated symptoms such as ataxia, papilledema and nerve palsies that may be caused by increased intracranial pressure. In infants, an increase in the size of the head or protrusions of the anterior fontanelle (membranous gaps between bones of skull during early development) may be observed (Quinlan & Rizzolo, 2017).

Whilst a significant number of symptoms of medulloblastoma are quite general, medulloblastoma patients will likely lack of any other apparent illnesses that may explain the symptoms at the time of physical examination. Concerning findings of the physical exam are compounded if the patient also presents with unusual results after neurological testing. Brain imaging scans are the first port of call, with CT and MRI showing excellent utility in confirming the presence of a cerebellar mass (Reulecke *et al.*, 2008). CT scans of medulloblastoma will show a mass at the vermis that extends into the fourth ventricle, causing it to look thinner than normal. Not all medulloblastomas will appear in the vermis. Less common areas such as the cerebellar hemisphere are typical of SHH subgroup tumours (Section 1.7.4.3). The area of mass is often hyperdense, and a contrast CT scan will show an area of high enhancement (contrast) between the suspected tumour and surrounding brain tissue (Figure 1.17). MRI scans, depending on the form ordered, will detect medulloblastomas efficiently, owing to its capability to show the area of interaction between the tumour and healthy brain in detail. MRI Type-1 scans, with and without gadolinium are most often used, and can accurately detect medulloblastoma in the brain (Millard & De Braganaca, 2016). If not performed before diagnosis. MR spectroscopy will often detect increased levels of choline, taurine peak as well as decreased levels of N-acetyl-acetate (Koeller & Rushing, 2003). MRI is often requested on the entire neuroaxis to evaluate metastatic behaviour before operating on the patient. Lumbar punctures, whilst effective to detect the presence of leptomeningeal spread, are not performed due to the already increased intracranial pressure in patients (Quinlan & Rizzolo, 2017). The identified location of a medulloblastoma via MRI has been shown to have a statistically significant relationship with molecular subgroup (Perreault *et al.*, 2014).

Whilst the above diagnostic tests are effective in diagnosing medulloblastoma, differential diagnoses to that of other masses in the posterior fossa can occur, such as atypical teratoid/rhabdoid tumour (ATRT) and ependymoma. ATRT is incredibly difficult to differentiate from medulloblastoma as both originate in similar areas and present with similar signs diagnostically (Koral *et al.*, 2008). This discerning between the two may only be achieved post-operation, when immunohistochemical tests can be performed on tumour tissue directly.

Due to the nature of the general symptoms of the early development of any brain tumour, the diagnosis of medulloblastoma can be delayed significantly. The median interval between the reporting of

symptoms to diagnosis of medulloblastoma has been reported as 65 days (Brasme *et al.*, 2012). A study into the relationship between the time taken to diagnose medulloblastoma and the subsequent effects on survival found complex results, with most of the variation concluded to be due to other factors such as the nature of the tumour (histology, subgroup) rather than the time of diagnosis (Brasme *et al.*, 2012). However, a median time of 2 months is a significantly long duration, and regardless of its effect on survival, it is important that this time is shortened so that suffering isn't prolonged unnecessarily.



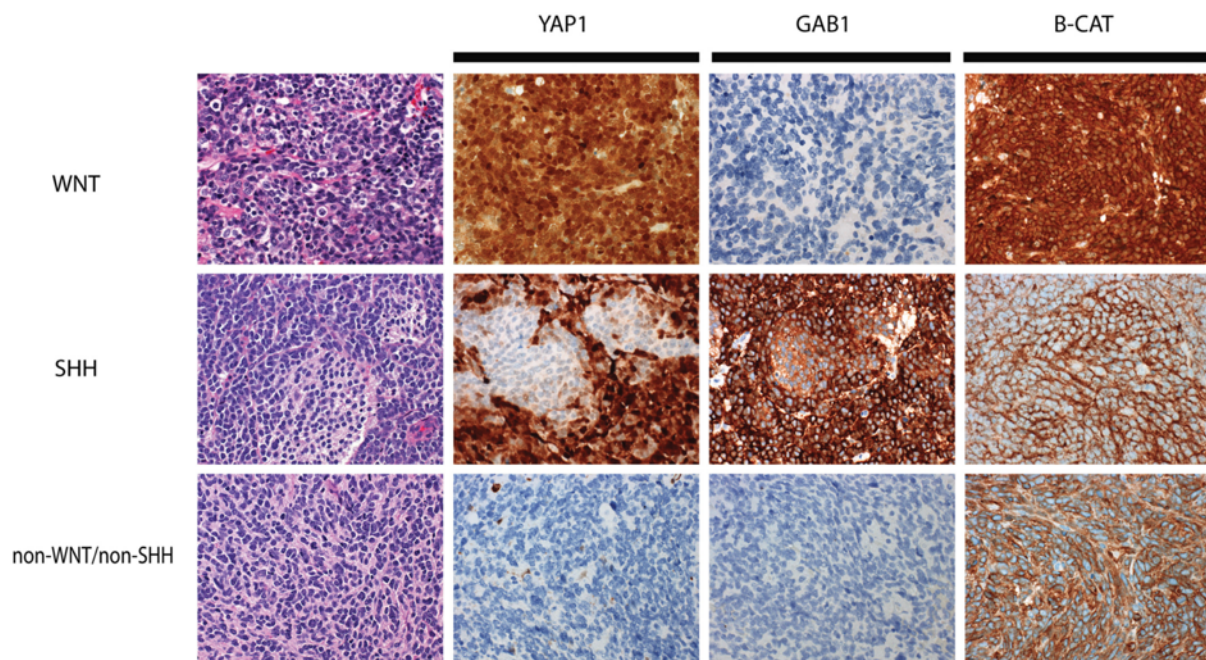
**Figure 1.17** – T1 weighted MRI of medulloblastoma (Carvalho Neto *et al.*, 2003). Medulloblastoma can be identified by the presence of high signal intensity (appearing white) in the cerebellum relative to healthy tissue.



## 1.7.6 - Determination of molecular subgroup

Given the clinical and prognostic significance of subgroups, it was necessary that assays were developed that could identify the molecular subgroup that a medulloblastoma tumour belongs to. Identifying subgroup (and their associated clinical features) subsequently informs clinical management for many patients and is of value to researchers for future studies wishing to further characterise the disease. There are several assays now developed, comprising immunohistochemistry, transcription, methylation, and sequencing methodologies, each with distinct advantages and disadvantages.

Immunohistochemistry assays are capable of discerning between the current WHO defined clinical groups of medulloblastoma – WNT, SHH and Non-WNT/non-SHH (Louis *et al.*, 2016). The assay is based upon expression patterns of 3 genes – *CTNNB1*, *GAB1*, and *YAP1*. This three gene assay is highly effective, where *CTNNB1* in both the cytoplasm and nucleus discerns WNT tumours from other types, and immunoreactivity for *GAB1* and *YAP1* and a lack of *CTNNB1* in the cytoplasm effectively identifying SHH type tumours (Figure 1.18). No reactivity is seen for non-WNT/non-SHH (i.e., Group 3 and Group 4) in any gene other than *CTNNB1* which is present only in the nucleus. IHC assays are often preferred in a clinical and research setting, owed to their low cost, compatibility with FFPE tissue and easy implementation (Ellison *et al.*, 2011). However, the lack of capability in discerning between Groups 3 and 4 is a significant limitation given the prognostic differences between the two subgroups and resulting subtypes and is generally not fit for purpose within a research setting.



**Figure 1.18** – ICH staining assay used to discern between WNT, SHH, Non-WNT/non-SHH medulloblastoma (Orr, 2020). WNT tumours display high immunoreactivity to *YAP1* and beta-catenin (*B-CAT*) and a lack of reactivity for *GAB1*. *SHH* tumours show high immunoreactivity to *YAP1* and *GAB1* and reactivity of beta-catenin is limited to the nucleus. Non-WNT/non-SHH show no reactivity in *YAP1* or *GAB1* and is the presence of beta-catenin is also limited to the nucleus.

The initial identification of molecular subgroups was based upon transcription profiling studies that utilised expression arrays (Cho *et al.*, 2011; Kool *et al.*, 2008; Northcott *et al.*, 2011). Northcott *et al.*, presented a NanoString assay, that measured the transcription levels of 22 genes to determine molecular subgroup – and presented very high classification rates and showed compatibility with FFPE derived DNA. However, as transcription-based assays are based upon the detection of RNA, which is inherently unstable, and the technology is not available readily in most clinical laboratories.

An effective alternative to transcription assays is the use of methylation arrays. The molecular subgroups of medulloblastoma were shown to display distinct methylation signatures, where tumours of each subgroup that share similar patterns of methylation cluster together and can be separated into class with high accuracy using supervised classifier models (Capper *et al.*, 2018). Methylation based classification can differentiate between all four molecular subgroups of medulloblastoma, and classifier models have been developed that can also identify subsequent molecular subtypes (Sharma *et al.*, 2019). Methylation assays have an added benefit in that they can also be used to detect chromosomal copy number aberrations, including focal gene amplifications such as *MYC* and *MYCN*. Methylation profiling is built upon the Illumina methylation array platform, and the accumulation of large datasets over the years due its cost compared to methylation sequencing assays has led to the accumulation of large reference cohorts across a range of paediatric tumour types. This has since culminated in the development of the DFKZ paediatric brain tumour classifier, which is used to diagnose the molecular subgroup of over 80 different paediatric brain tumour types. This capability has cemented methylation profiling via the array as the current gold standard within medulloblastoma diagnostics, but the platform is without limitations. Chromosomal copy number data is limited by the design of the array, and gene amplifications cannot be detected as reliably as karyotyping methods such as FISH across all genes due to uneven probe coverage. In rare tumour types like medulloblastoma it is also prone to long turnaround times, particularly in nations without centralised diagnostic infrastructures where samples cannot be pooled together into batches for analysis.

### **1.7.7 - Treatment: General protocols**

Over the decades, the general treatment strategy for medulloblastoma has become a trifecta approach, comprising maximal surgical resection followed post-operatively by a combination of radiotherapy and chemotherapy (unless otherwise specified in 1.7.7.1). Once the diagnosis of a cerebellar tumour is made clear, surgical resection serves as the initial therapy for patients. The goal of surgery is to safely remove as much of the tumour as possible, to relieve symptoms in the patient. Resection of the tumour also enables further diagnostic testing to be done on tumour biopsies. Patients who have the entirety of their tumours removed (gross total resection), are associated with improved overall survival compared to that of patients who do not (sub-total resection) (del Charco *et al.*, 1998).



After a patient has recovered from surgery, the patient is placed on a treatment regimen involving combination chemoradiotherapy followed by an extensive chemotherapy maintenance regimen. Typically, radiotherapy is administered for 7 weeks in conjunction with vincristine chemotherapy, to control micro-metastasis and locally control any residual tumorigenesis. Standard dosage of consists of craniospinal irradiation (CSI) to the entire neuroaxis at a dose of 23.4Gy, with a boost of 54Gy to the tumour bed in the cerebellum (Martin *et al.*, 2015). Adults often receive increased doses of CSI, receiving 36Gy, as the associated morbidity of CSI is not as high as that of children (Friedrich *et al.*, 2013). The administration of chemotherapy alongside radiotherapy in medulloblastoma patients has long since been shown to improve prognostic outcome in medulloblastoma patients (Tait *et al.*, 1990). The objective of chemotherapy is like that of radiotherapy, to control tumour growth and inhibit metastatic disease by controlling rapid cell division. A month after the initial 7-week chemoradiotherapy period, the typical standalone chemotherapy regimen for a standard risk MB consists of Cisplatin (75mg/m<sup>2</sup>), Lomustine (75mg/m<sup>2</sup>), Vincristine (1.5mg/m<sup>2</sup>) and Cyclophosphamide (1000mg/m<sup>2</sup>) doses over 44 weeks at varying combinations and doses.

### 1.7.7.1 – Assignment of risk

Historically, depending on demographic and prognostic factors, a patient with medulloblastoma is stratified into either “average” or “high” risk categories. All infants under the age of three are classed as “high-risk” patients, but due to the associated morbidity of radiotherapy are treated without it to minimise the risk of neurological defects. Patients over the age of three who have undergone a gross total resection or undergo sub-total resection with less than 1.5cm<sup>2</sup> remaining, as well as presenting with no signs of metastatic disease in the entire neuroaxis, are categorised as “average-risk” (Millard & De Braganaca, 2016). Patients who have undergone sub-total resection of the tumour with more than 1.5cm<sup>2</sup> remaining and/or are showing signs of disease dissemination are subsequently characterised as “high-risk”. These risk stratifications are based on prognostic evidence and dictate the resulting treatment that a patient will likely undertake. “Average-risk” patients are treated with the standard treatment protocols, whereas “high-risk” are considered for treatment escalation.

There have been calls for risk-stratification of patients to be refined in recent years, backed up by a wealth of new molecular evidence (Ramaswamy *et al.*, 2017). For example, the extent of resection is argued to not hold much, if any, prognostic significance, and that the molecular subgroup of a tumour is more informative of a patient’s chances of survival (Thompson *et al.*, 2016). In fact, the argument that surgical resection was a prognostic indicator was established back in 1999 and is at the very least severely outdated (Zeltzer *et al.*, 1999). Whilst the treatment strategies have yet to change, the clinical picture of risk stratification of medulloblastoma is now evolving to include molecular information (biomarkers, molecular subgroup). Such high-risk markers include the amplification of *MYC* or *MYCN* within a tumour, LCA histology subtype and *TP-53* mutation in SHH molecular subgroup tumours

(Crosier *et al.*, 2021). This information was compiled alongside histological data by the WHO in 2016 and 2021, producing a pioneering integrated method of diagnosis associated with new risk-stratifications (Louis *et al.*, 2016; Louis *et al.*, 2021). This has led to the classification of certain tumours as “Low-risk”, and patients that fall into these groups are now potential candidates for therapy de-escalation in clinical trials research. Whilst these new risk stratifications have not translated into altered treatment strategies just yet, it provides a unique window into the future, more personalised treatment landscape of medulloblastoma (Ramaswamy *et al.*, 2017).

### **1.7.7.2 - Complications of treatment**

The trifecta approach of surgery, chemo, and radio therapies to treat medulloblastoma has resulted in a relatively high survival rate for patients compared to that of other embryonal brain tumours (Packer, 2008). However, sequelae in survivors associated with the current treatments are significant, with patients suffering many short and long-term side effects. This problem is intensified in paediatric patients, for whom side-effects become a major obstacle to healthy brain development and long-term quality of life.

The invasive nature of surgical resection can damage tissue in the cerebellum, causing Posterior fossa syndrome. Posterior fossa syndrome affects ~25% of patients and impairs healthy brain function associated with the cerebellum. Symptoms present in the immediate days after surgery and include difficulties in speech, mutism, emotional instability, and ataxia (Doxey *et al.*, 1999). Patients appear to be more likely to develop this condition if the surgical resection is more aggressive, or if the tumour has begun to infiltrate or spread to the brainstem (Doxey *et al.*, 1999). A significant percentage of patients do not recover fully, with persistent speech impairment a significant problem (Robertson *et al.*, 2006).

As well as surgery, both chemotherapy and radiotherapy are associated with damaging side effects, causing significant impairment for patients. Patients can experience significant cognitive defects associated with radiation exposure to the brain and spine, of which is intensified if the dose is increased or if a patient is younger in age (Packer *et al.*, 2003). Radiation of the neuroaxis leaves patients at increased risk of numerous other conditions later in life, such as stroke, migraine, and other cerebrovascular diseases (Campen *et al.*, 2012). The effects of radiotherapy are intensified in younger patients, to the point where radiotherapy is avoided altogether in infants due to the associated damage to long term quality-of-life (Lafay-Cousin *et al.*, 2009). Neurotoxicity caused by chemotherapeutics are also an issue for medulloblastoma patients. Vincristine utilised significantly in standard medulloblastoma treatment protocols is associated with causing temporary neuropathy (nerve damage) in some patients (Starobova & Vetter, 2017). Cisplatin is also associated with causing side effects, particularly after breaching a cumulative dose of 300mg/m<sup>2</sup>. Like vincristine, neuropathies are

the main side effect for patients, but the recovery duration is longer than that of vincristine and is sometimes incomplete. Hearing loss is a major side effect associated with cisplatin, and its risk is highest in children and infants (Gurney *et al.*, 2014). General susceptibility to the side effects of chemotherapy, as well as recovery appears to be less in children compared to that of adults (Tabori *et al.*, 2005).

## 1.8 – Next generation sequencing & cancer research

The previous decades have seen significant advances in the field of DNA sequencing. The previous gold standard methods of chain-termination (Sanger) and shotgun sequencing used in the Human Genome Project in the early 2000's (International Human Genome Sequencing Consortium *et al.*, 2001, Venter *et al.*, 2001) have since been replaced by similar, but higher throughput methods, collectively termed 'Next Generation Sequencing' (NGS). These new methods are characterised by the capability to sequence large numbers of fragments in parallel, enabling the cost of sequencing a genome to be reduced from the billions spent on the Human Genome Project to prices as low as \$1000 per genome with the latest Illumina HiSeqX Ten platform.

There are multiple companies vying for market share in the NGS market (e.g., Illumina, Oxford Nanopore), and the technologies that they collectively offer can be divided into two iterations, short-read, and long-read sequencing. Short-read sequencing, or 'Sequencing by synthesis' involves the use of reversible fluorescent terminator dNTPs and PCR amplification of small DNA fragments ligated onto a sequencing chip into clusters (<300bp). Fluorescent imaging during each application cycle enables each fragment to be read at single base resolution, allowing for highly accurate alignments to be created (<0.1Kb error rate) (Kamps *et al.*, 2017). Short-read sequencing is the most frequently used sequencing technology today, offering a low cost per Gb. Illumina, the current market leader, claimed to have produced 90% of all sequencing data in 2015 and it is their platforms (NovaSeq) that are offering the most value for money in terms of cost per GB. In contrast, long-read sequencing enables reads of >2.5kb to be sequenced. There are several advantages to sequencing longer reads. Long reads are more easily aligned, enabling downstream analyses such as the detection of chromosomal rearrangement, DNA structural variation and epigenetic characterisation at lower levels of coverage than its short-read counterpart (Kamps *et al.*, 2017). PacBio offers this technology, termed 'single molecular real-time', or SMRT sequencing. However, it is currently associated with a higher error rate (>1Kb) and high cost per Gb, hindering its mainstream adoption in research. Oxford Nanopore technology takes long-read sequencing a step further enabling sequencing of reads >10Kb at a significantly lower cost but is associated with even higher error rates.

The utility of NGS has rendered it useful across many areas of research (e.g., phylogenetics, bio-synthetic gene cluster identification). This has since extended into the clinic where NGS of biopsies

(both tumour and liquid) have been used to screen for germline and/or somatic mutations in many diseases, including cancer. Gene panels generated by NGS analysis have shown utility in drug target identification (Sahm *et al.*, 2016) and chromosomal copy number variation can be detected with high resolution throughout the genome, even at low pass (Talevich *et al.*, 2016; Raman *et al.*, 2019). Its use in a research setting is now well established and it is expected to be used prominently in clinical diagnostics in the coming years. In the UK, this is already becoming a reality – Genomics England and Illumina have established a partnership to develop seven genomics laboratories within the National Health Service (NHS), with the aim of generating up to 500,000 whole genome sequences, ultimately with the goal of introducing whole-genome sequencing into routine diagnostics (Genomics England, 2020).

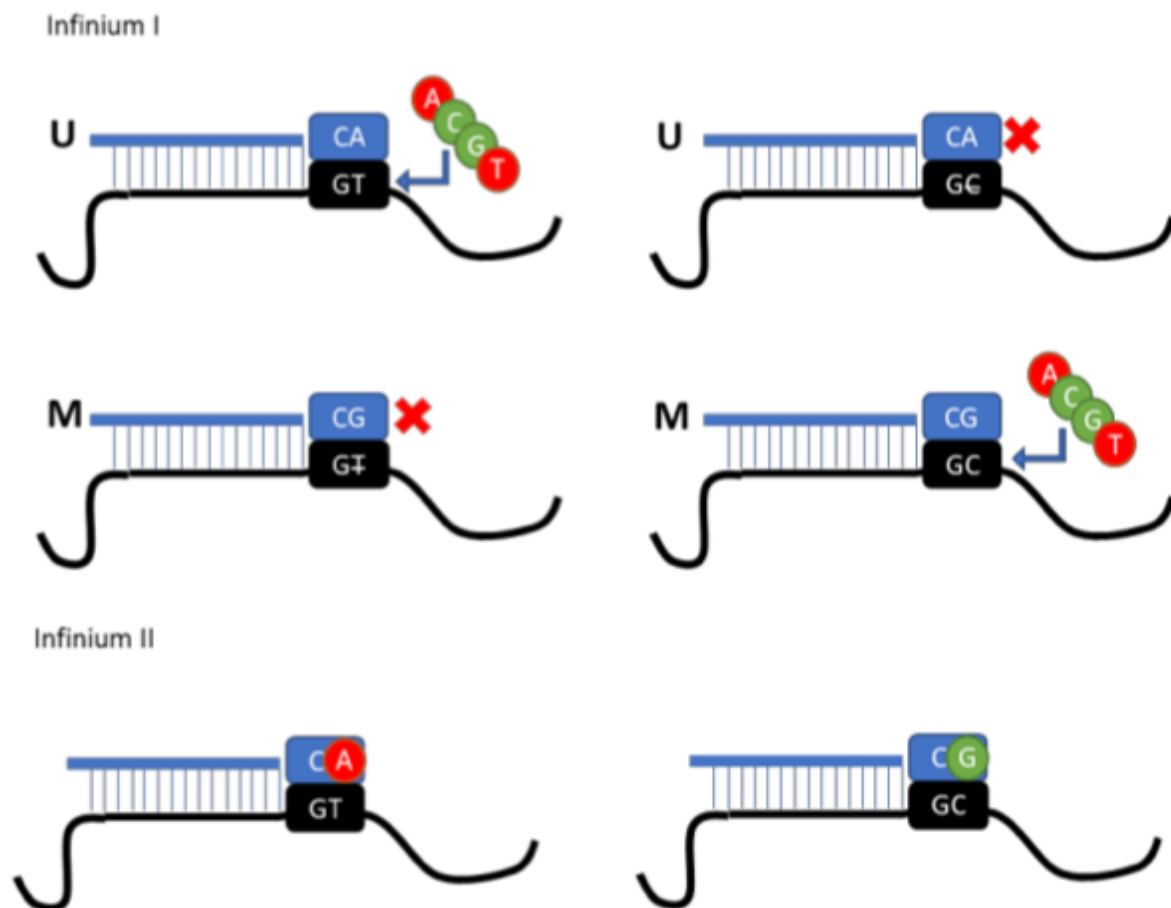
## 1.9 – Assessing DNA methylation

As mentioned in 1.7.6, the assessment of DNA methylation has formed a vital diagnostic component of subgroup detection in medulloblastoma and it can also be used to identify many more paediatric brain tumour entities (Capper *et al.*, 2018). The assessment of methylation – the addition of a methyl group (CH<sub>3</sub>) to the 5<sup>th</sup> carbon position of the cytosine base within DNA – is typically achieved via bisulfite treatment of DNA. Bisulfite treatment of DNA takes advantage of the different products that are obtained via de-amination of cytosine and 5-methylcytosine, of which cytosine is converted into uracil and 5-methylcytosine retains its integrity (Frommer *et al.*, 1992). This difference enables methylated residues to be distinguished from unmethylated positions, as the only remaining cytosines in a fragment of DNA are methylated. Since the 1990's several methodologies have been developed that assess DNA methylation. They can broadly be split into two types based upon the resolution at which methylation is interrogated – single-base resolution and region-based (Sun *et al.*, 2015)

Region-based methods comprise both MeDIP (methyl-DNA-immunoprecipitation) and MBD-seq (methyl-CpG-binding domain *MBD2* protein) technologies (Yong, Hsu & Chen, 2016). Bound regions are then sequenced, producing enrichment peaks throughout the genome, where the intensity of a peak corresponds to the level of methylation. MeDIP and MBD-seq utilise antibodies and methylate-CpG-binding protein to. Region based methods show good correlation to array-based methodologies and offer increased coverage throughout the genome. However, the method is not as high-throughput or can capture methylation information at as high a resolution as single-base resolution method.

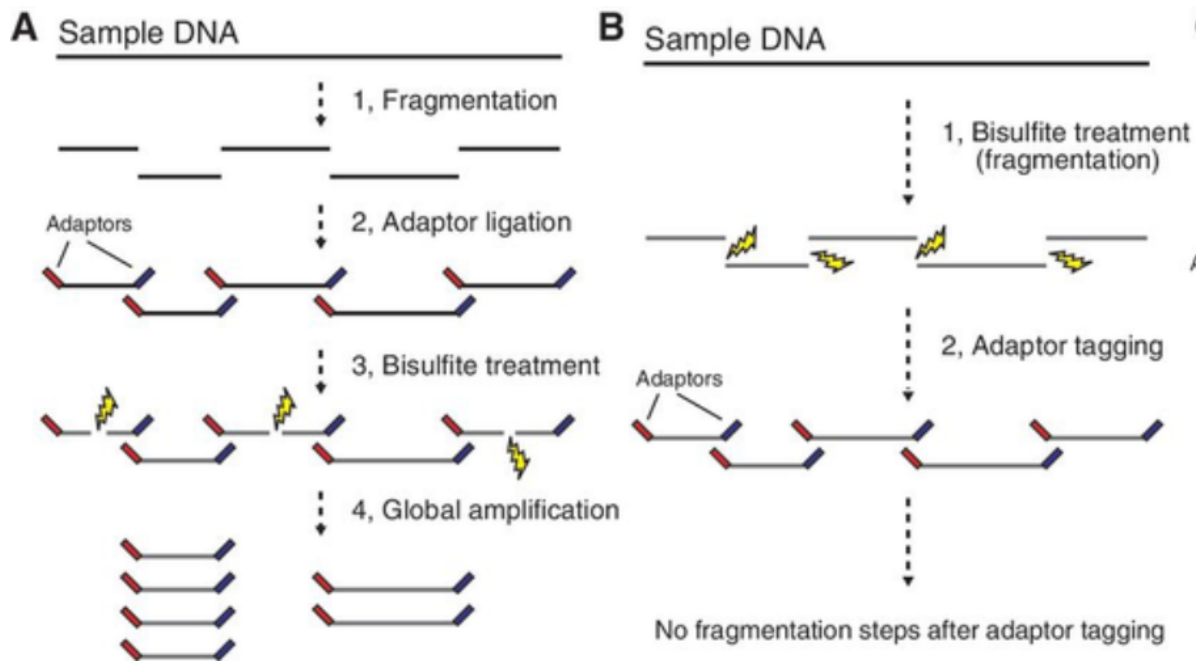
Single base resolution methods can be split into 2 types – array and NGS based. Microarray platforms like the Illumina Infinium series interrogate methylation status of CpG sites at many pre-determined probe sites, allowing methylation status to be determined throughout the genome. The sites that probe interrogate are selected carefully and are designed to encompass key regulatory and genic loci that are often implicated in methylation, including CpG islands, enhancer, and promoter regions (Bibkova

*et al.*, 2011). The latest iteration, the illumine Methylation EPIC chip, is designed to interrogate over 850,000 positions throughout the genome. The DNA methylation microarray is the platform of choice in methylation diagnostics currently due to its low cost and sample input requirements compared to sequencing-based approaches. All microarrays apply similar bead-based technologies to assess methylation. Split into two iterations – Type I and Type II. Each 50-mer probe (ligated to the array chip surface) sequence is designed to bind to bisulfite-treated DNA regions throughout the genome. At the 3' end of each probe is a CpG site. Following DNA hybridisation, single base extension with fluorescently labelled bases enable methylation to be quantified by measuring of fluorescence intensity. Two fluorescent labels are used to discern between methylated and unmethylated base positions – biotin for ddCTP/ddGTP nucleotides and 2,4-dinitrophenol (DNP) for ddATP/ddUTP respectively. Probe types differ based on the number of probes used to assess methylation at each CpG locus (Figure 1.19).



**Figure 1.19** – Difference between type I and type II probes on DNA methylation array chips (Bibkova *et al.*, 2011). Type I probes have two independent probe sequences, designed to assess unmethylated and methylated CpG status separately at a particular locus. Type II probes have just one probe for a CpG site and utilises a two-coloured bead approach to assess methylation status at the same locus. Both probes are useful – type I for interrogation of CpG-rich regions, where a more robust signal is required, and type II probes save space on array chips, assessing CpGs located outside of dense regions.

NGS based approaches supersede array technologies by surveying all CpGs of the methylome rather than a pre-determined amount. Methylation can be analysed at all CpG sites, including low-density regions, and distal regulatory elements. Various CpG contexts can also be interrogated – CpG, CHG and CHH context methylation. The gold standard NGS method of assessing methylation is whole genome bisulfite sequencing (WGBS), although its adoption until recently has been limited due to high costs associated with library preparation and sequencing. Various other methodologies have been proposed to circumvent many of the issues associated with NGS methylation sequencing – namely high economic costs and high input DNA requirements. Reduced-representation bisulfite sequencing (RRBS) utilises *Msp1* restriction enzyme digestion to shear DNA into fragments, followed by bisulfite treatment and sequencing to survey for methylation. *Msp1* recognises 5'-CCGG-3' sites, producing at the very least two cytosines per fragment. The development of RRBS was primarily initiated to maximise sequencing coverage in CpG dense regions and minimise costs by. Not sequencing the entire genome – representing a halfway point between array and whole genome-based approaches (Doherty *et al.*, 2014). In recent years, whole-genome bisulfite sequencing has also seen development that has made the platform more attractive for use. The issue associated with methylation sequencing is that large amounts of high-quality DNA is required for sequencing due to damage caused during bisulfite treatment. Post-bisulfite adaptor tagging (PBAT) places the bisulfite treatment step at the beginning of library preparation, avoiding the risk of bisulfite degradation of adaptor-tagged fragments (Figure 1.20). This has resulted in significantly reduced sample DNA requirements, reaching as low as 100ng (Miura *et al.*, 2012). New methodologies are also being that circumvent the need for bisulfite treatment. Enzymatic methyl-seq (EM-seq) utilise *TET2* and *APOBEC2* enzymes to first oxidise methylated CpGs followed by the conversion of unmethylated cytosines into uracil, removing the bisulfite treatment step entirely (Feng *et al.*, 2020). EM-seq has been shown to be a superior alternative to conventional WGBS – offering more amenable library preparation conditions, lowering DNA input and PCR amplification requirements. Crucially, the costs of bisulfite sequencing approaches are now also beginning to fall thanks to the optimisation of library preparation methods. The cost of low-pass sequencing (0.1 – 10x) is reaching comparable levels to microarray approaches - the current diagnostic gold standard (WGBS: ~£430 / Array: £300) and has since led to its increasing usage in research, including paediatric oncology in recent years (van Paemel *et al.*, 2020; Kuschel *et al.*, 2021).



**Figure 1.20** – Comparison between conventional WGBS and post-bisulfite adaptor tagging DNA library preparation approaches (Miura *et al.*, 2012). **A)** Conventional WGBS library preparation involves the bisulfite treatment of adaptor tagged fragments of DNA. This can result in the splitting of fragments due to bisulfite induced DNA damage, reducing the amount of viable DNA for sequencing. **B)** Post-bisulfite adaptor tagging treats bisulfite DNA with bisulfite in place of a sonication step prior to adaptor tagging. This significantly limits the further breakdown of DNA, producing much greater amounts of viable DNA for sequencing.

## 1.10 – Limitations of DNA methylation microarrays / advantages of methylation sequencing

Despite the advantages of microarray technology and its significant contributions to advancing the field of epigenetics in cancer research, the technique is not without its limitations. Microarray applications are limited in the regions of DNA that they interrogate by the number of probes used. The Methylation EPIC chip by Illumina has over 850,000 probes, deliberately designed to interrogate >95% of CpG islands and >90% of CpG shore regions, supplemented with a further ~300,000 sites designed to interrogate FANTOM5 enhancer regions (de Hoon *et al.*, 2015). Whilst comprehensive, the overall coverage throughout the genome is tiny. The 450k and EPIC arrays cover just ~3-6% of all ~28 million CpGs, ignoring a substantial amount of methylation data that may hold biological significance. Array platforms can be leveraged to derive copy number estimates via popular software packages like conumee (Hovestadt & Zapatka, 2017), but is inherently held back by the low number of probes, resulting in low resolution calling that impacts the ability of the platform to detect focal regions of amplification/deletion. Currently, the array remains the gold standard for methylation diagnostic, and is now the staple platform for subgroup determination in medulloblastoma. However, turnaround times



can be variable when not submitting samples in batches, which can often occur when working with rare tumour types (Perez & Capper, 2020).

Critically, the platform is proprietary, running the risk of being discontinued. As mentioned above, the Illumina methylation array platform has been iterated upon twice throughout its lifetime, with each new iteration not covering its predecessors design completely, affecting comparability. Illumina appears to anticipate a move into the methylation sequencing space, and now offer a methyl capture sequencing kit based upon RRBS that offers a superior alternative to the EPIC chip (Illumina, 2022). Methylation sequencing presents many superior solutions to all the disadvantages associated with the methylation array platform. The gold standard method, WGBS, offers whole genome coverage, interrogating nearly all ~28.3 million CpGs on a per-sample basis. Data is produced in the form of beta values like the array, meaning that in theory most analyses used on array data would be analogous to methylation sequencing data. Popular array analyses such as differential methylation and copy number calling could be performed at significantly greater resolution, offering the capability to detect additional regions of differential methylation and leverage sequencing reads to derive copy number estimates throughout the genome and not just probe sites.

Despite being the gold standard method of assessing methylation, its application in research has been limited due to high cost and sample input DNA requirements. However, recent advances like that of PBAT have reduced these barriers to entry significantly in recent years, and the price of low-pass WGBS is now reaching a level that is accessible to researchers. The platform presents significant untapped potential in methylation research, particularly within paediatric oncology where methylation is at the forefront of tumour analysis and diagnostics. RRBS of cell-free DNA has been used to classify a range of paediatric brain tumour types with high accuracy with minute amounts of DNA (< 10ng) (van Paemel *et al.*, 2020). Methylation profiling of paediatric brain tumours, including medulloblastoma samples via nanopore sequencing was used in conjunction with array data for classifier development – demonstrating a proof-of-concept that methylation sequencing would be amenable to current array-based diagnostic methods (Kuschel *et al.*, 2021).

## 1.11 – Summary & aims

Recent advances in bisulfite sequencing methods have subsequently removed many of the barriers that have limited its use in a diagnostic and research settings. Alternative library preparation methods such as post-bisulfite adaptative tagging have driven down costs of the sequencing significantly. Generating data at high coverage remains costly, but the costs of low-pass sequencing are now becoming comparable to per-sample costs of DNA methylation microarrays (10x WGBS = ~£430 / Array = £300). The DNA methylation microarray forms crucial component of subgroup diagnostics within medulloblastoma and indeed many other paediatric brain tumour types but is inherently limited by design and is continually subject to the risk of discontinuation as it is a privately owned platform. Low-pass WGBS offers a platform-free alternative of interrogating methylation throughout the methylome, and could, in theory, be used as a replacement for the array in a research and diagnostic setting. However, no consensus methodology currently exists for handling clinical samples sequenced using WGBS at low-pass, and the capability of the platform in being used to detect molecular subgroups of medulloblastoma and generate other clinically meaningful readouts such as copy number, mutation and differential methylation data has not been comprehensively assessed. Should such an assay for data handling and clinical analysis of WGBS in medulloblastoma be developed, it would offer a powerful alternative to the DNA methylation microarray, where WGBS samples could be used to predict subgroup and generate clinical readouts with whole-genome coverage at single-base resolution. Surveying the entire methylome also alludes to the possibility that additional biological heterogeneity could be identified, which may present opportunities to refine or optimise the current methylation profiles of subgroups/subtypes present within the disease.

This project aimed to investigate the capability of low-pass WGBS in being used to identify the subgroups of medulloblastoma and to generate clinically meaningful readouts that are associated with subgroup diagnostics compared to the current gold standard DNA methylation microarray platform, with the following specific aims:

1. Develop an optimised low-pass bioinformatic processing pipeline capable of producing both high quality sequencing read data that can be used for NGS-based analyses downstream, and non-missing methylation data through the introduction of a viable imputation strategy that demonstrates high correlation with array data in a matched medulloblastoma cohort (Chapter 3).
2. Investigate the degree of variability present throughout the medulloblastoma methylome and investigate the degree to which the current subgroups of medulloblastoma can be identified using low-pass WGBS both within and independent of medulloblastoma array cohorts. (Chapter 4)

3. Demonstrate that the molecular subgroups of medulloblastoma can be predicted with high accuracy by applying WGBS sample data to pre-existing array classifier models and show that WGBS trained models perform equivalently to like-for-like models trained using matched array samples. (Chapter 4)
4. Develop and optimise a strategy for estimating chromosomal copy number throughout the genome that performs equivalently to array-based methods in identifying large-scale events and demonstrates increased sensitivity calling smaller focal/single-gene changes (Chapter 5).
5. Test the ability for low-pass WGBS to be used to call variants using specialised BS-seq software and determine its utility in a diagnostic setting (Chapter 5).
6. Survey the landscape of differential methylation throughout the medulloblastoma methylome and provide an overview of the degree to which additional differential methylation can be detected using sequencing platforms compared to the array (Chapter 5).

## **Chapter 2 – Materials & Methods**

## 2.1 – Study Cohorts

Two cohorts – NMB and ICGC - comprising 70 (90%) primary medulloblastoma and 8 (10%) control cerebellum samples were used for the investigations reported within this study. As well as WGBS sample data, matched DNA methylation Microarray data was available for all samples.

Our NMB cohort comprised of 36 primary medulloblastoma samples, of which key characteristics are described below (Table 2.1). Samples were selected deliberately to ensure that subgroup composition reflected that of the wider literature, with all samples associated with high classification probability (>0.9) for either subgroup or subtype as per the MNP classifier. When possible, samples were chosen that harboured molecular features key to each subgroup (e.g., monosomy 6 in WNT, TP53/MYCN amplification in SHH). 27 (75%) males and 9 (25%) females were included, consisting of 18 (50%) classic, 3 biphasic classic (8%), 3 DN (8%), 7 LCA (19%), 1 MBEN (3%) and 3 with unknown histology (8%) respectively. Age at diagnosis was not a factor during cohort selection, where 15 (42%) were classed as infants (< 3 at diagnosis), 19 (53%) children (aged >3-15), 1 (3%) adult (> 15 years) and 1 (3%) with age unknown. M-stage according to Chang's staging system (Chang *et al.*, 1969) was available. M-stage was categorised into 2 categories: M-, encompassing M0/1 / M0 and M1 and M+ - encompassing M2 and M3 respectively. 26 cases (72%) were classed as M- and 10 (28%) classed as M+. Survival data was present for 35 (97%) samples, where a mean survival time of 3.325 years was reported. Relapse data was available, where 10 (28%) out of 36 samples relapsed later following disease progression. Molecular subgroup was determined via application of raw array sample data to the DFKZ paediatric brain tumour classifier (Capper *et al.*, 2018), where 5 (14%) WNT, 5 (14%) SHH, 12 (33%) Group 3 and 14 (31%) Group 4 subgroup classifications were assigned.

Sample ID	Sample Description	Gender	Age	Survival Status	Last Known Disease Status	Survival Time (Days) (PFS)	Histological Subgroup	Molecular Subgroup	Molecular Subtype	Sequencing Coverage	Tumour Stage	Relapse (Yes/No)	Relapse Type	Relapse Interval (Days)	MYC Status	MYCN Status
<b>NMB_17</b>	Primary Medulloblastoma	Male	4.697	Alive	N/A	<b>19.7 years</b>	Classic	Group 3	IV	10x	M0/1	No	N/A	N/A	Balanced	Balanced
<b>NMB_77</b>	Primary Medulloblastoma	Female	8.53	Alive	N/A	<b>10.613 years</b>	Classic	Group 4	VII	10x	M3	No	N/A	N/A	Balanced	Amplified
<b>NMB_79</b>	Primary Medulloblastoma	Female	3.438	Alive	N/A	<b>8.844 years</b>	Desmoplastic Nodular	SHH	SHH_Inf-I	10x	M3	No	N/A	N/A	Balanced	Balanced
<b>NMB_169</b>	Primary Medulloblastoma	Male	8.91	Deceased	Relapse	<b>0.272 years</b>	LCA	Group 3	II	10x	M0	Yes	Distant Recurrence	6	Amplified	Balanced
<b>NMB_178</b>	Primary Medulloblastoma	Male	4.95	Alive	N/A	<b>6.475 years</b>	Classic	Group 4	VII	10x	M3	No	N/A	N/A	Balanced	Balanced
<b>NMB_181</b>	Primary Medulloblastoma	Male	8.41	Deceased	Relapse	<b>0.144 years</b>	LCA	SHH	SHH_Child	10x	M0	Yes	Local Recurrence	196	Balanced	Amplified

<b>NMB_185</b>	Primary Medulloblastoma	Male	9.65	Alive	N/A	<b>9.92 years</b>	Classic	Group 4	VII	10x	M0	No	N/A	N/A	Balanced	Balanced
<b>NMB_187</b>	Primary Medulloblastoma	Male	3.7	Alive	N/A	<b>9.18 years</b>	Classic	Group 4	VIII	10x	M0	No	N/A	N/A	Balanced	Balanced
<b>NMB_203</b>	Primary Medulloblastoma	Male	6.24	Alive	N/A	<b>6.7 years</b>	Biphasic Classic	Group 4	I	10x	M0	No	N/A	N/A	Balanced	Balanced
<b>NMB_250</b>	Primary Medulloblastoma	Male	4.8	Alive	N/A	<b>7.15 years</b>	Classic	Group 4	VII	10x	M0	No	N/A	N/A	Balanced	Balanced
<b>NMB_273</b>	Primary Medulloblastoma	Male	5.71	Deceased	N/A	<b>5.71 years</b>	Unknown	Group 3	V	10x	M3	No	N/A	N/A	Amplified	Balanced
<b>NMB_318</b>	Primary Medulloblastoma	Male	5.91	Deceased	Relapse	<b>0.897 years</b>	LCA	Group 3	II	10x	M3	Yes	Distant Recurrence	19	Elevated (MLPA)	Balanced
<b>NMB_362</b>	Primary Medulloblastoma	Male	7	Deceased	Relapse	<b>1.753 years</b>	Desmoplastic Nodular	Group 4	VI	10x	M0	Yes	Distant Recurrence	180	Balanced	Balanced
<b>NMB_363</b>	Primary Medulloblastoma	Male	0.6	Alive	N/A	<b>4.938 years</b>	MBEN	SHH	SHH_Inf-II	10x	M0	No	N/A	N/A	Balanced	Elevated (MLPA)



<b>NMB_378</b>	Primary Medulloblastoma	Male	16.827	Alive	N/A	<b>6.241 years</b>	Classic	Group 3	IV	10x	M0	No	N/A	N/A	Balanced	Balanced
<b>NMB_390_FFPE</b>	Primary Medulloblastoma	Male	16.827	Alive	N/A	<b>2.636 years</b>	Unknown	WNT	WNT	10x	M0	No	N/A	N/A	Balanced	Balanced
<b>NMB_390</b>	Primary Medulloblastoma	Male	16.827	Alive	N/A	<b>2.636 years</b>	Unknown	WNT	WNT	10x	M0	No	N/A	N/A	Balanced	Balanced
<b>NMB_416</b>	Primary Medulloblastoma	Female	4.64	Alive	N/A	<b>6.914 years</b>	Classic	Group 4	VII	10x	M0	No	N/A	N/A	Balanced	Balanced
<b>NMB_433</b>	Primary Medulloblastoma	Male	1.894	Alive	N/A	<b>8.719 years</b>	Biphasic Classic	Group 3	IV	10x	M0	No	N/A	N/A	Balanced	Balanced
<b>NMB_436</b>	Primary Medulloblastoma	Male	4.97	Alive	Relapse	<b>3.917 years</b>	Classic	WNT	WNT	10x	M0	Yes	Local Recurrence	N/A	Balanced	Balanced
<b>NMB_529</b>	Primary Medulloblastoma	Female	8.11	Alive	N/A	<b>6.675 years</b>	LCA	Group 4	VII	10x	M0	No	N/A	N/A	Balanced	Balanced
<b>NMB_549</b>	Primary Medulloblastoma	Female	5.68	Deceased	Relapse	<b>2.202 years</b>	LCA	SHH	SHH_Child	10x	M0	Yes	Distant Recurrence	201	Balanced	Amplified

<b>NMB_610</b>	Primary Medulloblastoma	Male	8.875	Alive	N/A	<b>3.538 years</b>	Classic	Group 3	II	10x	M0	No	N/A	N/A	Balanced	Balanced
<b>NMB_725</b>	Primary Medulloblastoma	Male	5.75	Deceased	Relapse	<b>1.233 years</b>	Classic	Group 4	V	10x	M3	Yes	Distant Recurrence	N/A	Balanced	Balanced
<b>NMB_729</b>	Primary Medulloblastoma	Male	6.91	Deceased	Relapse	<b>1.13 years</b>	Classic	Group 3	I	10x	M0	Yes	Distant Recurrence	58	Balanced	Balanced
<b>NMB_733</b>	Primary Medulloblastoma	Female	6.67	Alive	N/A	<b>5.717 years</b>	Classic	Group 4	VII	10x	M2	No	N/A	N/A	Balanced	Balanced
<b>NMB_758</b>	Primary Medulloblastoma	Male	5.213	Deceased	N/A	<b>0.756 years</b>	LCA	Group 3	III	10x	M1	Yes	Distant Recurrence	34	Amplified	Balanced
<b>NMB_769</b>	Primary Medulloblastoma	Male	4.01	Deceased	Relapse	<b>N/A</b>	LCA	Group 3	III	10x	M3	Yes	Local + Distant	44	Amplified	Balanced
<b>NMB_772</b>	Primary Medulloblastoma	Male	7.93	Deceased	Relapse	<b>1.825 years</b>	Classic	Group 3	V	10x	M3	Yes	Unknown	323	Balanced	Balanced
<b>NMB_774</b>	Primary Medulloblastoma	Male	5.89	Alive	N/A	<b>4.961 years</b>	Classic	Group 3	III	10x	M0	No	N/A	N/A	Elevated (MLPA)	Elevated (MLPA)

<b>NMB_787</b>	Primary Medulloblastoma	Female	13.54	Alive	N/A	<b>4.822 years (ongoing)</b>	Biphasic Classic	Group 4	VIII	10x	M0	No	N/A	N/A	Elevated (MLPA)	Elevated (MLPA)
<b>NMB_803</b>	Primary Medulloblastoma	Male	2	Alive	N/A	<b>1.688 years</b>	Desmoplastic Nodular	SHH	SHH_Inf-I	10x	M2	No	N/A	N/A	Balanced	Balanced
<b>NMB_858</b>	Primary Medulloblastoma	Female	6.428	Alive	N/A	<b>1.030 years</b>	Classic	WNT	WNT	10x	M0	No	N/A	N/A	Elevated (MLPA)	Balanced
<b>NMB_867</b>	Primary Medulloblastoma	Male	6.556	Deceased	N/A	<b>2.636 years</b>	Classic	Group 4	VII	10x	M1	No	N/A	N/A	Balanced	Elevated (MLPA)
<b>NMB_870</b>	Primary Medulloblastoma	Female	11.144	Alive	N/A	<b>0.35 years</b>	Classic	WNT	WNT	10x	M0	No	N/A	N/A	Elevated (MLPA)	Elevated (MLPA)
<b>NMB_890</b>	Primary Medulloblastoma	Male	7.833	Alive	N/A	<b>3.325</b>	Classic	Group 4	V	10x	M0	No	N/A	N/A	Balanced	Balanced

**Table 2.1** - Clinical Demographics of NMB Cohort (n=36). Sample type, gender, age at diagnosis (in years), last known survival status, survival time (years) is shown. Histological subtype is provided in addition to molecular subgroup (WNT/SHH/Group 3/Group 4) and subtype (WNT / SHH child / SHH Infant subtype I and subtype II / Subtypes I – VIII of Group 3 and 4). Sequencing depth is provided in addition to relapse status, relapse interval (days) and MYC / MYCN amplification status. Missing data is denoted with N/A

A publicly available external cohort (n=42) comprised of 34 (81%) primary medulloblastoma and 8 (19%) control cerebellum samples was also used in this study (Hovestadt *et al.*, 2014). Data was acquired via the ICGC data portal (Zhang *et al.*, 2019) using the PBCA-DE project code. Gender information was available for all 42 samples – 24 (57%) males and 18 (43%) females were included. For our control samples, 4 were adult cerebellar samples from 26-, 41-, 33- and 87-year-old individuals and 4 were foetal cerebellar tissue samples (age < 0).

For the 34 primary medulloblastoma samples in this cohort limited additional clinical information was available, whose key characteristics are described below (Table 2.2). Histological subtype information was available for all 34 samples, consisting of 5 (15%) desmoplastic, 9 (26%) LCA and 20 (59%) classic tumours respectively. Age at diagnosis – split into infant, child and adult as above comprised 8 (24%) infants, 25 (74%) children and 1 (3%) adult case present. Chang's M-stage data was available, where 22 (65%) tumours were M-, 10 (29%) were M+ and 2 (6%) tumours had unknown M-status. Survival time at the time of data publish was available for 31 samples, where a mean survival time of 3.841 years was reported. Relapse data was also available, with 6 (18%) patients reported as relapsing. Molecular subgroup was determined was originally determined as previously described (Hovestadt *et al.*, 2013), but has since been updated with classifier calls via the DFKZ paediatric brain tumour classifier (Capper *et al.*, 2018), where 5 (15%) WNT, 6 (18%) SHH, 11 (32%) Group 3 and 12 (35%) Group 4 subgroup classifications were assigned.

Sample ID	Sample Description	Gender	Age	Survival Status	Last Known Disease Status	Survival Time (Days)	Histological Subgroup	Molecular Subgroup	Molecular Subtype	Sequencing Coverage	Tumour Stage (pM)	Relapse (Yes/No)	Relapse Type	Relapse Interval	MYC Status	MYCN Status
<b>ICGC_A1</b>	Normal Cerebellum - Adult	Male	33	N/A	N/A	N/A	N/A	N/A	NA	30x	N/A	N/A	N/A	N/A	NA	
<b>ICGC_A2</b>	Normal Cerebellum - Adult	Female	87	N/A	N/A	N/A	N/A	N/A	NA	30x	N/A	N/A	N/A	N/A	NA	
<b>ICGC_A3</b>	Normal Cerebellum - Adult	Male	26	N/A	N/A	N/A	N/A	N/A	NA	30x	N/A	N/A	N/A	N/A	NA	
<b>ICGC_A4</b>	Normal Cerebellum - Adult	Male	41	N/A	N/A	N/A	N/A	N/A	NA	30x	N/A	N/A	N/A	N/A	NA	
<b>ICGC_F1</b>	Normal Cerebellum - Foetal	Male	< 0	N/A	N/A	N/A	N/A	N/A	NA	30x	N/A	N/A	N/A	N/A	NA	
<b>ICGC_F2</b>	Normal Cerebellum - Foetal	Female	< 0	N/A	N/A	N/A	N/A	N/A	NA	30x	N/A	N/A	N/A	N/A	NA	
<b>ICGC_F3</b>	Normal Cerebellum - Foetal	Female	< 0	N/A	N/A	N/A	N/A	N/A	NA	30x	N/A	N/A	N/A	N/A	NA	

<b>ICGC_F4</b>	Normal Cerebellum - Foetal	Female	< 0	N/A	N/A	N/A	N/A	N/A	NA	30x	N/A	N/A	N/A	N/A	NA
<b>ICGC_MB1</b>	Primary Medulloblastoma	Female	6	Alive	Stable	<b>5.255</b>	Desmoplastic	SHH	SHH_Child	30x	M0	No		Balanced	Balanced
<b>ICGC_MB2</b>	Primary Medulloblastoma	Male	7	Alive	Stable	<b>8.841</b>	Classic	Group 4	VII	30x	M0	No		Balanced	Balanced
<b>ICGC_MB5</b>	Primary Medulloblastoma	Female	6	Alive	Complete Remission	<b>6.255</b>	Classic	Group 3	MBNOS	30x	M0	No		Balanced	Balanced
<b>ICGC_MB6</b>	Primary Medulloblastoma	Female	14	Alive	Complete Remission	<b>6.504</b>	Classic	Group 4	VI	30x	M3	No		Balanced	Amplified
<b>ICGC_MB7</b>	Primary Medulloblastoma	Male	9	Alive	Stable	<b>5.255</b>	Classic	Group 4	VIII	30x	M3	No		Balanced	Balanced
<b>ICGC_MB9</b>	Primary Medulloblastoma	Male	6	Alive	Complete Remission	<b>6.255</b>	LCA	Group 3	II	30x	M3	No		Balanced	Balanced
<b>ICGC_MB14</b>	Primary Medulloblastoma	Female	14	Alive	Stable	<b>3.419</b>	LCA	WNT	WNT	30x	M0	No		Balanced	Balanced



<b>ICGC_MB15</b>	Primary Medulloblastoma	Male	17	Alive	Stable	<b>1.918</b>	Classic	Group 4	VIII	30x	M0	No			Balanced	Balanced
<b>ICGC_MB16</b>	Primary Medulloblastoma	Male	3	Deceased	Relapse	<b>1.167</b>	Classic	Group 3	III	30x	M2	Yes	Local Recurrence	365	Balanced	Balanced
<b>ICGC_MB17</b>	Primary Medulloblastoma	Female	15	Alive	Complete Remission	<b>3.170</b>	Classic	PIN T, PB B	NA	30x	M3	No			NA	NA
<b>ICGC_MB18</b>	Primary Medulloblastoma	Female	13	Alive	Stable	<b>3.501</b>	LCA	Group 3	I	30x	M0	No			Balanced	Balanced
<b>ICGC_MB19</b>	Primary Medulloblastoma	Female	13	Alive	Complete Remission	<b>3.501</b>	LCA	Group 4	VII	30x	M0	No			Balanced	Balanced
<b>ICGC_MB24</b>	Primary Medulloblastoma	Male	7	Alive	Relapse	<b>3.085</b>	Classic	Group 4	VIII	30x	M0	Yes	Local Recurrence	487	Balanced	Balanced
<b>ICGC_MB26</b>	Primary Medulloblastoma	Male	8	Alive	Complete Remission	<b>5.255</b>	Classic	Group 4	VII	30x	M1	No			Balanced	Balanced
<b>ICGC_MB28</b>	Primary Medulloblastoma	Male	1	Alive	Complete Remission	<b>7.255</b>	Desmoplastic	SHH	SHH_Inf-II	30x	M0	No			Balanced	Balanced

<b>ICGC_MB32</b>	Primary Medulloblastoma	Female	8	Alive	Complete Remission	<b>7.923</b>	Classic	Group 4	VII	30x	M3	No			Balanced	Balanced
<b>ICGC_MB34</b>	Primary Medulloblastoma	Female	13	Alive	Relapse	<b>3.003</b>	LCA	SHH	SHH_Child	30x	M0	Yes	Local Recurrence	761	Balanced	Balanced
<b>ICGC_MB35</b>	Primary Medulloblastoma	Female	4	Alive	Complete Remission	<b>2.501</b>	LCA	SHH	SHH_Inf-NOS	30x	M1	No			Balanced	Balanced
<b>ICGC_MB36</b>	Primary Medulloblastoma	Male	4	Alive	Complete Remission	<b>3.003</b>	LCA	Group 3	II	30x	M3	No			Balanced	Amplified
<b>ICGC_MB37</b>	Primary Medulloblastoma	Female	1	Alive	Stable	<b>2.668</b>	Desmoplastic	SHH	SHH_Inf-II	30x	M0	No			Balanced	Balanced
<b>ICGC_MB38</b>	Primary Medulloblastoma	Male	6	Alive	Complete Remission	<b>2.252</b>	Desmoplastic	Group 4	VI	30x	M0	No			Balanced	Balanced
<b>ICGC_MB39</b>	Primary Medulloblastoma	Male	4	Alive	Complete Remission	<b>2.334</b>	Classic	Group 3	II	30x	M0	No			Balanced	Balanced
<b>ICGC_MB40</b>	Primary Medulloblastoma	Male	3	Alive	Stable	<b>2.753</b>	Classic	Group 4	VII	30x	M0	No			Balanced	Balanced

<b>ICGC_MB45</b>	Primary Medulloblastoma	Male	3	Alive	Stable	<b>2.668</b>	Classic	Group 3	VI	30x	M0	No			Balanced	Balanced
<b>ICGC_MB46</b>	Primary Medulloblastoma	Female	6	Alive	Stable	<b>5.337</b>	Classic	WNT	WNT	30x	M2	No			Balanced	Balanced
<b>ICGC_MB49</b>	Primary Medulloblastoma	Male	10	Alive	Relapse	<b>5.586</b>	Classic	Group 4	VIII	30x	M0	Yes	Local Recurrence	1248	Balanced	Balanced
<b>ICGC_MB50</b>	Primary Medulloblastoma	Female	10	Deceased	Relapse	<b>1.000</b>	Classic	Group 3	MBNOS	30x	M3	Yes	Local Recurrence	213	Amplified	Balanced
<b>ICGC_MB51</b>	Primary Medulloblastoma	Male	7	Alive	Stable	<b>1.501</b>	Classic	MB, G4	VI	30x	M0	No			Balanced	Balanced
<b>ICGC_MB88</b>	Primary Medulloblastoma	Male	2	Alive	Relapse	<b>2.252</b>	Desmoplastic	SHH	SHH_Inf-I	30x	M0	Yes	Local Recurrence	244	Balanced	Balanced
<b>ICGC_MB90</b>	Primary Medulloblastoma	Male	1.6	Alive	Stable	<b>1.836</b>	LCA	Group 3	VI	30x	M0	No			Balanced	Balanced
<b>ICGC_MB95</b>	Primary Medulloblastoma	Male	3	Alive	Complete Remission	<b>1.836</b>	Classic	Group 3	II	30x	M0	No			Balanced	Balanced

<b>ICGC_MB107</b>	Primary Medulloblastoma	Male	15	Deceased	N/A	<b>N/A</b>	Classic	WNT	WNT	30x	M3	No	Balanced	Balanced
<b>ICGC_MB112</b>	Primary Medulloblastoma	Male	7	Alive	N/A	<b>N/A</b>	Classic	WNT	WNT	30x	Unknown	No	Balanced	Balanced
<b>ICGC_MB113</b>	Primary Medulloblastoma	Female	14	Alive	N/A	<b>N/A</b>	LCA	WNT	WNT	30x	Unknown	No	Balanced	Balanced

**Table 2.2** - Clinical Demographics of ICGC Cohort (n=42). Sample type, gender, age at diagnosis (in years), last known survival status, survival time (years) is shown. Histological subtype is provided in addition to molecular subgroup (WNT/SHH/Group 3/Group 4) and subtype (WNT / SHH child / SHH Infant subtype I and subtype II / Subtypes I – VIII of Group 3 and 4). Sequencing depth is provided in addition to relapse status, relapse interval (days) and MYC / MYCN amplification status. Missing data is denoted with N/A.

## 2.2 – Sample preparation

Our NMB cohort was sequenced via low-pass WGBS sequencing (Section 2.1). 35 samples were from DNA extracted from fresh-frozen tissue and 1 was matched sample DNA extracted from formalin-fixed, paraffin-embedded tissue. An additional 3 FFPE samples were selected for sequencing but had failed library preparation due to a laboratory error by the sequencing company. A further 35 previously analysed matched methylation microarray samples were selected to ensure comparability at each stage of methylation data processing when applicable. Ethical approval was obtained for the collection, storage, and study of all material prior to this study by the Newcastle & North Tyneside Research Ethics Committee (REC: 07/Q0905/71). Wet lab work was conducted by Jemma Castle (laboratory technician) at Newcastle University.

To extract Genomic DNA from fresh-frozen tissue, the Qiagen DNeasy® blood and tissue extraction kit (Qiagen) was applied using the standard protocol provided by the manufacturer. DNA obtained from FFPE tissue blocks was extracted with the Qiagen QIAmp® DNA FFPE tissue extraction kit (Qiagen) using the manufacturer-supplied protocol. All extracted DNA was eluted in DNase/RNase free water. DNA prepared for sequencing was quantified using the NanoDrop 1000 Spectrophotometer to evaluate quality at the ratio 260nm/280nm. For previously analysed methylation microarray samples, double-stranded DNA was quantified using the Qubit® PicoGreen dsDNA broad range assay kit (ThermoFisher).

Following sample preparation, DNA was packaged and cold-shipped for whole genome bisulfite sequencing (Novogene). Prior to library preparation, agarose gel electrophoresis was applied by the genome sequencing company to assess sample integrity and purity. DNA quantity was checked a second time using a Qubit® fluorometer.

## 2.2.1 – Library preparation & sequencing

Library preparation and sequencing was conducted externally by Novogene. After DNA preparation, DNA was spiked with unmethylated lambda DNA to control for unbalanced base composition and fragmented into 200-400bp lengths using the Covaris® S220 Focused-ultrasonicator (Covaris). Sheared DNA fragments were then end-repaired, and bisulfite treated using the EZ DNA Methylation Gold Kit (Zymo Research). After bisulfite treatment, a bisulfite DNA library (selected fragment size = 150bp) was prepared using the Accel-NGS® Methyl-Seq DNA Library Kit (Swift Biosciences). Briefly, the library kit attaches a low complexity polynucleotide tail to both strands 1 and 2 in a 3-step reaction (Adapter ligation, size selection, PCR amplification). Following adapter tagging, an indexing PCR step is performed to increase DNA yield and incorporate full length adapters on both strands.

Following library preparation, DNA library concentration was first quantified using a Qubit® 2.0 fluorometer before being diluted to a concentration of 1ng/ul. Insert size was then checked using an Agilent® 2100 bioanalyzer (Agilent) and quantified a second time using Q-PCR. Sample libraries were then pooled and sequenced using the Illumina NovaSeq 6000 platform via a paired-end 150bp (PE150) strategy (Illumina).

## 2.2.2 – Sequencing data quality control (external)

Prior to data return, raw sequenced reads underwent basic quality control analysis and filtering by the sequencing company. Briefly, sequencing base quality value distribution (Phred score), error rates, GC content and Bisulfite conversion rates were calculated and tabulated for each sample. Reads with > 50% low quality nucleotides (Phred < 5) and/or > 10% N-content were discarded.

## 2.3 – Bioinformatic techniques

Throughout chapters 3, 4 and 5, I provide the relevant materials and methodologies associated with the bioinformatic analyses conducted within each chapter. A range of techniques are applied when concerned with data handling and clinical analysis of samples analysed by WGBS and DNA methylation microarray. Where necessary, the following section describes techniques in more detail, providing a background and understanding that underly their functional role in this study prior to their description in the relevant results chapter.

### 2.3.1 – Array data processing: normalisation

Methylation microarray methylation data (beta values) are calculated by estimating the ratio between both methylated (M) and unmethylated (U) signal intensities derived from each respective probe/beat type respectively:

$$M/(M+U)$$

Often a constant offset is applied to the denominator to avoid negative values following background adjustment. In ideal cases, a beta value of 1 (or 100%) would represent complete methylation, whereas a value of 0 (or 0%) represents a completely unmethylated probe site. However, various technical issues associated with the array prevent this from occurring. Variations in background fluorescence often occur, causing an additive error to occur across affected probes, which effectively reduces the range which beta values reach. This background signal is termed as the 'foreground' signal and attempts to correct beta value measurements for it are collectively termed 'background correction' methods and forms a crucial component of sample processing of array sample data.

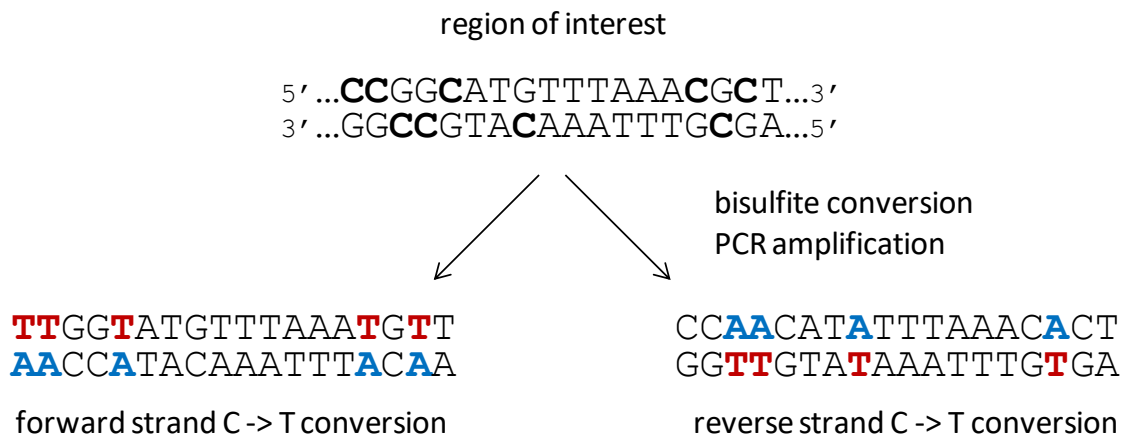
In this study, we apply normal-exponential out-of-band (noob) normalisation (Fortin, Triche Jr., & Hansen, 2017). This technique was deemed most suitable for this project as it allows for datasets to be processed separately, which was a necessity in this study given our use of an external pre-processed cohort. Noob normalisation corrects foreground signal by estimating mean background out-of-bound intensity using control probes via normal-exponential convolution model (Triche Jr. *et al.*, 2013). Out-of-bound intensity denotes the signal of fluorescence that is obtained from the opposite fluorophore at the binding of each probe site. Type I probes, which calculate methylated and unmethylated signal intensity separately, are used to measure this, which fluoresce at the same wavelength. To prevent any influence from test probes (which measure signal at biological loci), this correction is calculated from control probe sites, which denote probes that contain no CpGs that are deliberately included on the array chip for normalisation purposes (186 are included within the 450k array chip). Based on the mean intensity calculated from out-of-band probes, all probe channels can then be corrected accordingly.

Normalisation methods aim to correct experimental variability, but they also impact the variability in methylation seen as a result. Rather than reaching beta values that denote fully methylated or unmethylated sites (0-1), array derived beta values typically are limited to a range of 0.1-0.9, constituting an inherent limitation when using the platform.



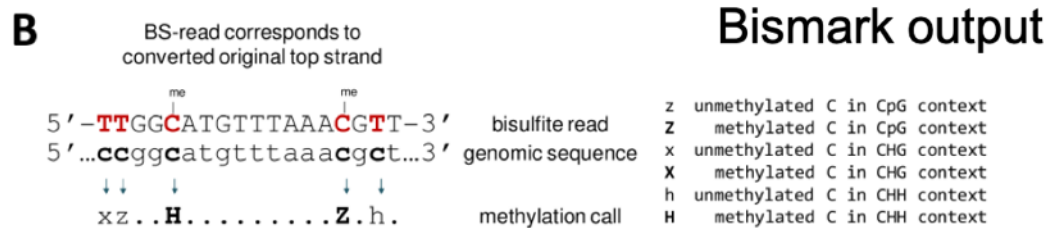
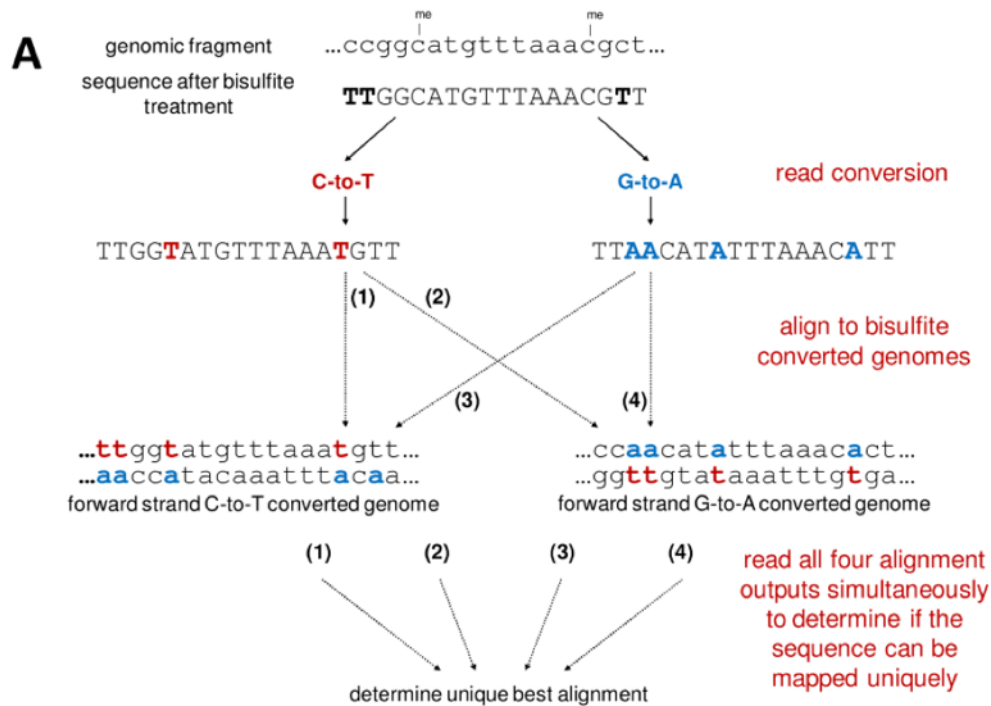
## 2.3.2 – Alignment of bisulfite treated DNA

Bisulfite treatment converts non-methylated cytosines into uracil, which is subsequently converted into thymine upon PCR amplification during library preparation. Methylated cytosines remain unaffected, resulting in the production of four DNA strands for each locus throughout the genome (Figure 2.1).



**Figure 2.1** - Bisulfite treatment of DNA produces up two different PCR products, resulting in the production of four different strands of DNA for each locus after amplification (Krueger & Andrews, 2011). Illumina sequencing protocols the original top (OT) and original bottom (OB) strands for sequencing.

Aligning bisulfite converted sequencing data is significantly more challenging due to the DNA code being reduced from four to 3 bases. The risk of incorrect bisulfite conversion presents the possibility that reads for each locus can exist in both methylation states, presenting an additional challenge. As shown in figure 1, reads taken forward for sequencing will correspond to converted versions of the original forward and reverse strands. Whether a strand is forward, or reverse is unknown, and alignment tools like Bismark typically establish this by performing four alignment processes in unison (Krueger & Andrews, 2011). Briefly, bisulfite reads are converted twice, producing C->T and G->A versions for each read (Figure 2.2). Each read is aligned to the reference genome (the human genome reference version hg19 was used in this project), which is also converted as above. All four alignments are outputted simultaneously, and unique alignments are subsequently outputted. By converting reads *in silico* and by performing four separate alignments, the best unique alignments can be derived irrespective of whether a read position is methylated or not, allowing for methylation values to easily be calculated accurately and extracted efficiently.



**Figure 2.2** - Bismark alignment process (Krueger & Andrews, 2011).

**A)** C->T and G->A converted reads(n=4) are aligned to C->T and G->A converted versions of the reference genome. Unique alignments for between all four strands (1+3 / 2+4) are used to identify unique alignments. **B)** Methylation state is identified by comparing aligned 'bisulfite reads' to the 'genomic sequence' Methylated and unmethylated cytosines can then be successfully extracted for all positions and are subsequently used for beta value calculation during methylation extraction.

## 2.3.3 - Imputation

Whilst methylation assays intend to provide measurements for all intended loci of interest (all probe sites for the array and all CpGs during WGBS), many measurements can be deemed as unreliable and are therefore removed from a sample of interest. For methylation microarrays, missing data is largely created following filtering of measurements deemed to be unreliable. Unreliable probe measurements are determined based on the calculation of detection p-values, which distinguish probe measurements from background signals (determined during background correction). As a result of the removal of unreliable measurements, methylation array data often contains missing values. Missing data can significantly affect analysis downstream, many of which do not tolerate missing data.

To deal with missing data, values are imputed. Imputation denotes the process of assigning replacement values for missing ones and various methodologies exist for imputing methylation values. The two methods used in this study are discussed further below.

### 2.3.3.1 – Imputation of array beta values via k-nearest neighbours

A popular approach for imputing missing data for methylation microarray cohorts is the k-nearest neighbour (kNN) algorithm. For methylation data matrices, where rows correspond to individual probes and columns correspond to samples, it is likely that there will be one or more missing values per row. As its name suggests, kNN considers neighbouring data points when inferring the value of any given missing value. In methylation datasets, neighbouring values correspond to neighbouring CpGs. An average value is calculated from them, and the subsequent output is used for imputation. In this study we use 10 neighbouring CpG values (5 either side). kNN is comparatively simple compared to other imputation algorithms but is quite effective when dealing with methylation data as neighbouring CpGs typically exhibit correlated methylation (Illumina, 2021). However, the algorithm is subject to limitations. Values cannot be imputed if neighbouring values are missing, and such an approach disregards any potential biological variability at a sample of interest.

---

### 2.3.3.2 – Imputation of WGBS beta values via BoostMe

Dealing with WGBS datasets sequenced at low pass produces sample sets with large numbers of missing values following processing. Due to matrix size dimensions exceeding 20 million rows, many well established imputation algorithms optimised for smaller matrices become unusable due to high processing requirements, even for HPC infrastructures. As an alternative, machine learning methods can be applied to resolve this issue, with Random Forest and XGBoost algorithms having been developed to handle missing methylation values from WGBS (Stekhoven *et al.*, 2011; Zou *et al.*, 2018).

XGBoost – or “Extreme Gradient Boosting” is used for supervised learning problems, where information is provided to ‘train’ the model to predict a target variable. The algorithm can be used for classification and regression-based predictions, the latter of which is used for the prediction of continuous methylation values ranging from 0-1. In this study we applied BoostMe (Zou *et al.*, 2018), which utilises XGBoost to impute missing values in WGBS datasets. BoostMe prepares a training data set consisting of several variables that are associated with each selected CpG from the original WGBS dataset to ensure highly accurate prediction of missing data. Prediction features are as follows:

- The nearest non-missing methylation values both up and downstream of the missing CpG of interest.
- The respective distance (in bp) to each neighbouring CpG.
- The average methylation value of the missing CpG calculated from all other samples where it is present.

For training, a regression model is fitted for each CpG and its associated predictors. This model is then applied to the same predictors where CpG values are missing. Regression based imputation has been shown to highly effective when concerned with methylation data, which typically exhibit short and long-range correlations – both of which can be characterised by regression (Lovkist *et al.*, 2016; Zhang *et al.*, 2019). For model training, BoostMe uses 1 million randomly selected CpGs above a pre-defined sequencing depth. This has been shown to be highly effective in predicting missing methylation data in WGBS cohorts, with excellent validation performance observed by the author in independent datasets (Zou *et al.*, 2018). However, like kNN, this model ignores any inherent variability that may be present. This concern is magnified in heterogenous datasets like that used in this study, where it is in our interest to capture as much variability as possible.

## 2.3.4 – Dimensionality reduction / visualisation techniques

Array and WGBS methylation datasets consist of many hundreds of thousands, to tens of millions of data points (beta values) per sample for each platform respectively. Working with datasets of this size presents significant challenges to downstream analysis. The scale of data points present often requires significant processing power to analyse and inferring any meaningful relationship based upon excessive amounts of data becomes more difficult. This problem is known as the ‘curse of dimensionality’, and methods to circumvent this are termed as “dimensionality reduction techniques”.

Dimensionality reduction techniques all work to reduce the size (dimension) of a dataset without sacrificing any variability within it. The use of these techniques is highly applicable to methylation datasets, where the vast majority of CpGs are concurrently methylated and do not contribute to any variability. Dimensionality reduction techniques are incredibly useful for ‘noise’ reduction, classification, and visualisation, reducing a dataset to only its most variable data points. For visualisation, dimensionality reduction aims to capture the variation found within a dataset and present it in 2 or 3 dimensions. By doing this, we can visualise our data on 2D or 3D scatter plots – providing insights into underlying data structures that are difficult to isolate at higher dimensions. The usage of these visualisation techniques is highly amenable to medulloblastoma, where subgroup clusters can be visualised effectively through the application of these techniques – which are described in more detail below.

### 2.3.4.1 – Principal component analysis

Principal component analysis (PCA) is one of the most common approaches used for dimensionality reduction and visualisation. The technique reduces dimensionality by transforming the original set of variables to a smaller set of ‘components’ that retains the greatest amount of variability from the original dataset as possible (Pearson, 1901; Hotelling, 1933). Each component is ranked in order of how much variance it captures – where the first component captures the greatest variance, with each subsequent component capturing decreasing amounts of variance. PCA is an effective algorithm to use when wanting to summarise variance present within a dataset and can capture phenotypic differences quite effectively. However, care must be taken when interpreting its output. Biological datasets are often complex and much of the underlying structure present between samples cannot be captured in a linear fashion. If the underlying structure of input data is not linear, clusters may not easily be identified. Although standardisation measures can be used to attempt to mitigate this, the goal of the PCA algorithm is to maximise variance, rather than identify clusters (Lever, Krzywinski & Altman, 2017), and may not be the most suited algorithm for visualisation if cluster identification is the intended goal.

### **2.3.4.2 – t-distributed stochastic neighbour embedding**

Where PCA comprises a linear approach to dimensionality reduction, there are numerous non-linear algorithms, collectively termed as ‘manifold learning’ techniques. Manifold learning techniques are inherently suitable for biological datasets, allowing for local structure present to be preserved between samples whilst maximising the distance between dissimilar groups. t-distributed stochastic neighbour embedding (t-SNE) is one such method and works by modelling the probability distribution of pairwise distances between high dimensional points (van der Maaten, 2008). The degree to which local structure is preserved can be changed (via the perplexity parameter), making it useful for optimising the degree of separation seen by the end-user. However, this also means that results must not be over-interpreted, as the separation observed may not accurately reflect the differences or similarities observed in the original dataset.

### **2.3.4.3 – Uniform manifold approximation and projection**

Uniform manifold approximation and projection (UMAP) seeks to preserve local relationships within a dataset whilst retaining global structure to a greater degree than t-sne (McInnes, Healy & Melville, 2018). UMAP estimates the topology (preserved structure) of high-dimensional data and uses it to present the data in lower dimensions, where similar data points generally cluster closer together. UMAP and t-sne are similar in the way that they can be used to identify the degree of similarity between neighbouring data points but differ when it comes to interpreting separation between clusters. The degree of distance between clusters identified by t-sne does not necessarily correspond to the degree of dissimilarity, but this is the case with UMAP where the degree of separation between points is generally indicative of dissimilarity. Both t-sne and UMAP are highly amenable to biological datasets particularly when one is interested in clustering and/or discovery, with the former used prominently as part of unsupervised clustering of the original DFKZ paediatric brain tumour classifier reference cohort (Capper *et al.*, 2018) and the latter used to define cell types in mixed population single-cell RNA-seq experiments (Wang *et al.*, 2018; Cao *et al.*, 2019).

### 2.3.5 – DNA methylation-based classification

The generation of large methylation datasets over numerous years of paediatric oncology research has led to various tumour entities and subgroups being identified, including the molecular subgroups of medulloblastoma. When coupled with machine learning techniques, these datasets can be leveraged diagnostically, where algorithms are trained to define the methylation profiles of reference samples based upon class labels (i.e., tumour type/subgroup). Trained algorithms can then be used to identify the class of a test sample (i.e., diagnostic tumour) into one of the defined labels. Machine learning algorithms are typically referred to as “classifiers”. Classifiers can be trained to serve different purposes. As seen in the DFKZ classifier, broad entity types and subgroups can be classified across a wide range of tumours (Capper *et al.*, 2018), whereas specialised models can be used to identify more subtle differences present within subtypes. Application of both types is seen in medulloblastoma, where the DFKZ classifier is used to assign molecular subgroup, and further molecular subtypes of groups 3 and 4 are defined using a separate model (Sharma *et al.*, 2019).

The diagnostic classifiers mentioned above utilise multiclass classification, and crucially use algorithms that assign an associated probability measure of class assignment. An associated class probability measure is important in a diagnostic setting, enabling clinicians to take the likelihood of assignment into account and make more informed clinical decisions (Simon, 2014). There are several classification algorithms that have been shown to be effective with DNA methylation data, with the more prominent being the random forest algorithm, upon which the DFKZ classifier is built (Capper *et al.*, 2018). Other algorithms include support vector machines (SVM), extreme gradient boosting (XGBoost), elastic net and logic regression, all of which are tested in this study.

Classifier performance (how effective it is in assigning class to unknown samples) is assessed both during training via cross-validation and by the application of a test set of sample data that is not involved in the training process. Cross-validation is a technique often employed to avoid over-fitting, a term used to describe a classifier that whilst performing well on the training set, the same performance cannot be recaptured on test data. The risk of overfitting is high when concerned with high-dimensional datasets such as methylation data, where many thousands of probes are used for training to discern between tumour types. In this project we apply k-fold cross-validation. K-fold cross validation partitions the training set into k number of folds, before proceeding to train the classifier on k-1 folds which is deliberately held out. The fold left out is then used for testing as a ‘validation set’, allowing performance to be evaluated for that fold. After performance measures are collected, the model is trained again until each fold has been held out and used for performance validation. The mean performance across all validated models is then taken forward as the overall performance of the trained model.



### **2.3.5.1 – Elastic net**

The elastic net algorithm incorporates multiple shrinkage regression methods (both lasso and ridge regression) – which works to limit sample variation and reduce the potential bias that a training set can introduce on a model (Friedman, Hastie & Tibshirani, 2010). The elastic net effectively applies a penalty term ( $\alpha$ ) to both regression methods to balance the degree to which the model limits variance (lasso regression) or mitigates bias (ridge regression) – with the aim of producing a highly generalisable model that is least prone to overfitting. Elastic nets can be adjusted using two parameters,  $\alpha$  and  $\lambda$ , both of which were tuned in this analysis. The degree to which each regression method is applied is controlled by the variable  $\alpha$ , where 0 corresponds to a complete ridge regression and 1 corresponds to complete lasso regression.  $\lambda$  corresponds to the level of shrinkage that is applied by each regression type. Classification algorithms built upon elastic net have been used to great effect in biological datasets, where the amount of variables often heavily outweigh the number of classes (Engelbrechtsen & Bohlin, 2019).

### **2.3.5.2 – Random Forest**

Random forests construct multiple decision trees, each comprising a random subset of features, with the feature most correlated with each class being used to split the node and form another branch. As the features used are randomly selected for each tree, each individual tree is typically a weak predictive model. However, random forests produce many thousands of trees, which are all used to derive a majority class vote. This aggregation of weak learners collectively creates strong predictions, forming models that are highly diverse and generalisable, mitigating the risk of overfitting. Parameters turned in this study were  $mtry$  (the number of randomly selected features for each node), and  $ntrees$ , the overall number of trees grown by the model. Random forest classifiers have been used with great success in epigenetics, where the current DFKZ paediatric brain tumour classifier is built upon the algorithm (Capper *et al.*, 2018)

### **2.3.5.3 – Extreme gradient boosting**

Extreme gradient boosting classifiers were trained using either boosted decision trees or linear function (Chen & Guestrin, 2016). Unlike random forests, which use an ensemble of separate weak models for aggregate prediction, gradient boosting fits weak learners sequentially, adding weight to the most predictive features within each model. The next model is then built upon these results, with the aim of continuously improving the model until a performance convergence is reached. Extreme gradient boosting can utilise decision trees, which are shallower (contain fewer branches/nodes) compared to the larger decision trees seen in random forest models, or via linear regression where an

elastic net-like approach is taken, where multiple separate linear regression models are build, with the aim of capturing the best correlation between each feature and specific class. Generally, linear boosted classifiers perform better in simple classification problems whereas trees perform better in more complex datasets. However, no such study comparing the two methods has been conducted in methylation datasets, and it is generally recommended to assess both algorithms to determine which is most suitable for each dataset.

### **2.3.5.4 – Support vector machines**

Support vector machines (SVM) distinguish between class by projecting training variables into a higher dimensional space, and then identifying a decision boundary (hyperplane) that discerns between classes. Variables which are located closest to those most associated with other classes are termed ‘support vectors’, which are used to guide the placement of the hyperplane and maximise separation between classes. SVMs are typically used for binary classification problems but do have application in multi-class problems, where models for every class combination pair are created and a classification is derived via majority vote. To project data into higher dimensional spaces, SVMs utilise kernels – a term to describe a collection of mathematical functions used for data transformation. In this study we tested classifiers that utilised linear and radial kernels. Linear kernels are the most basic kind and are often the best function to use when dealing with large numbers of features (Hearst, 1998). Radial kernels are often employed when the relationships present in the data are not always linear – both of which are amenable when concerned with methylation datasets. SVM classifier models can be tuned using cost and gamma parameters. Cost denotes the penalty value given to incorrect classifications – where higher cost values produce models that are highly biased to training sets. The gamma parameter controls the degree of influence for each specific decision boundary, where a low gamma value can separate variables that are located far apart.

### **2.3.5.5 – Logic regression**

Logic regression is an algorithm based upon adaptive regression, which aims to capture linear and non-linear relationships and determine how they interact between classes (Lu *et al.*, 2021). Regression algorithms by nature aim to reduce any relationship between predictors and class and are highly suited to binary classification problems or when dealing with data that is highly correlated to each specific class. Instead of drawing simplistic correlation, logic regression aims to find combinations of parameters that are highly correlated with a class of interest and assign Boolean statements to feature combinations (i.e., true/false), allowing for more nuanced relationships to be drawn between features and class rather than simplifying it. This collection of statements can then be applied to decision trees, allowing for large numbers of feature combinations to have an influence on the final class prediction.

This makes the algorithm inherently more suited to biological datasets, which often comprise binary predictors and the interaction between them is often complex and cannot be captured by simple regression easily. Beta values are binomially distributed, and those derived from the WGBS platform are majority 1s and 0s, rendering this data type highly suited to the logic regression algorithm.

### **2.3.6 – NGS and methylation array-based copy number estimation**

Alluded to in 1.10, methylation array data can be leveraged to estimate variations in copy number. As probes are typically interrogated by two probes, rather than calculate a ratio of intensity, the two intensities can be added together, and a ratio formed against that of a healthy reference set of samples that have a stable diploid genome. To increase resolution, probes are combined to produce bins – 50kb or higher to standardise coverage and reduce variability prior to segmentation. There are many tools available for copy number estimation with the DNA methylation microarray, of which ‘conumee’ has shown optimal calling performance in diagnostic tumour sample data compared to other algorithms (Kilaru *et al.*, 2020). Copy number segments are identified using circular binary segmentation (CBS) (Olshen *et al.*, 2004), a method which partitions a genome of binned ratio values by identifying statistically significant change points in mean copy number.

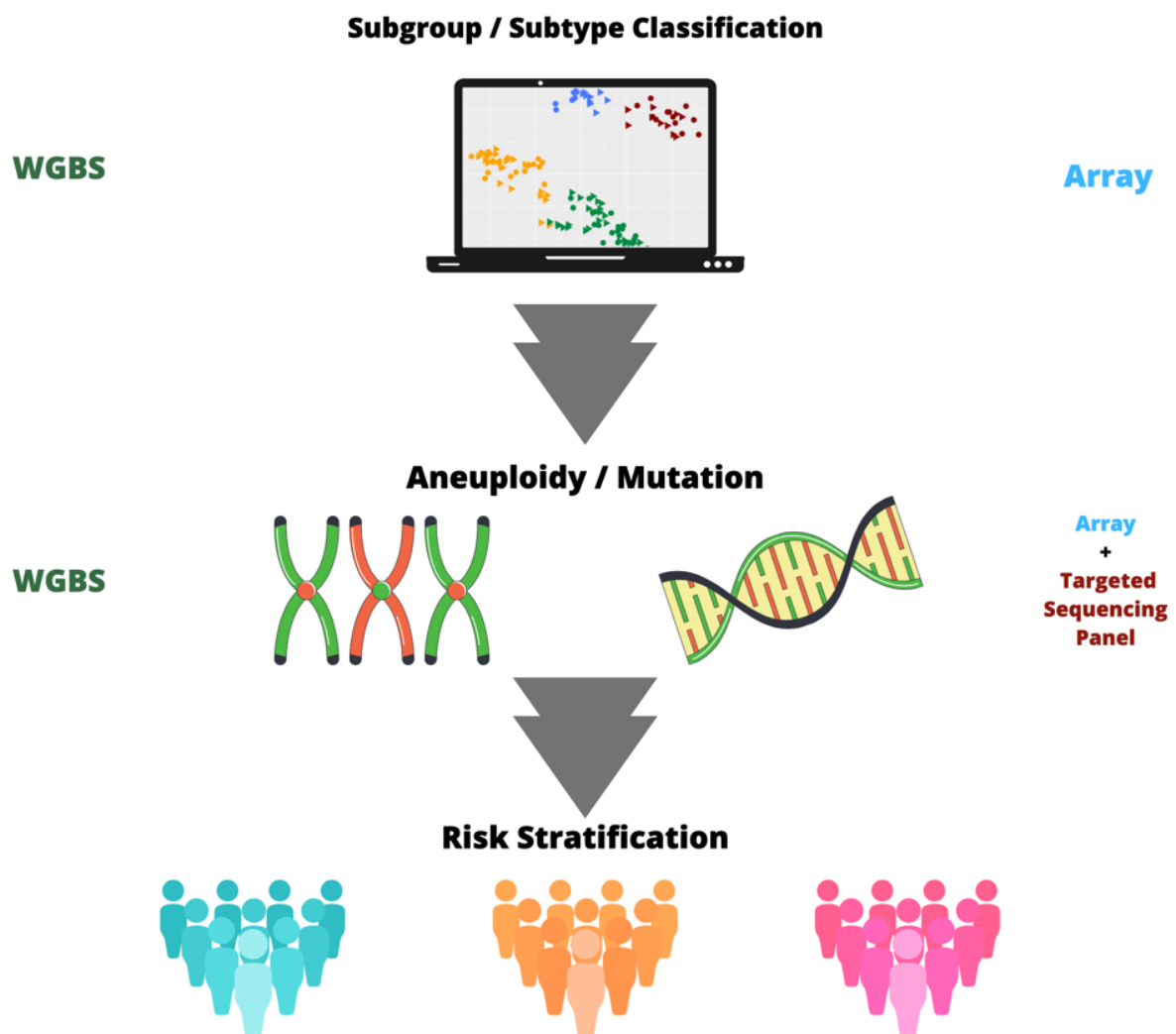
Copy number calling from sequencing data is segmented using CBS in this study via CNVKit (Talevich *et al.*, 2016). Where the methods differ is that rather than inferring a copy number ratio, they can directly infer copy number using the depth of coverage calculations from the many millions of sequenced reads generated. Segments can then be plotted graphically corresponding to genomic location, with higher ratios corresponding to areas of gain and low ratios corresponding areas of loss. Copy number estimation using the array provides a reasonable summary of large structure changes throughout the tumour genome. Where probes are concentrated, amplifications and deletions can be detected, although the resolution of these events is low, and this capability is lost in areas lacking probe coverage. NGS approaches to estimating copy number are generally more precise, with far greater numbers of data points throughout the genome from which to detect segment change-points.

**Chapter 3 – Establishment of a robust  
processing methodology for the clinical  
analysis of WGBS samples sequenced at low-  
pass**

### 3.1 – Introduction

DNA methylation is a robust epigenomic indicator of cell development and plays a key role in paediatric brain tumour classification, particularly for the designation of rare tumour types into known tumour classes for the identification of new tumour entities (Capper *et al.*, 2018). Current classification models are built upon the DNA methylation microarray platform, a cost-effective assay that interrogates up to 850,000 CpGs, including 99% of genes and 95% of CpG islands (Pidsley *et al.*, 2018). However, the platform is proprietary and therefore at risk of being discontinued. It is also prone to variable turnaround times when used diagnostically due to the specialist laboratory infrastructure required and costs can increase significantly when not submitting samples in batches for multiplex analysis (Perez & Capper, 2020). This can be problematic when dealing with rare disease entities like medulloblastoma, where a medical institution lacking a national, co-ordinated infrastructure will typically only see a few cases in any given year. The methylation microarray platform can also be used to derive low-resolution chromosomal copy number, an important prognostic feature of the molecular subgroups of medulloblastoma. Utilising the methylation microarray in conjunction with a targeted sequencing panel for subgroup specific variants enables both clinicians and researchers to obtain a definitive medulloblastoma subgroup diagnosis, with a range of additional aneuploidy and mutation data available to aid clinical risk stratification. This combination has elevated the microarray platform to its status as the current ‘gold-standard’ method for subgroup diagnosis in medulloblastoma and tumour classification across the field of paediatric neuro-oncology (Figure 3.1).

Whilst the diagnostic utility of the methylation microarray platform is clear, it is not a solution for the interrogation of genome-wide DNA methylation (the methylome). Whole genome bisulfite sequencing (WGBS) offers the potential to interrogate nearly all ~28.3 million CpGs present in the human genome at single-base resolution on a per-sample basis (Frommer *et al.*, 1992). This far exceeds the Illumina Infinium MethylationEPIC (EPIC) platform which, whilst its assessed loci are carefully selected, is restricted to just ~3% of total CpGs. The platform can also be used to conduct “read-based” analyses that are used in conventional NGS studies such as high-resolution aneuploidy detection and variant calling. WGBS therefore unites both NGS and methylation technologies within a single platform, offering a robust, platform independent means of conducting all the analyses that are required to reach a definitive subgroup diagnosis in medulloblastoma, supplemented further with a wealth of extra data of potential biological significance within the human methylome that is not assessable using current methods.



**Figure 3.1** - Standard medulloblastoma subgroup & patient risk stratification analytical pipeline. WGBS has the potential to be used for all 3 analytical steps involved in medulloblastoma risk stratification. Currently, methylation microarray and targeted gene sequencing panels are used in unison to achieve this.

Despite being established in the 1990s, the adoption of WGBS in a research setting has been limited due to sample input requirements and high financial cost. Only in recent years with the advent of alternative library preparation techniques such as post-bisulfite adaptor tagging (PBAT) (Olova *et al.*, 2018) has the need for prohibitively large amounts of input DNA been eliminated. This makes the technique a far more attractive proposition for use in clinical research, but the generation of high-coverage data using WGBS remains costly. However, the price of generating low-pass data (0.1-<10x) is now reaching a comparable price with that of the DNA methylation microarray platform (In 2021: 10x WGBS = ~£430 / Array = £300). Utilising low-pass sequencing is at odds with many of the current sequencing guidelines for WGBS and NGS in general. For example, the ENCODE project sequencing project guidelines (2017) recommend that samples sequenced using WGBS be at a minimum depth of 30x to ensure that any methylation measurements are reliable and well correlated with any

sequenced replicates. Although none would disagree, the economic and clinical need to reduce costs has encouraged researchers to work with low-pass data to attempt to generate assays that can provide robust data of clinical benefit. The use of low-pass NGS has grown significantly in popularity over the past decade and has generated important results. In 2015, a landmark paper published by the 1000 Genomes Project Consortium demonstrated that whole genome sequencing (WGS) could be used to obtain accurate aneuploidy measurements at just 7x coverage (Auton *et al.*, 2015). This has since driven magnitudes lower, with copy number variants being detected at depths as low as 0.25x (Dong *et al.*, 2017). Low-pass WGBS has also been used to successfully call regions of differential methylation and copy number at 5x (Laufer *et al.*, 2021). The evidence that low-pass NGS technologies can be used to generate clinically useful data combined with the potential of WGBS as a platform independent means of interrogating the methylome provided a rationale that low-pass WGBS could feasibly be used as a means of subgroup determination in medulloblastoma.

Currently, no consensus methodology for processing low-pass WGBS *in-silico* currently exists in the literature, and many current clinical papers that utilise the technique do not address a major drawback of working with low-pass data – data missingness. When dealing with individual samples, CpGs missing at random are not much of a concern as sample-specific analyses like aneuploidy detection can still be performed without issue. However, the problem is compounded significantly as cohort sizes increase, causing a drastic decrease in the number of measured CpGs that are common across all samples. For example, a study involving low-pass WGBS of *A.thalina* showed that over 92% of CpGs were missing in at least one sample (Kawakatsu *et al.*, 2016). Many of the analyses that typically comprise an epigenetic study, particularly those involved in subgrouping studies such as clustering, and classifier model training become impractical when working with datasets littered with large numbers of missing values. This is because most algorithms involved cannot tolerate missing data, meaning only CpGs measured in common can be analysed. As data missingness increases, the number of CpGs that can be analysed decreases, causing significant information loss and the potential for bias in any derived data. One means to circumvent this issue is to impute any missing data. imputation (explained further in 2.3.3) is defined as the process of replacing missing data with computed values. More complex imputation algorithms can impute data that effectively mimics the binomial distribution of methylation beta values (0 = non-methylated, 1 = methylated), by considering known data from other samples within the input dataset. Imputation is now a common part of handling missing data when working with methylation microarray data but is computationally expensive as algorithm complexity and input data size increases. This becomes prohibitive for researchers that lack high performance computing (HPC) infrastructure when wishing to apply the same methodologies to matrices that contain methylation data for every CpG in the methylome (~28,000,000 vs ~850,000). In recent years, researchers have begun to develop algorithms that are more suited to the WGBS platform, enabling samples sequenced at lower levels of coverage to be feasibly used in research studies (Zou *et al.*, 2018; Stekhoven *et al.*, 2012; Taudt *et al.*, 2018). However, the usage of imputation



in WGBS research is limited outside of the software creators' publications and has not been comprehensively tested in clinical studies.

In summary, WGBS offers a means of interrogating the entire methylome at single-base resolution, eclipsing the coverage offered by methylation microarray. Previous barriers to entry are now beginning to be removed, with low DNA input and economic costs now becoming feasible for clinical research at low pass. Studies involving low-pass NGS data have successfully collected data of clinical benefit, with many assays now published that are both a cheaper and faster alternative than current assays used in the clinic today. The development of a robust low-pass WGBS processing methodology for the clinical analysis of WGBS has yet to be established within paediatric oncology, and many of the associated concerns with low-pass methodologies such as missingness and defining the lowest acceptable level of coverage is have not been addressed. The DNA methylation microarray platform remains the gold-standard method of classifying the molecular subgroups and subtypes of medulloblastoma. Many of the drawbacks associated with the platform could be mitigated through the development of an assay based upon WGBS, which would offer a powerful platform-independent alternative. Data derived from the platform would be compatible with existing DNA methylation microarray-based classifiers and could be used to generate high-resolution copy number data as well as variant calling on a per sample basis. In addition, surveying the entire methylome may elucidate additional epigenetic heterogeneity present between subgroups, which may further aid the identification of difficult to classify samples or refine and/or identify further subtypes present within the disease.

## 3.2 – Aims

The primary analytical goal of this chapter was to optimise a robust processing methodology for low-pass WGBS, developed using a combined cohort of low-pass, internally sequenced (NMB) and high-depth externally sequenced (ICGC) primary medulloblastoma and cerebellar tissue samples. As well as generating high quality sequencing data, we aimed to develop an effective methylation data processing pipeline that includes imputation, enabling downstream epigenomic investigations such as clustering and classification to be conducted without being hampered by the ever-present issue of missing data when working with low pass data. In this chapter, a detailed analysis of imputation performance in low-pass WGBS sample data is described and it is demonstrated that methylated CpG methylation values, calculate using low-pass WGBS, are highly correlated to corresponding probe CpGs located on the methylation microarray, using a matched cohort of samples. Finally, using randomised proportional downsampling, a comprehensive analysis of the effect of lowering coverage on both datasets down to 1x was undertaken, providing valuable insights into the quality of sample data returned as sequencing depth decreases.

## 3.3 – Materials & Methods

### 3.3.1 – Analysis Overview

	<b>Data Format</b>	<b>Analytical method : Software used</b>	<b>URL</b>
1.	<b>Raw Reads (.fastq)</b>	<b>Quality Control :</b> <i>FastQC</i> <b>Read Trimming :</b> <i>TrimGalore!</i> <b>Plotting :</b> <i>MultiQC</i>	<a href="http://bioinformatics.babraham.ac.uk/projects/fastqc">bioinformatics.babraham.ac.uk/projects/fastqc</a> <a href="http://bioinformatics.babraham.ac.uk/projects/trim_galore">bioinformatics.babraham.ac.uk/projects/trim_galore</a> <a href="http://multiqc.info">multiqc.info</a>
2.	<b>Aligned Reads (.bam)</b>	<b>Alignment :</b> <i>Bismark</i> <b>Deduplication :</b> <i>Bismark</i> <b>Methylation Extraction :</b> <i>Bismark</i>	<a href="http://bioinformatics.babraham.ac.uk/projects/bismark">bioinformatics.babraham.ac.uk/projects/bismark</a>
3.	<b>Methylation Ratio (.txt)</b>	<b>Quality Control :</b> <i>RnBeads 2.0</i> <b>CpG Filtering :</b> <i>RnBeads 2.0</i> <b>Beta Value Matrix Generation :</b> <i>RnBeads 2.0</i>	<a href="http://rnbeads.org">rnbeads.org</a>
4.	<b>Beta Values (.csv)</b>	<b>Data Handling :</b> <i>BSseq</i> <b>Imputation :</b> <i>BoostMe</i>	<a href="http://hansenlab.org/software.html">hansenlab.org/software.html</a> <a href="https://github.com/lulizou/boostme">github.com/lulizou/boostme</a>
5.	<b>Analysis-Ready Beta Values (.csv)</b>	<b>LiftOver :</b> <i>MethylLiftover</i> <b>Interplatform Correlation :</b> <i>ggplot2</i> <b>Plotting :</b> <i>ggplot2</i>	<a href="https://github.com/Christensen-Lab-Dartmouth/methylLiftover">github.com/Christensen-Lab-Dartmouth/methylLiftover</a> <a href="http://ggplot2.tidyverse.org">ggplot2.tidyverse.org</a>

**Figure 3.2** – Graphical overview of all conducted analyses and associated software (including relevant download links) used within chapter 3. Further details (e.g., references, parameters used) are provided in subsequent sections.

## 3.3.2 – Raw data processing – WGBS

### 3.3.2.1 – Computing resources/software

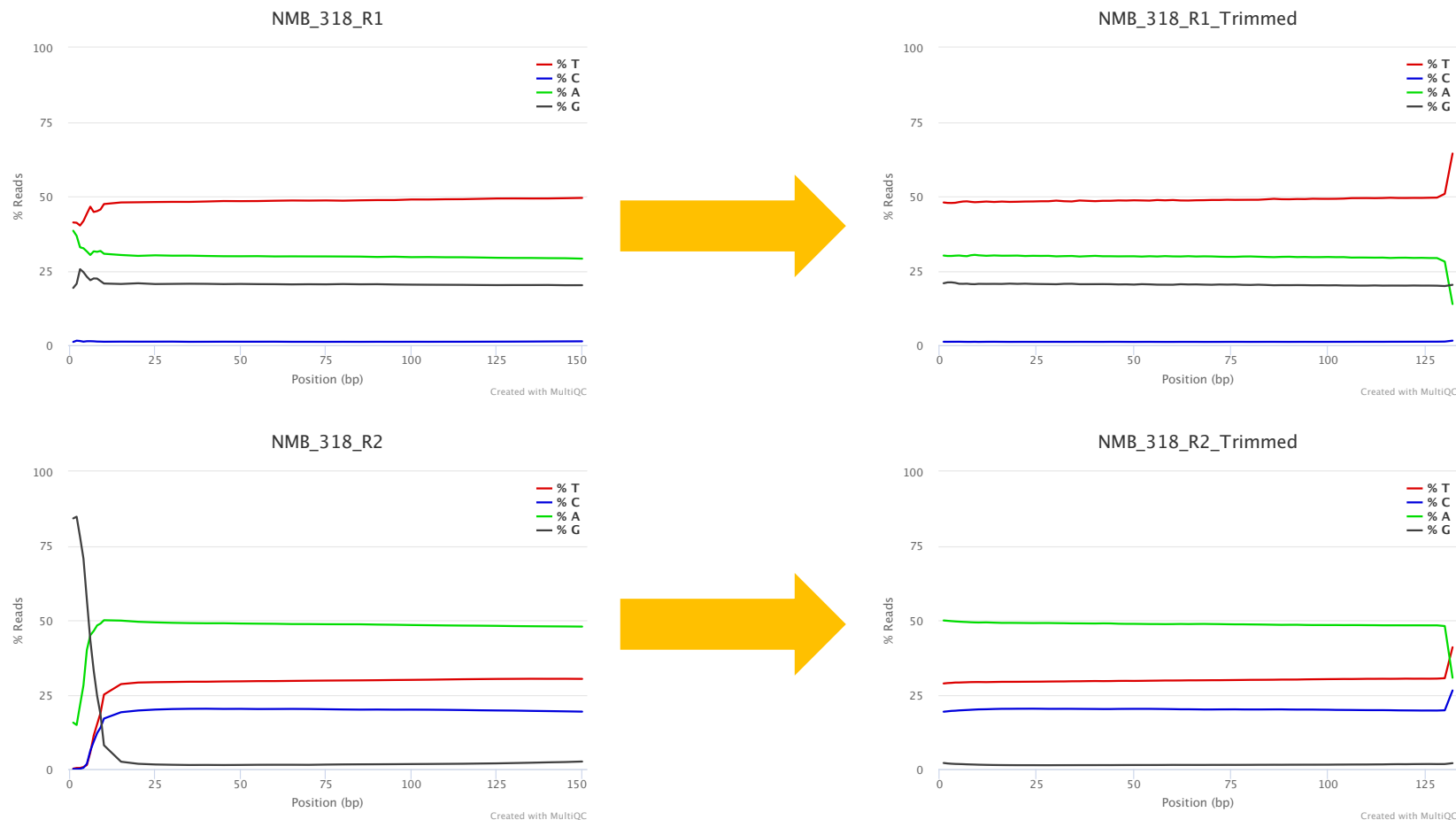
Upon data return, raw sequencing data for all samples was transferred to the Newcastle University secure file store. All subsequent analyses, detailed in section 3.3.5, were conducted using the ROCKET High Performance Computing service (Newcastle University). The ROCKET hardware infrastructure consists of 123 compute nodes in total, with the largest compute nodes consisting of 56 cores (27.4GB per core) and 8.7TB scratch space. A shared main data file store of 500TB is available to users. Unless stated otherwise, all software was installed and executed within the Linux CentOS 7 operating system. All scripts were written using VIM 8.2.

### 3.3.2.2 – Quality control & read trimming

Raw fastq files were first subjected to internal quality control using FastQC version 0.11.8 (Andrews, 2010). Raw data statistics and plots from all modules (including true sequencing coverage determination) were generated using MultiQC and tabulated. Quality trimming was then conducted using TrimGalore! 0.6. and Cutadapt version 1.18 (Martin, 2011) using the following filters:

- Quality Phred score cut-off (-q): 20
- Adapter sequence (auto detected): 'AGATCGGAAGAGC' (Illumina TruSeq)
- Maximum trimming error rate: 0.1 (default)
- Minimum required overlap: 1bp
- Minimum required sequence length for both reads after trimming: 20bp

Both reads 1 and 2 were trimmed by 17bp (5' = 10bp, 3' = 7bp) after analysing the distribution of base sequence content of both reads in all samples and considering the trimming recommendation that 10bp be cut from both reads at the 5' end by the sequencing company (Figure 3.3).



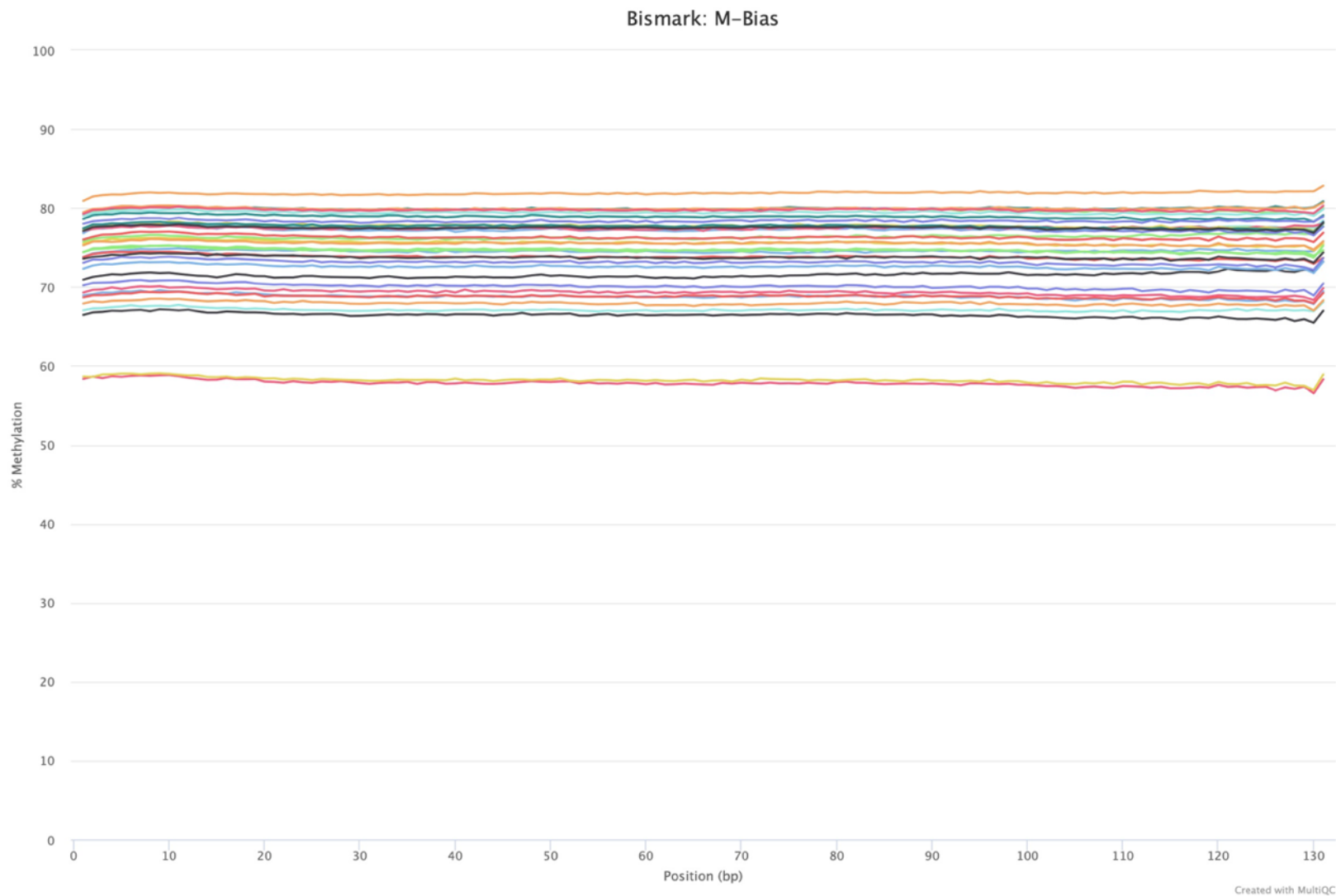
**Figure 3.3** - Effect of quality trimming on per-base sequence content distribution for NMB\_318. FastQC per-base sequence content plot for raw reads 1 and 2 are shown on the left-hand side. Both plots highlight unstable base composition within the first 25bp of both reads, including a significant G bias in read 2. Both plots on the right show the resulting per-base sequence content of each respective read following trimming. Base content has stabilised across all base pairs at the 3' of reads, and a slight bias introduced because of strict adapter trimming at the 5' end.

### **3.3.2.3 – Alignment and deduplication**

Following quality control, trimmed reads were then aligned using Bismark version 0.22.1 (Krueger & Andrews, 2011) (Section 2.3.2). Bismark was run in conjunction with Bowtie 2 (Langmead & Salzberg, 2012) against a bisulfite converted hg19 reference genome. Alignment was run with zero permitted mismatches for all samples. After successful alignment, aligned reads were deduplicated using Bismark.

### **3.3.2.4 – Methylation bias assessment & methylation data extraction**

To check for the presence of any remaining technical biases in the data, the methylation proportion across each possible read position was determined via methylation bias assessment using Bismark. A small bias was identified at the 3' of read 2 which was subsequently ignored during methylation extraction (Figure 3.4). Methylation count data was then extracted into bedGraph, bismark.cov and CpG\_report.txt formats for downstream analyses for all samples.



**Figure 3.4** - Observed methylation bias identified in read 2 of all samples at the final 2 bases of the 3' end. This methylation bias corresponds to the base composition bias introduced at the 3' of reads following adapter trimming in section 3.3.1.2.



### 3.3.3 – External cohort data handling

Prior to further processing, both external WGBS and array cohorts were downloaded via the ICGC Data Portal (Zhang *et al.*, 2019). 42 samples sequenced using high-depth WGBS with matching methylation microarray data were available (Hovestadt *et al.*, 2014) under the project code PBCA-DE and downloaded (Section 2.1). Externally downloaded methylation array data was available as a matrix of pre-processed beta values – the final format following methylation data processing. Both external WGBS sequencing data and Array data were processed as previously described (Hovestadt *et al.*, 2014).

Externally acquired WGBS sample data (n=42) was downloaded in the ICGC Data Portal submission file format. The format is tab-separated and has 26 columns, the majority of which is project and sample information. Within the format is the key experimental information pertaining to processed WGBS data – Chromosome number, start position, end position, strand, methylated read count, unmethylated read count and methylation ratio (beta value). These columns were extracted from the format to mimic the Bismark.cov format created via the processing of internal samples, creating two cohorts that can be processed further in unison. Going forward, internally sequenced samples are termed as ‘NMB’ and externally acquired data termed as ‘ICGC’

### 3.3.4 – Methylation data processing

#### 3.3.4.1 - Computing resources / software

Following raw sequencing data processing and data download, methylation data could be processed in unison. All subsequent analyses in section 3.3.7 were conducted with the R statistical computing environment version 4.0.3 – “Bunny-Wunnies Freak Out” within RStudio (R Core Team, 2021). All processing was conducted within the R/Bioconductor package RnBeads 2.0 (Müller *et al.*, 2019), a comprehensive package for the analysis of both methylation microarray and bisulfite sequencing data.

#### 3.3.4.2 – QC & filtering of DNA methylation microarray data

Bead Array IDAT files were obtained for all 35 matched samples were analysed using the Quality Control and pre-processing modules within RnBeads 2.0. Quality was assessed by checking the signal distribution of quality control probes across all samples. Probe detection p-values were checked in all samples using a filter of < 0.05. If 5% of probe detection p-values were >0.05, the sample would be removed. All samples passed quality control. Next, samples were filtered for known SNPs (mapping within 2 nucleotides) and normalised via the normal-exponential out-of-band (noob) method using

single sample normalisation (Fortin, Triche. Jr & Hansen, 2017). This is in keeping with the processing methods used for the ICGC array cohort. Following normalisation, probes located on the X and Y chromosomes were removed and beta value matrices generated for downstream analysis. Missing data was then imputed using the kNN algorithm (Altman, 1991).

### 3.3.4.3 – QC & filtering of WGBS methylation data

Methylation read counts for both NMB and ICGC cohorts were analysed using the quality control and pre-processing modules within RnBeads 2.0. CpGs overlapping with known SNPs (within 3 nucleotides), high coverage outlier sites and sex chromosome sites were removed. High coverage outlier CpGs were defined as any CpG where coverage exceeds 50 times the 0.95 quantile of coverage values in its sample. Sites with low coverage designated by a specified filter were removed. Anticipating high numbers of low coverage sites due to the nature of low-pass data, we tested multiple coverage filters for each dataset (Table 3.1). Following processing, beta value matrices were generated for downstream analysis.

Coverage	Coverage Threshold(s) Applied	
<b>30x</b>	5	
<b>10x</b>	5	3
<b>9x</b>	5	3
<b>8x</b>	5	3
<b>7x</b>	5	3
<b>6x</b>	5	3
<b>5x</b>	3	1
<b>4x</b>	3	1
<b>3x</b>	1	
<b>2x</b>	1	
<b>1x</b>	1	

**Table 3.1** – Coverage filters applied to each downsampled dataset. Lower filters were also used to attempt to mitigate the expectedly high degree of missing data.

### 3.3.5 – Data downsampling

NMB samples were downsampled in whole number increments from 10x to 1x using SeqKit v2.1.0 (Shen *et al.*, 2016). Rather than proportionally downsampling all levels of coverage from the original 10x data, we iteratively downsampled from each corresponding high coverage dataset to emulate the random missingness of reads expected at lower sequencing depths. Following downsampling, a total of ten datasets were generated at 10x-1x. All sample data was processed using the methodologies previously mentioned in 3.3.1.

Due to ICGC sample data being acquired in a pre-processed, methylation read count format, it was necessary to use an alternative means of downsampling to generate matched datasets. Methylated read counts matrices were extracted for each sample and downsampled using the `downsampleMatrix` function within the `DropletUtils` package (version 1.10.3) in RStudio (Griffiths *et al.*, 2018). For each sample, methylation read counts matrices were iteratively downsampled to depths corresponding to that of the NMB cohort (10x-1x).

### 3.3.6 – Imputation via BoostMe

To generate imputed data, processed methylated read counts for both NMB and ICGC WGBS cohorts (in `bismark.cov` format) were read and combined into a single RObject using `bsseq` v1.26.0 (Hansen *et al.*, 2014). `BSseq` class objects enable methylated count matrices to be manipulated in a computationally efficient manner by creating separate objects for genomic co-ordinates, CpG coverage and Methylation values (both matrices). To test the effect of sequencing coverage on imputation, downsampled objects were created using the downsampled datasets defined in section 3.3.8.

Following data preparation, each combined dataset was then imputed using the `BoostMe` package (Zou *et al.*, 2018). Only CpGs below a chosen depth are imputed – specified using the same coverage filters applied using processing (Table 3.1). `BoostMe` applies the XGBoost gradient boosting algorithm to impute low coverage CpGs by leveraging known methylation data in other samples within the cohort. Machine learning imputation is explained in detail in chapter 2; briefly for a missing CpG locus, `BoostMe` uses the average beta value at the same loci from neighbouring samples, the beta values of the nearest non-missing CpGs within the same sample and their respective distances as features for training, testing and validation. Imputation is then performed on a per sample basis. If there is not enough information to impute a specific CpG (e.g., if no CpG is above the specified coverage filter in at least two other samples), then imputation is skipped at that locus. For training, testing and validation of the model, the default option of 1 million randomly selected CpGs was used.

Model performance was assessed by the following measures:

- **Root mean square error (RMSE)** – a standard deviation of all prediction errors. RMSE is an effective measure for regression classification, allowing us to measure how far the predicted data point (in this case our methylation values) differ from the line of best fit determined by the trained model. In our case of imputation, a lower RMSE indicates better model performance, with 0 being the perfect score.
- **Accuracy** – A measure depicting the proportion of predicted values (rounded to 1, 0.5 and 0) compared to that of known values (rounded to 1, 0.5 and 0). 100% accuracy denotes a perfectly performing model, with 1 being a perfect score.
- **Area under the receiving operator characteristic curve (AUROC)** – A measure of the ability of the model to correctly predict fully methylated (beta = 1) and fully unmethylated (beta = 0) values. To determine this, a curve is plotted denoting the true positive rate (how many true 1s and 0s are predicted compared to the number of actual 1s and 0s in the training set, against the false positive rate (the proportion of values that are not correctly predicted by the model). Better performing models display AUC values closer to 1, with 1 being a perfect score.
- **Area under the precision-recall curve (AUPRC)** – Similar to AUROC, but a curve is drawn against the precision (a ratio of the number of accurately predicted values divided by the sum of all values) and recall (analogous to true positive rate) rather than true positive and false negative rates. Whilst similar, AUPRC provides a measure of model calibration as opposed to measuring its ability to discriminate, which AUROC achieves. A perfect model would yield AUPRC values of 1.

RMSE, accuracy, AUROC and AUPRC statistics were compiled for all imputed datasets. Statistical differences in imputation performance between NMB and ICGC cohorts were calculated via t-test, whilst potential differences in performance due to coverage and applied filter within cohorts were calculated via paired t-test. Significant differences were determined using a threshold of  $p < 0.05$ .

### 3.3.7 – Dataset summary

In summary, imputed and non-imputed WGBS beta value matrices (where rows = CpG coordinates, columns = samples) were generated for use in downstream analyses. A single imputed beta value matrix was also generated for the matched array cohort, allowing for downstream methylation analyses to be conducted in parallel for both platforms. Downsampled WGBS datasets for coverages 10-1x were also created to examine the effect of lowering coverage on imputation model performance and cross-platform correlation (Figure 3.5).

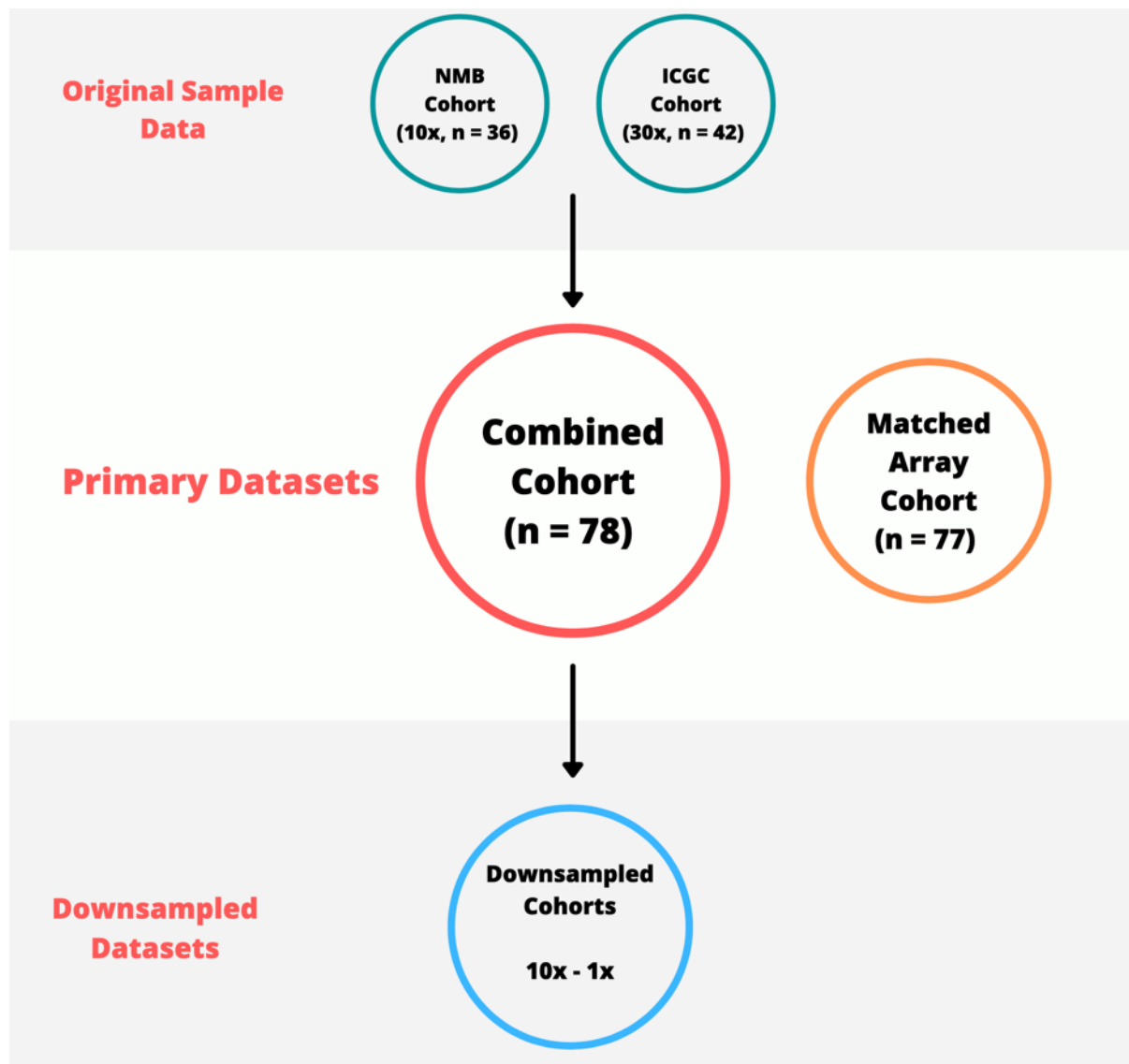


Figure 3.5 – Overview of generated datasets

### 3.3.8 – Liftover & interplatform correlation analysis

To enable a direct correlation to be drawn between WGBS and methylation microarray, a ‘Liftover’ cohort was generated for each WGBS dataset. This was done by sub-setting the data to only CpGs whose genomic coordinates mapped to that of the methylation array 450k probe manifest version 1.2 (Illumina). Any CpGs mapping to probes that were removed during pre-processing were removed. A global Pearson product moment correlation measure was applied to each sample for all corresponding CpGs/probes for each platform. This was also applied to all downsampled cohorts to investigate the effect of lowering coverage on methylation value correlation between the two platforms. Statistical differences in mean Pearson correlation values were calculated as specified in 3.3.6.

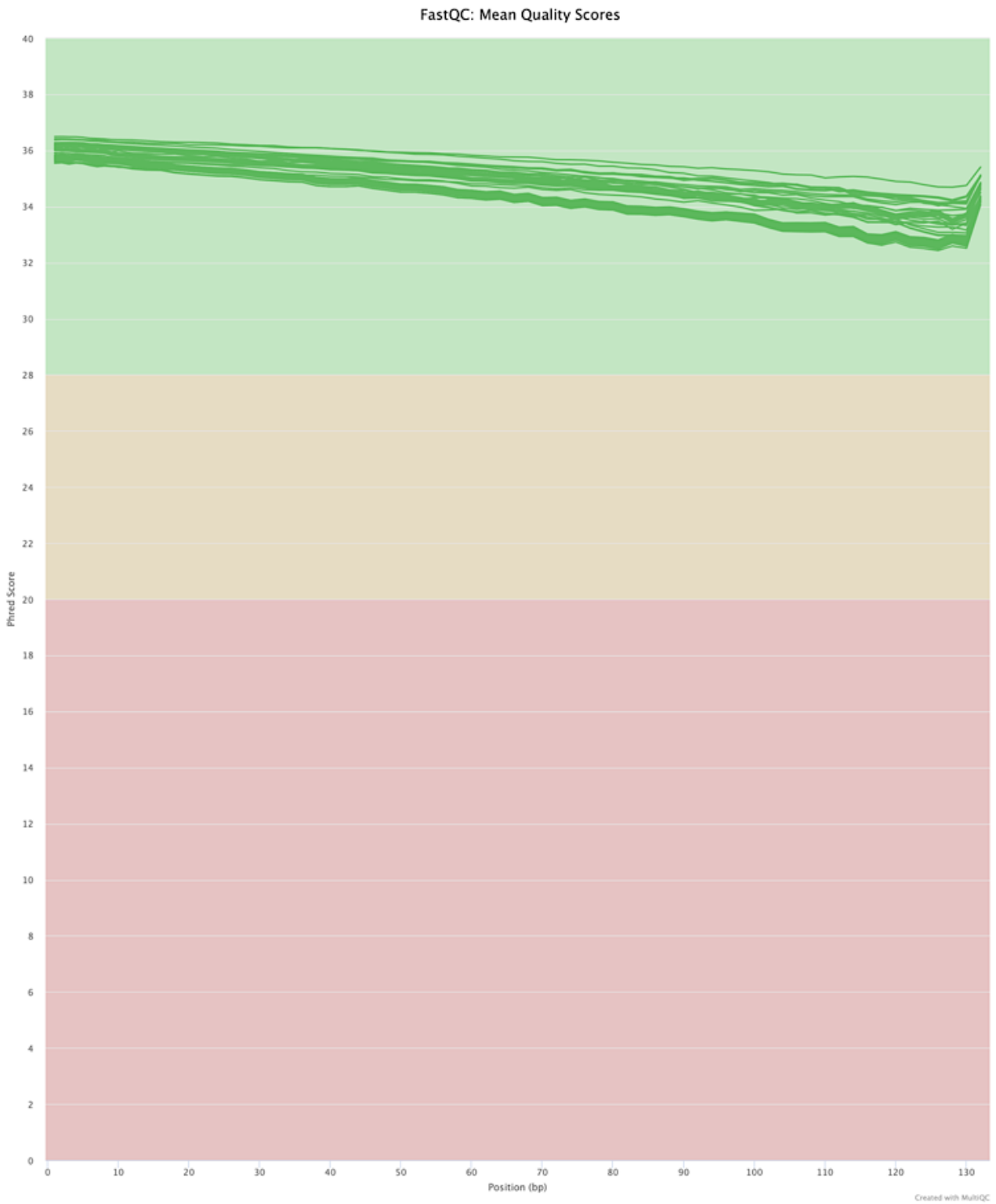
## 3.4 – Results & discussion

### 3.4.1 – Establishment of a robust processing pipeline for clinical low-depth WGBS samples

#### 3.4.1.1 – Quality control analysis of sequencing data

High quality data was obtained for all 36 samples selected for low-pass WGBS in this study. Whilst PBAT was the chosen library preparation method for all samples, four samples required a significantly larger amount of input DNA compared to the other 32 (Table 3.2). These four samples (NMB\_169, NMB\_181, NMB\_436, NMB\_787) were sequenced as part of a pilot study over 12 months earlier at which point the genome sequencing provider for this study required greater amounts of input DNA for sequencing. In addition to this, the remaining 32 samples initially failed library preparation with the sequencing provider, forcing the provider to attempt to use a low-input PBAT protocol that required 150ng DNA as a contingency measure. Ultimately, the quality of sequencing data outputted was judged to be satisfactory and appeared to be no different to that of the four samples prepared using far greater amounts of input DNA. This also had the unintended benefit of demonstrating that low-input WGBS could be applied effectively using clinical tissue samples, where DNA cannot often be supplied in such high amounts. An average paired-end sequencing yield of 217,290,129 reads were collected for all samples (Mean raw data size = 32.61GB). Sequencing statistics for the ICGC cohort are not comprehensive, but some are provided as part of the original study (Hovestadt *et al.*, 2014) (see appendix – 8.1.1). In this study, an average of 621,007,162 reads were generated across all 42 samples. The bisulfite conversion rates of all sequenced genomes in both studies were high, ranging between 99.32% to 99.62%.

Quality control of sequencing reads via FastQC indicates that all sample reads were of high quality. Mean raw per-base quality scores for the first 50 million reads in all samples was 34.955, far beyond the passing threshold of 28 (Figure 3.6). The single FFPE sample successfully sequenced was also high quality, with an average per-base quality score of 35.835. Following read trimming, a noticeable negative adenine bias is present in all reads at the 3' end. This is due to the stringent adapter trimming measures employed, removing any trailing adenine bases present at the end of reads. Whilst not an issue during primary quality control, adapter trimming also resulted in a slight positive cytosine bias in the final two bases of read 2. This may introduce a slight methylation bias during the methylation extraction step of processing and would subsequently be ignored to ensure stable methylation levels across all positions of all reads.



**Figure 3.6** – Mean per base quality score distribution for all NMB samples (n = 36).

The alignment rate for all remaining reads for each sample was high, averaging 78.75% across all samples. Similar alignment rates are also reported in other studies that utilise the same library preparation kits used by the genome sequencing company in this study (Raine *et al.*, 2017; Raine *et al.*, 2018). Of the reads that did not align uniquely, an average of 2,151,173 reads aligned ambiguously (aligned at more than one position in the genome), 20,831,656 did not align under any condition and 20 reads had no known genomic sequence to align to. A further 15,494,807 duplicate reads were removed on average from all samples prior to methylation extraction. An average CpG methylation rate of 73.8% (SD = 5.61) was observed in all samples, a similar rate to that of the external high-depth cohort used in this study (Hovestadt *et al.*, 2014). Average CHG and CHH methylation rates were 0.73% and 0.74% respectively. Two samples, NMB\_758 and NMB\_769 displayed CpG methylation rates greater than two standard deviations below the mean. Both samples were of LCA histology and belonged to the group 3 and subtype III methylation subgroup and subtype respectively. Both samples displayed CHG and CHH methylation rates far greater than the remaining samples, but no indicators of poor sample quality were identified. Average CpG coverage following methylation data extraction of processed reads was expectedly lower than the 10x depth originally sequenced, reaching an average depth of 7.18x across all samples



Sample ID	Library Preparation Method	Input DNA	No. of paired-end sequencing reads	No. of sequencing reads (single strand)	Raw Data Size (Gb)	BS-Conversion Rate (%)	Alignment %	Total No. of Cytosines (millions)	Average CpG Coverage	Average CpG Methylation Rate	Average CHG Methylation Rate	Average CHH Methylation Rate
<b>NMB_17</b>	Post-Bisulfite Adaptor Tagging	150ng	202,970,008	101,485,004	30.4	99.3	77.30%	3813	<b>6.6X</b>	69.1	1.8	1.8
<b>NMB_77</b>	Post-Bisulfite Adaptor Tagging	150ng	220,802,174	110,401,087	33.1	99.4	78.00%	3468.9	<b>7.2X</b>	79.2	0.7	0.7
<b>NMB_79</b>	Post-Bisulfite Adaptor Tagging	150ng	218,406,932	109,203,466	32.8	99.5	78.60%	3384.6	<b>7.2X</b>	72.3	0.5	0.5
<b>NMB_169</b>	Post-Bisulfite Adaptor Tagging	2.5ug	203,740,078	101,870,039	30.5	99.6	83.40%	3202.2	<b>7.0X</b>	74.3	0.4	0.4
<b>NMB_178</b>	Post-Bisulfite Adaptor Tagging	150ng	241,605,908	120,802,954	36.2	99.3	76.60%	3645.7	<b>7.8X</b>	77.2	0.7	0.7
<b>NMB_181</b>	Post-Bisulfite Adaptor Tagging	2.5ug	216,391,298	108,195,649	31.5	99.6	82.40%	3178.9	<b>7.2X</b>	71.3	0.4	0.4
<b>NMB_185</b>	Post-Bisulfite Adaptor Tagging	150ng	218,195,722	109,097,861	32.7	99.5	77.70%	3312.1	<b>7.2X</b>	75.4	0.8	0.9
<b>NMB_187</b>	Post-Bisulfite Adaptor Tagging	150ng	221,720,334	110,860,167	33.3	99.5	77.70%	3429.3	<b>7.3X</b>	78.7	0.7	0.7
<b>NMB_203</b>	Post-Bisulfite Adaptor Tagging	150ng	223,581,462	111,790,731	33.5	99.5	78.30%	4217.3	<b>7.3X</b>	73.6	0.6	0.6
<b>NMB_250</b>	Post-Bisulfite Adaptor Tagging	150ng	238,414,028	119,207,014	35.8	99.4	82.60%	4798.3	<b>8.3X</b>	77.4	0.6	0.6
<b>NMB_273</b>	Post-Bisulfite Adaptor Tagging	150ng	211,908,532	105,954,266	31.8	99.3	77.10%	3847.7	<b>6.9X</b>	77.2	0.7	0.7
<b>NMB_318</b>	Post-Bisulfite Adaptor Tagging	150ng	207,983,602	103,991,801	31.2	99.4	77.50%	3881.1	<b>6.8X</b>	73.6	2.3	2.4
<b>NMB_362</b>	Post-Bisulfite Adaptor Tagging	150ng	214,006,122	107,003,061	32.1	99.4	78.10%	4039.4	<b>7.0X</b>	79.6	0.7	0.7
<b>NMB_363</b>	Post-Bisulfite Adaptor Tagging	150ng	205,008,404	102,504,202	30.8	99.5	77.60%	3854.4	<b>6.7X</b>	75.4	0.7	0.7
<b>NMB_378</b>	Post-Bisulfite Adaptor Tagging	150ng	206,123,666	103,061,833	30.9	99.4	77.50%	3228.9	<b>6.8X</b>	69.8	0.6	0.6

<b>NMB_390_FF PE</b>	Post-Bisulfite Adaptor Tagging (FFPE)	150ng	202,021,308	101,010,654	31.5	99.535	82.90%	2715.1	<b>6.99X</b>	77	0.8	0.9
<b>NMB_390</b>	Post-Bisulfite Adaptor Tagging	150ng	225,371,846	112,685,923	33.8	99.337	77.50%	3428.2	<b>7.31X</b>	77.5	0.7	0.7
<b>NMB_416</b>	Post-Bisulfite Adaptor Tagging	150ng	209,393,592	104,696,796	31.4	99.433	81.60%	4072	<b>7.16X</b>	79.7	0.6	0.6
<b>NMB_433</b>	Post-Bisulfite Adaptor Tagging	150ng	199,674,444	99,837,222	30	99.435	77.40%	3064.8	<b>6.46X</b>	76.1	0.7	0.7
<b>NMB_436</b>	Post-Bisulfite Adaptor Tagging	2.5ug	246,922,252	123,461,126	37.0 4	99.61	81.50%	4360.9	<b>8.40X</b>	76.1	0.4	0.4
<b>NMB_529</b>	Post-Bisulfite Adaptor Tagging	150ng	259,848,460	129,924,230	39	99.43	81.70%	5171	<b>8.94X</b>	66.8	0.6	0.6
<b>NMB_549</b>	Post-Bisulfite Adaptor Tagging	150ng	229,704,050	114,852,025	34.5	99.4	78.40%	3544.7	<b>7.57X</b>	68.6	0.6	0.6
<b>NMB_610</b>	Post-Bisulfite Adaptor Tagging	150ng	205,686,756	102,843,378	30.9	99.51	77.00%	3827.7	<b>6.67X</b>	66.2	0.5	0.5
<b>NMB_725</b>	Post-Bisulfite Adaptor Tagging	150ng	200,189,558	100,094,779	30	99.415	76.80%	3101.9	<b>6.51X</b>	74.4	0.7	0.7
<b>NMB_729</b>	Post-Bisulfite Adaptor Tagging	150ng	243,202,318	121,601,159	36.5	99.575	78.10%	4587.1	<b>8.04X</b>	67.7	0.5	0.5
<b>NMB_733</b>	Post-Bisulfite Adaptor Tagging	150ng	213,862,174	106,931,087	32.1	99.465	81.00%	3277.7	<b>7.29X</b>	78.1	0.9	1
<b>NMB_758</b>	Post-Bisulfite Adaptor Tagging	150ng	207,052,574	103,526,287	31.1	99.543	77.20%	3840.6	<b>6.73X</b>	57.6	0.4	0.4
<b>NMB_769</b>	Post-Bisulfite Adaptor Tagging	150ng	201,004,734	100,502,367	30.2	99.57	77.20%	3768.2	<b>6.53X</b>	58	2	2
<b>NMB_772</b>	Post-Bisulfite Adaptor Tagging	150ng	213,820,964	106,910,482	32.1	99.41	77.00%	3998.3	<b>6.97X</b>	77.6	0.6	0.6
<b>NMB_774</b>	Post-Bisulfite Adaptor Tagging	150ng	207,929,412	103,964,706	31.2	99.487	76.30%	3846.1	<b>6.69X</b>	68.5	0.5	0.5
<b>NMB_787</b>	Post-Bisulfite Adaptor Tagging	2.5ug	209,844,192	104,922,096	31.4 8	99.6	82.30%	3189.5	<b>7.22X</b>	81.7	0.5	0.5
<b>NMB_803</b>	Post-Bisulfite Adaptor Tagging	150ng	221,292,150	110,646,075	33.2	99.505	78.00%	3409.5	<b>7.30X</b>	77.3	0.5	0.5
<b>NMB_858</b>	Post-Bisulfite Adaptor Tagging	150ng	253,421,272	126,710,636	38	99.385	77.30%	4707.7	<b>8.25X</b>	74.5	0.7	0.7

<b>NMB_867</b>	Post-Bisulfite Adaptor Tagging	150ng	210,136,200	105,068,100	31.5	99.463	81.50%	4153.6	<b>7.20X</b>	79.7	0.6	0.6
<b>NMB_870</b>	Post-Bisulfite Adaptor Tagging	150ng	208,355,470	104,177,735	31.3	99.38	77.40%	3942.1	<b>6.81X</b>	72.9	0.7	0.7
<b>NMB_890</b>	Post-Bisulfite Adaptor Tagging	150ng	202,852,640	101,426,320	30.4	99.443	76.60%	3748.9	<b>6.55X</b>	79.5	0.6	0.6

**Table 3.2** – Sequencing statistics of NMB cohort (10x).

### 3.4.1.2 – Quality control analysis of methylation data – WGBS

Following methylation extraction, methylation read counts for all captured CpGs were prepared for all NMB samples. ICGC samples were collected in the same format, enabling methylation data QC to be performed in unison (Table 3.3). A total 27,860,185 CpGs were read at least once across all samples in the NMB cohort out of a total ~28.3 million CpGs estimated to be in the human genome (Babenko, Chadaeva & Orlov, 2017). The total number of CpGs covered in the ICGC cohort was slightly lower, with 27,603,130 covered in total. The greater number of CpGs covered in the internal cohort is reflected in the greater number of SNP-enriched sites during filtering, with 358,319 and 362,531 sites removed on average per sample respectively. The same difference in the total number of CpGs located on the sex chromosomes was also observed, with 1,242,708 and 1,281,868 identified in both cohorts (Table 3.3). Greater numbers of high coverage outlier sites (defined as any CpG where coverage exceeds 50 times the 0.95 quantile of coverage values in its sample) were removed in the external cohort compared to internally sequenced samples, with 13,755 and 6,341 sites filtered respectively. This is due to the higher sequencing depth of samples in the external cohort, resulting in greater numbers of sites being sequenced at ultra-high depths. Average CpG coverage following CpG filtering for both cohorts was 29.48x and 6.80x respectively.

The effect of higher sequencing depth naturally persists when analysing the total number of CpGs with coverage above specific thresholds. For ICGC samples, 96.07% of CpGs were sequenced at a depth greater than or equal to 5x prior to filtering (Table 3.3). Following filtering, no sites were identified below 5x. For NMB samples, the number of sites identified as low coverage is magnitudes greater, with just 63% of reads being sequenced at greater than or equal to 5x. When applying the recommended 5x coverage filter, a total of 11,523,459 CpGs are removed on average per sample. It is at this stage of filtering where data missingness presents itself as a significant barrier to downstream analysis, as all low coverage CpGs are randomly distributed in different areas of the methylome in each sample. This scale of data loss is emphasised when subsetting the dataset for only common CpGs, identifying just 69,101 CpGs that are shared amongst all samples. This presents a significant analytical challenge when handling this dataset, as common downstream analyses used in for molecular subgrouping do not tolerate missing values.

In attempt to mitigate the loss of over one third of total CpGs from samples, a second 'low filter' NMB cohort was processed where a lower depth of 3x was specified for the masking of low-depth CpGs. As expected, this resulted in a significant reduction in the total number of CpGs removed, with 4,574,233 removed per sample on average, representing a recovery of 60.31%. This recovery is underlined when filtering for CpGs in common across all samples, where 1,967,445 CpGs are now shared. Regardless, this number remains low compared to the total number of CpGs present in the dataset and highlights the pressing need for an effective imputation strategy to be used when handling

low-coverage WGBS data. Applying a lower coverage filter clearly results in a substantial recovery of data, but it is recognised that including methylation data calculated using lower depth CpG counts may result in the inclusion of inaccurate data. However, such a decision may be necessary and studies utilising low coverage WGBS have applied filters below 5x for the differential methylation analysis of breast cancer samples previously (Hoppe *et al.*, 2021), providing support for its application when optimising a processing pipeline for this study. Regardless, it was deemed necessary that both datasets continue be analysed to comprehensively assess the effect that applying a lower coverage threshold would have on both imputation and cross-platform correlation.

	ICGC	NMB (High Filter)	NMB (Low Filter)
<b>No. of Sites Covered (Across all samples)</b>	27,603,130		27,860,185
<b>Mean Cytosine Coverage</b>	29.5X		6.8X
<b>No. CpGs with coverage <math>\geq 5</math> (Avg per sample)</b>	26,518,524		17,648,537
<b>No. CpGs with coverage <math>\geq 10</math> (Avg per sample)</b>	25,366,501		5,484,378
<b>No. CpGs with coverage <math>\geq 30</math> (Avg per sample)</b>	11,692,040		26,102
<b>No. CpGs with coverage <math>\geq 60</math> (Avg per sample)</b>	706,239		11,154
<b>SNP-enriched sites removed</b>	358,319		362,531
<b>High Coverage Outlier Sites Removed</b>	13,755		6,341
<b>Low Coverage Sites Masked (Avg per sample)</b>	0	<b>11,524,459</b>	<b>4,574,233</b>
<b>Sex Chromosome Sites Removed</b>	1,242,708		1,281,868
<b>Total Samples Retained</b>	42		36

**Table 3.3** – Methylation processing statistics for ICGC (filtered at 5x) and NMB cohort (High filter = 5x / Low filter = 3x).

### 3.4.1.3 – Differences between internal and external cohort downsampling approaches

Coverage analysis of both downsampled NMB and ICGC cohorts highlighted discrepancies between the intended and actual CpG coverage of each dataset due to the issues caused by working with raw and pre-processed sequencing data. Randomised iterative downsampling of raw sequencing reads (as applied to internally sequenced samples) performed as expected, with the average number of paired-end reads per sample falling proportionately with each drop in coverage. This removal of reads prior to processing in turn causes a fall in the total number of CpGs covered in the methylome (Table 3.4).

<b>Sequenced Coverage</b>	<b>No. of paired-end sequencing reads (mean)</b>	<b>Alignment % (mean)</b>	<b>Total No. of Cytosines (millions, mean)</b>	<b>Average CpG Coverage (Actual)</b>	<b>Average CpG Methylation Rate</b>	<b>Average CHG Methylation Rate</b>	<b>Average CHH Methylation Rate</b>
<b>10x</b>	217.3	78.75%	3751.6	<b>7.19X</b>	73.9	0.7	0.7
<b>9x</b>	195.6	78.75%	3060.5	<b>6.47X</b>	73.9	0.7	0.7
<b>8x</b>	173.9	78.75%	2776.4	<b>5.75X</b>	73.9	0.7	0.7
<b>7x</b>	152.2	78.75%	2476.9	<b>5.03X</b>	73.9	0.7	0.7
<b>6x</b>	130.4	78.75%	2165.2	<b>4.31X</b>	73.9	0.7	0.7
<b>5x</b>	108.7	78.75%	1850.7	<b>3.59X</b>	73.9	0.7	0.7
<b>4x</b>	86.9	78.75%	1502.6	<b>2.88X</b>	73.9	0.7	0.7
<b>3x</b>	65.2	78.75%	1150.3	<b>2.16X</b>	73.9	0.7	0.7
<b>2x</b>	43.5	78.75%	782.9	<b>1.44X</b>	73.9	0.7	0.7
<b>1x</b>	21.7	78.75%	399.9	<b>0.72X</b>	73.9	0.7	0.7

Table 3.4 – Downsampled NMB cohort sequencing statistics (10-1x)

As highlighted in 3.4.1.1, sequenced coverage does not translate into actual coverage after trimming, alignment and deduplication is applied to reads. This bears out in all downsampled datasets, where the observed 'actual' coverage is lower than the intended sequenced (Table 3.4). This drop in coverage is expected when working with sequencing platforms, where duplicate reads are a common symptom of WGBS library preparation methodologies (Zhou *et al.*, 2019). This disparity in coverage was not observed when working with ICGC sample data, which was only available in a pre-processed format. As the data was already pre-processed prior to acquisition, any downsampling methodology would not remove reads, resulting in no loss of CpG sites covered like that observed in the NMB cohort (Table 3.5). The lack of downsampling at the sequence read level results in highly accurate CpG coverage for each downsampled dataset compared to the disparity observed in internally sequenced samples.

Whilst pre-processed data was downsampled to the target level of coverage accurately, it does not truly reflect the CpG coverage that would be obtained during sequencing like that of the NMB cohort. It is therefore highly likely that these samples were in fact sequenced at a higher level of coverage than 30x to achieve a CpG coverage at that depth. Regardless, recapitulating this disparity in coverage manually in the ICGC cohort is not feasible and sacrificing known CpGs in a non-randomised manner would introduce bias into any downstream analysis conducted. To control for the stable number of CpGs covered in ICGC samples, cohorts were combined at sites only present in NMB sample data at each designated level of coverage, ensuring that CpG information lost due to downsampling would also be missing from ICGC samples. Despite efforts to mitigate these issues, this disparity in coverage is a limitation of this study, caused by having to handle different datasets acquired at different stages of processing.

	<b>ICGC</b>	<b>NMB</b>	
<b>No. of Sites Covered (Across all samples)</b>	27,603,130	27,860,185	
<b>Mean Cytosine Coverage</b>	29.5x	6.8x	
<b>No. CpGs with coverage <math>\geq</math> 5 (Avg per sample)</b>	26,518,524	17,648,537	
<b>No. CpGs with coverage <math>\geq</math> 10 (Avg per sample)</b>	25,366,501	5,484,378	
<b>No. CpGs with coverage <math>\geq</math> 30 (Avg per sample)</b>	11,692,040	26,102	
<b>No. CpGs with coverage <math>\geq</math> 60 (Avg per sample)</b>	706,239	11,154	
<b>SNP-enriched sites removed</b>	358,319	362,531	
<b>High Coverage Outlier Sites Removed</b>	13,755	6,341	
<b>Coverage Filter Applied</b>	5	5	3
<b>Low Coverage Sites Masked (Avg per sample)</b>	<b>0</b>	<b>11,524,459</b>	<b>4,574,233</b>
<b>Sex Chromosome Sites Removed</b>	1,242,708	1,281,868	
<b>Total Samples Retained</b>	42	36	

Table 3.5 – Methylation processing statistics for both ICGC and NMB cohorts.



### 3.4.1.4 – Quality control analysis of methylation data - Array

Processing of matched methylation microarray samples yielded high quality data. To enable direct comparison, the pre-processing methodologies applied to NMB array samples was kept concordant with the methodologies used for ICGC sample data, which could only be acquired as a pre-processed beta value matrix (Hovestadt *et al.*, 2014). Control probe quality control highlighted no issues with any samples in the dataset. A total of 63,550 probes were removed, including 29,669 cross-reactive, 10,329 unreliable (detection p-value > 0.05) and 12,430 SNP-enriched probe sites. To ensure only common probe sites were analysed in both cohorts, ICGC samples were merged only at common probe sites, with any remaining probes removed. No additional probes were removed after combining cohorts, with a total of 422,027 probes taken forward for analysis (Table 3.6).

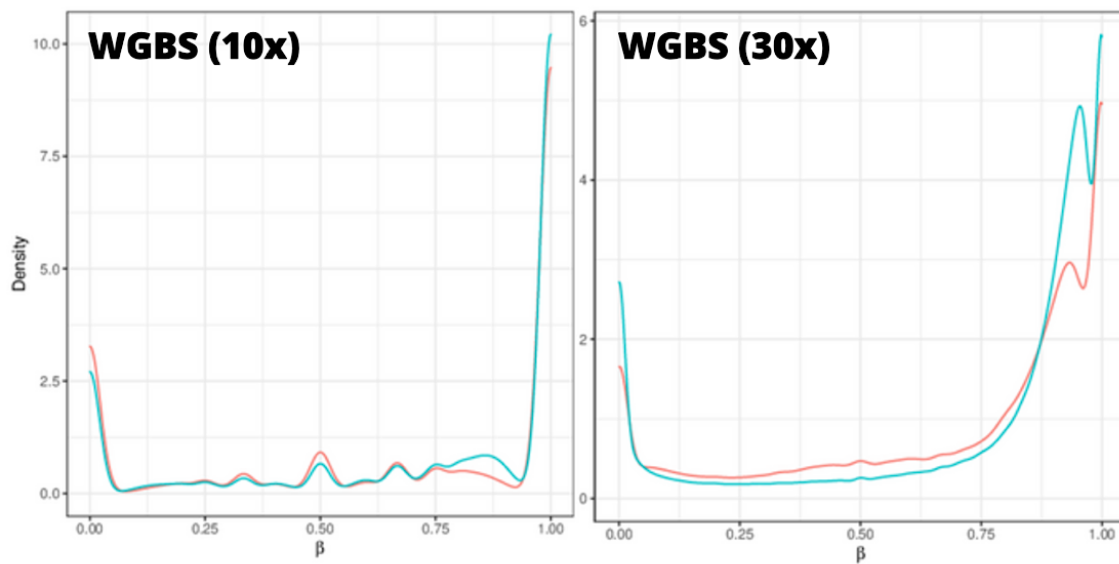
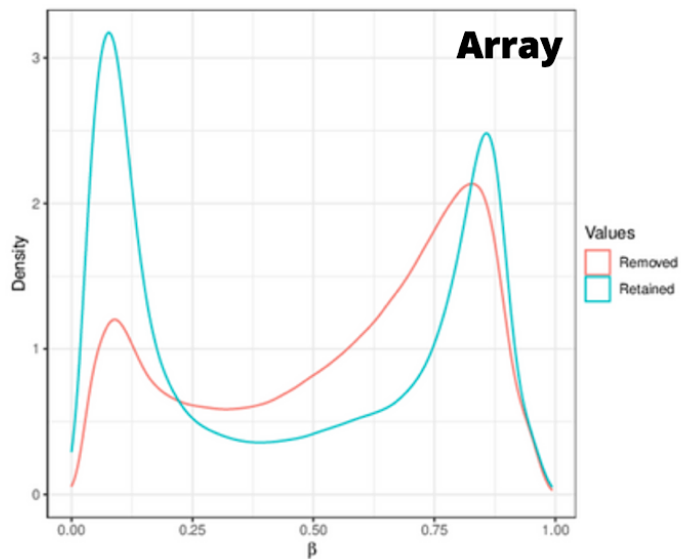
	<b>NMB</b>
<b>No. of Sites Covered (Across all samples)</b>	485,577
<b>SNP-enriched sites removed</b>	12,430
<b>Cross-reactive Probes removed</b>	29,669
<b>Greedy cut probes removed</b>	10,329
<b>Sex Chromosome Sites Removed</b>	9,912
<b>Context-specific Probes Removed</b>	1,211 (CAG: 995, CAH: 145, CTG: 7, Other: 64)
<b>Total Probes Removed</b>	63,550
<b>Total Samples Removed</b>	0
<b>Total Probes Retained</b>	<b>422,027</b>
<b>Total Samples Retained</b>	36

**Table 3.6** – Methylation processing statistics for matched NMB methylation microarray cohort.

### 3.4.1.5 – Methylation values calculated from WGBS sample data display increased bimodal distribution

Observing the beta value distributions of samples analysed from WGBS and methylation microarray revealed differences in methylation values characteristic to each platform. Greater numbers of CpGs are methylated for WGBS samples than that of the array. This is because all CpGs are interrogated, of which 70-80% are methylated (Jang *et al.*, 2017). Conversely, most probe sites interrogated by the methylation microarray are unmethylated, which is explained by deliberate targeting of gene promoters and CpG islands (Bibkova *et al.*, 2011). Array probe beta values are bi-modally distributed, with the greatest number of probe beta values peaking at between 0.7-0.9 and 0.05-0.15 respectively (Figure 3.7). WGBS derived beta values display increased polarisation, with the highest peaks being at both 0

and 1 respectively. This difference is due to the nature of how methylation values are calculated in each platform, where array values are calculated via normalised probe intensities from methylated and unmethylated readouts, and WGBS determined using a read count-based calculation. Differences between low and high coverage WGBS data are also prevalent due to the way they are calculated. Reads that have failed bisulfite conversion, and are incorrectly identified as unmethylated, result in a second peak of methylation values in the 0.9-0.99 range in high-depth WGBS data. Additional peaks are observed at other decimal intervals in low depth data, where an incorrect CpG has a larger effect on the calculated beta value (Figure 3.7). It is currently unknown what effect that the greater polarisation of methylation values may have on downstream analysis, if any, but it is likely that any CpG variance (or lack thereof) will be more pronounced in WGBS data, represented by a greater/lower standard deviation value calculated for any given row of CpG methylation values than would be for any given probe site analysed using the array. Given that CpG variance filtering is widely applied as a method of feature selection and dimensionality reduction in molecular subgrouping studies (Northcott *et al.*, 2017; Capper *et al.*, 2018; Sharma *et al.*, 2019), an investigation into CpG variability between both platforms for matched cohorts would be beneficial to assess any differences that may be observed in analyses that leverage selected CpG-sets such as visualisation, classification, and clustering.



**Figure 3.7** – Beta value density distribution plot for WGBS (ICGC = 30x / NMB = 10x) and Array methylation data. For each cohort – Density lines are drawn for both removed and retained beta values.

## **3.4.2 – Machine learning imputation predicts robust, accurate methylation values in low-pass WGBS samples**

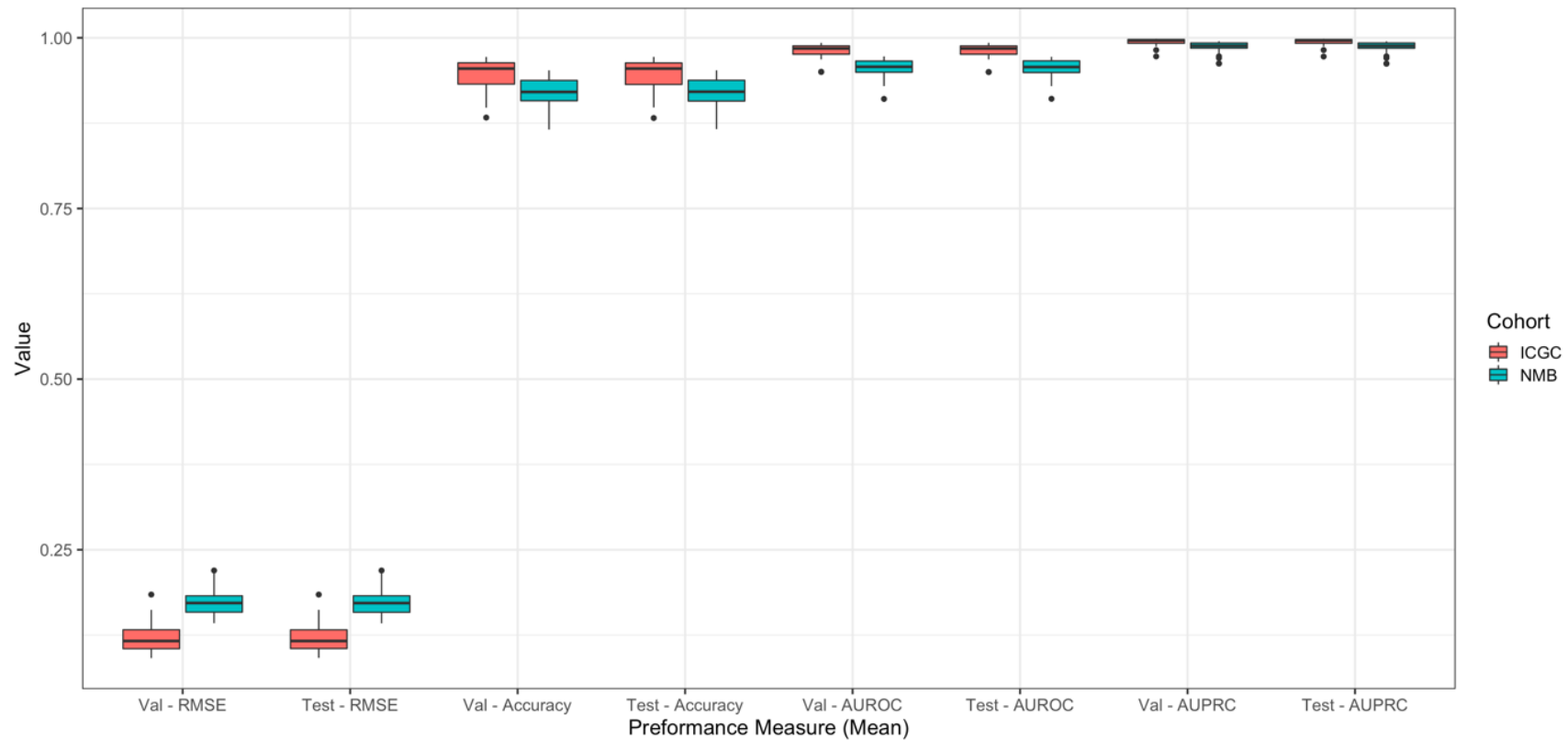
### **3.4.2.1 – Model performance**

The increased coverage offered by WGBS gives the platform a definitive advantage over methylation microarray platforms, which cover just 3% of CpGs in the methylome. Although low-pass WGBS is becoming a feasible proposition for researchers, high-depth WGBS remains costly. Data missingness presents an analytical challenge when working with low-pass WGBS data, where many common subgrouping analyses such as visualisation, clustering and classification do not tolerate missing values. When applying the recommended 5x filter to the low-pass samples in this study, just 69,101 CpGs in common across all assessed samples are identified out of a possible 27,860,185, with an average 11,523,459 CpGs missing per sample. Over 60% of missing CpGs were recovered when applying a lower filter of 3x, but the risk of including methylation values at lower read depths has not yet comprehensively been assessed. Missing data handling for methylation microarray platforms form a straightforward component of most processing pipelines, with a wide array of optimised algorithms available, including k-nearest neighbours, linear regression, and random forest. However, most of these approaches become computationally unfeasible when handling WGBS data where the beta value matrices involved exhibit significantly increased dimensionality (over millions of CpG rows per sample column) and sparsity. Current studies testing the capability of methylation sequencing for brain tumour classification have avoided imputation entirely, drastically subsetting the total number of CpGs for downstream analysis (Kuschel *et al.*, 2021; van Paemel *et al.*, 2020). Whilst they have demonstrated success to a certain degree, we deemed it necessary to impute missing data to provide a comprehensive analysis of the tumour and to test the fidelity of low-pass imputation.

Imputation algorithms for WGBS are available, with linear regression-based machine/deep learning models showing the most promise in providing robust methylation value prediction for sparse datasets (Stekhoven *et al.*, 2012; Zou *et al.*, 2018). Regression based methods have been shown to be highly effective for the imputation of methylation beta values, where the long-and short-range correlations are easily recapitulated by simple linear regression (Lena *et al.*, 2019). Machine learning algorithms such as Random Forest and XGBoost can increase the accuracy of such predictions via training on neighbouring features (CpGs) within the same sample and common features in other samples. One imputation software based on the XGBoost algorithm, BoostMe, has been shown to outperform other WGBS imputation algorithms and exhibited the highest computational efficiency (Zou *et al.*, 2018). This was confirmed after a pilot study tested our NMB cohort (n=36) within the HPC infrastructure available, which showed that imputation could be performed with less than 1TB RAM. Whilst this level of memory is high, it remains below that which has been available to that of previously mentioned

projects that have not implemented imputation and is well within the computational capabilities of most clinical HPC architectures (Kuschel *et al.*, 2021).

Application of the BoostMe algorithm to the combined WGBS cohort used in this study predicted methylation data with excellent performance (Figure 3.8). For test CpG sets ( $n = 1,000,000$ ) for all samples, the imputation model achieved an average root-mean-squared error (RMSE) of 0.15 (SD = 0.03), accuracy of 0.93 (SD = 0.03), area under the receiver operating characteristic curve (AUROC) of 0.97 (SD = 0.02) and area under the precision recall curve (AUPRC) of 0.99 (SD = 0.01). Both NMB and ICGC cohorts were combined prior to model implementation to leverage the increased CpG coverage present in the high-depth cohort. When focusing on each cohort individually RMSE performance for NMB samples was 0.18 (SD = 0.02, significantly higher than that of imputed ICGC sample data with an RMSE of 0.12 (SD = 0.02, t-test,  $p = < 0.001$ ). An observed fall in performance for NMB samples was not unexpected, given that far greater numbers of missing values are present than that of high coverage data. Additionally, imputation performance for the chosen algorithm in this study exceeded that achieved by other algorithms dealing with lower amounts of missing data in other studies, providing evidence that BoostMe is one of the top performing algorithms for WGBS data, even at low pass (Zou *et al.*, 2018).

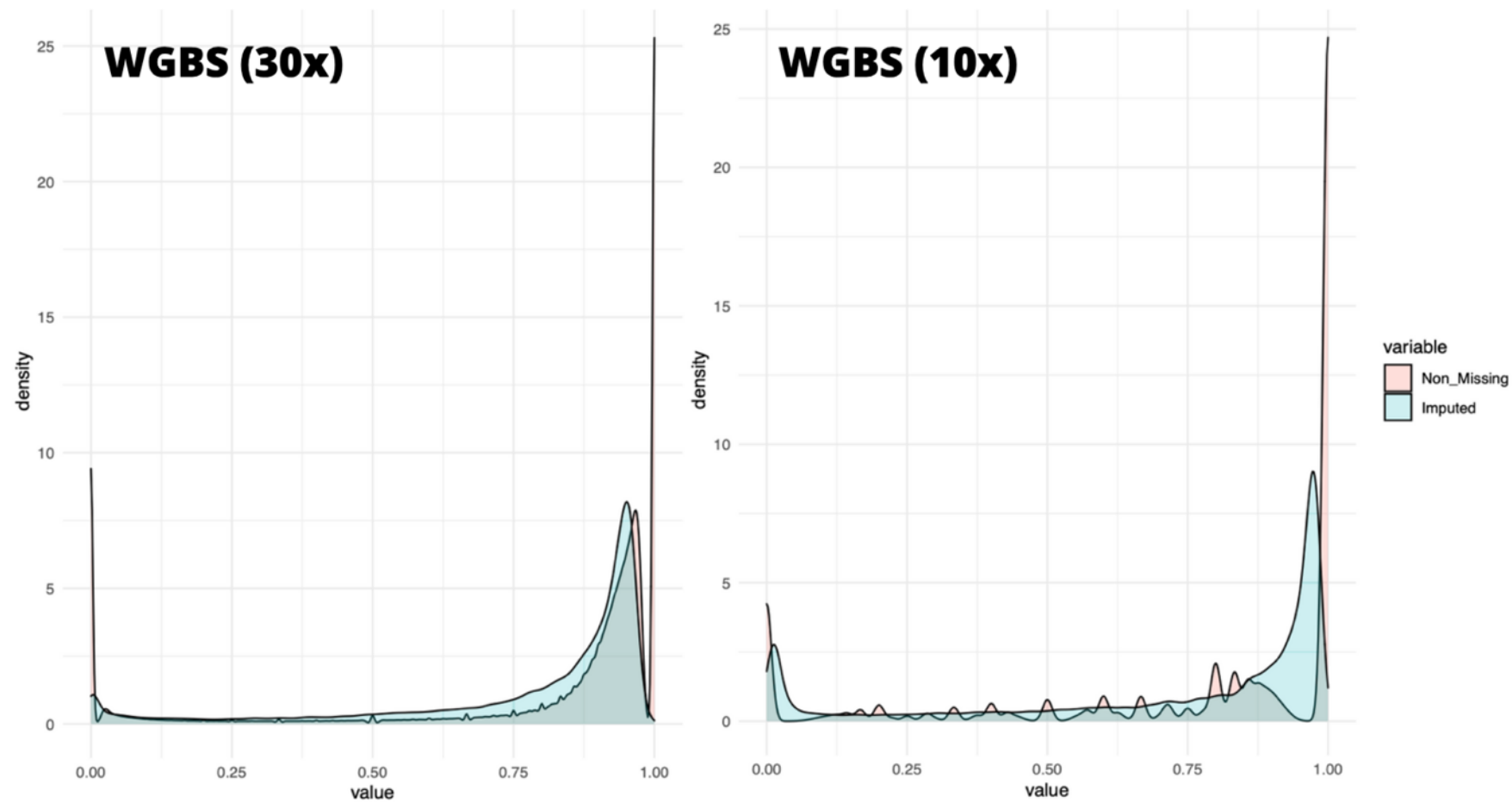


**Figure 3.8** - Validation and test CpG set imputation model performance for all samples both high and low-pass WGBS data (split into NMB and ICGC cohorts).

**Key:** The black line denotes the median value across all validation/test sets. Either side of the box denote the 25th and 75th percentiles, and the whiskers correspond to the inter quartile range. Dots outside of this range are considered outliers.

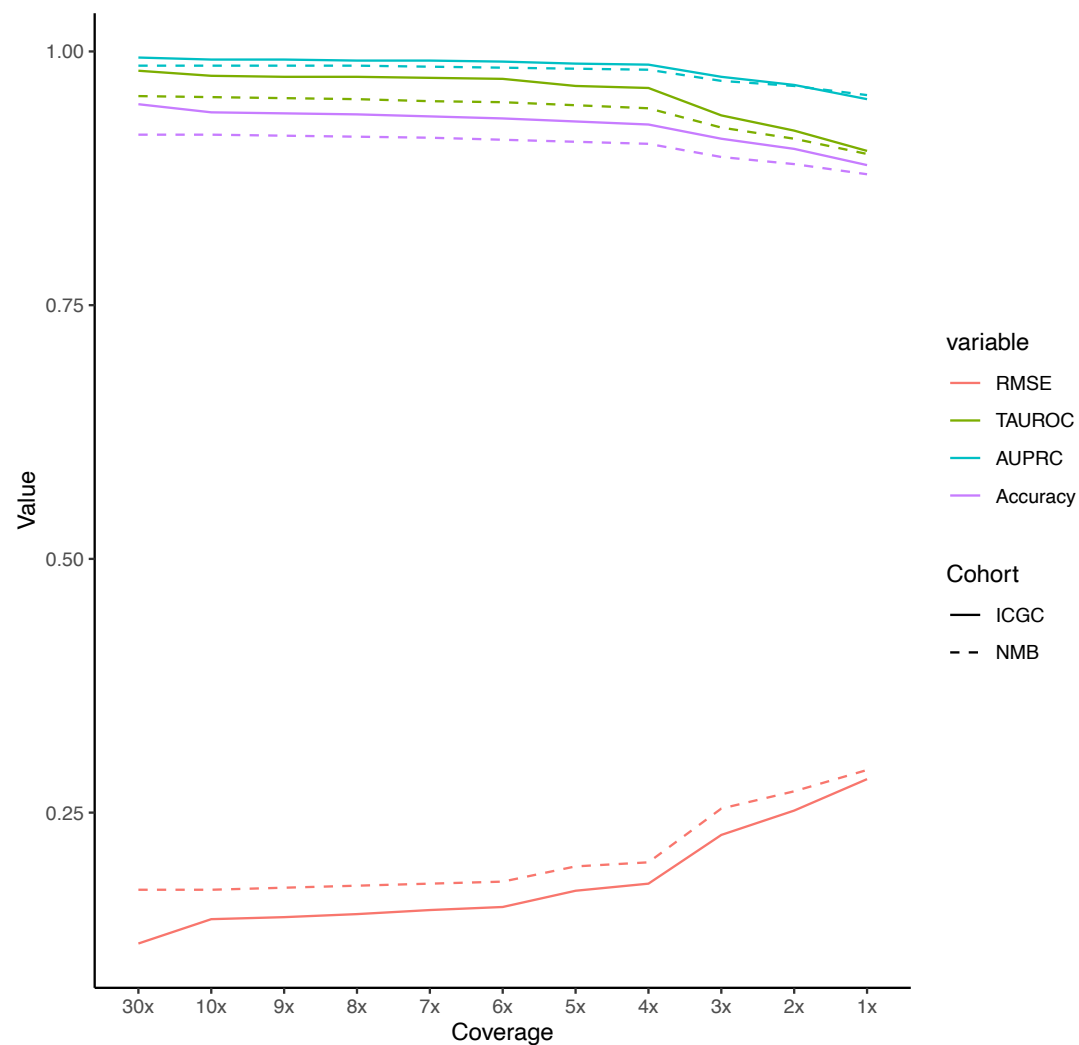
The excellent imputation performance is further demonstrated by the distribution of imputed values compared to known values. Imputed values mimic the bimodal distribution of known methylation values almost perfectly, where a continuous imputed value distribution was observed with peaks close to 0 and 1 (Figure 3.9).

The coverage filter applied to the combined WGBS cohort had a small, yet significant effect on model performance, where a higher coverage filter performed slightly better (Average RMSE = 0.145, SD = 0.03) than a lower coverage filter (Average RMSE = 0.148, SD = 0.03,  $p < 0.001$ ). Whilst significant, the difference is small, and it remains difficult to comment on whether the total number of CpGs that are recovered when using a lower coverage threshold offer additional benefits that offset the poorer imputation model performance identified at this stage of the analysis. Model performance significantly falls with each drop in coverage ( $p < 0.001$ ) with a more pronounced drop in coverage from 4x onwards (Figure 3.10). Imputation of ICGC sample data performs significantly better in all tested metrics than NMB sample data at each downsampled level of coverage ( $p < 0.001$ ). Whilst similar, the difference between the two serves to highlight the differences taken in each approach taken to downsampling, with greater levels of missingness present in internally sequenced data than that of externally sequenced samples.



**Figure 3.9** - Beta value distributions of non-missing (red) and imputed (blue) CpG data for high and low-pass WGBS sample data.





**Figure 3.10** – Mean test CpG set performance for downsampled high and low-pass WGBS data. Imputation model performance was measured using root mean squared error (RMSE), area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC) and accuracy.

### 3.4.2.2 – Potential for future optimisation

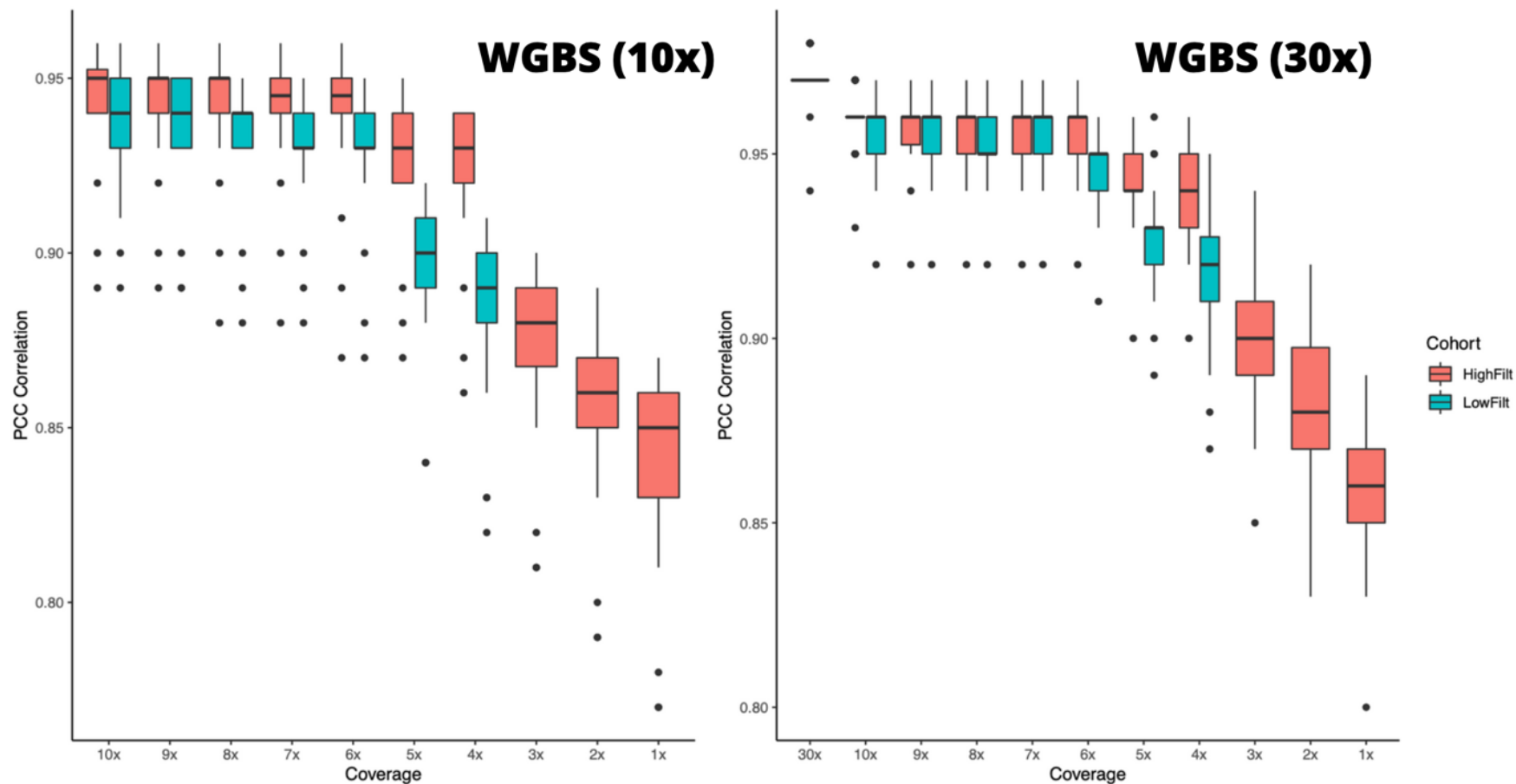
Imputation enables nearly all processed CpGs of the methylome to be brought forward for downstream analyses, alleviating many of the downstream analytical hurdles that are presented when working with sparse methylation datasets. Despite this, memory consumption for imputation, even for the relatively modest sized cohort in this study is prohibitively high for research groups that do not use HPC infrastructures. The development of low-memory imputation methodologies is therefore crucial if low-pass WGBS studies involving much larger cohorts are to be conducted. One such solution may be the introduction of a sliding window approach, where only a small subset of a matrix is loaded in memory at one time, drastically reducing the amount of memory required. Such an approach applying kNN has recently been offered as part of a WGBS processing pipeline named BSBolt, which may offer a solution to this computational obstacle (Farrell *et al.*, 2021). Regardless, the implementation of imputation via the BoostMe algorithms is a robust and effective method of handling missing data and should be included as part of processing pipelines for low-pass WGBS where possible.

### 3.4.3 – WGBS platform produces highly correlated methylation data to corresponding CpGs located on the array platform at low-sequencing depths

#### 3.4.3.1 – Cross-platform correlation analysis

Following the generation of high-throughput, non-missing methylation data for both WGBS and Array platforms, the final analytical step was to perform a cross-platform correlation analysis for all matched samples. By subsetting WGBS CpG sites to the genomic co-ordinates of all probe sites located on the Illumina Methylation450k manifest, a Pearson correlation coefficient value was derived for matched samples to characterise the concordance of WGBS and Microarray beta value estimates. The subsetting method used was adapted from *methyLiftover*, a prevalent function for creating WGBS-derived beta matrices that map to array probe manifests (Titus *et al.*, 2016). Liftover of all WGBS samples in this study matched non-missing CpGs to 469,341 probe sites out of a total 485,577. When disregarding for 3,091 non-CpG probes and 65 random SNP probes included as part of the 450k manifest (Bibkova *et al.*, 2011), over 96% of probe sites are covered by all samples in this study, representing a suitably high coverage suitable for correlation. When accounting for probes filtered during array cohort processing, the proportion of covered probes increases, with a total 420,809 CpGs present out of a total 422,027 (99.7%) brought forward for analysis.

Overall inter-platform correlation between WGBS and Methylation microarray for non-imputed sample data used in this study was very high. Average Pearson product-moment correlation coefficient (PCC) for ICGC sample data was 0.97 (SD = 0.006), replicating the correlation observed by the original study (Hovestadt *et al.*, 2014). PCC for NMB sample data was significantly lower owing to lower sequencing depth (PCC = 0.95, SD = 0.014, t-test -  $p < 0.001$ ). When including WGBS CpGs sequenced at 3x or above, the inter-platform PCC is significantly lower (PCC = 0.94, SD = 0.014,  $p = 0.02$ ). Average PCC values for downsampled data remains high for both sample cohorts (Figure 3.11). PCC for NMB samples do not exhibit a significant fall until reaching 5x coverage (PCC = 0.93, SD = 0.016,  $p < 0.001$ ), at which point PCC falls with each sequential drop in depth, reaching 0.84 at 1x (SD = 0.026,  $p = 0.004$ ). PCC for ICGC samples significantly falls when downsampling from 30x to 10x, falling from 0.97 (SD = 0.006) to 0.96 (SD = 0.006) respectively ( $p < 0.001$ ). PCC for each subsequent drop in coverage remains the same until reaching 6x, similar to the NMB cohort ( $p < 0.001$ ).



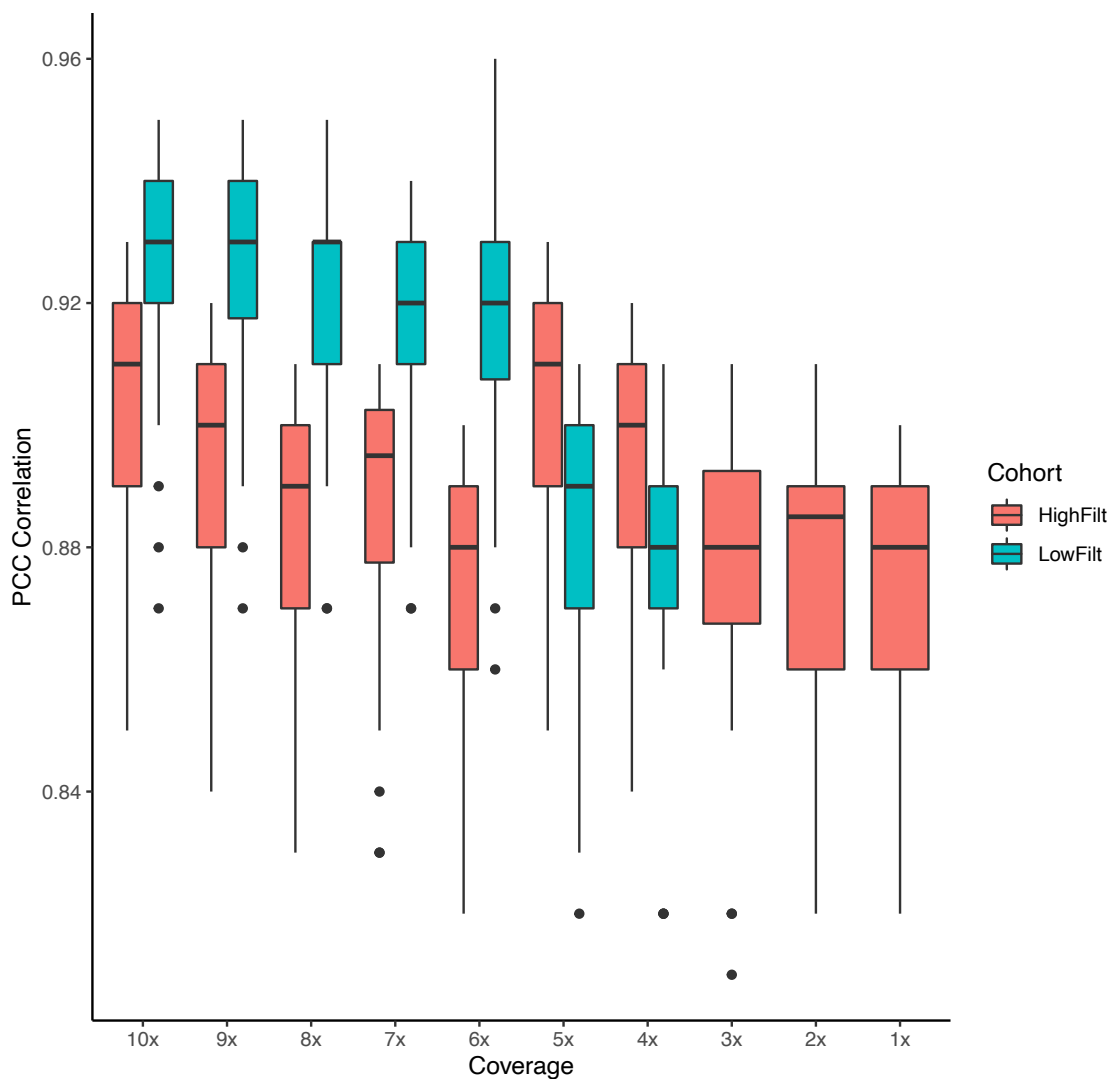
**Figure 3.11** - Cross-platform Pearson correlation values for downsampled high and low-pass WGBS cohorts.

**Key:** The black line denotes the mean value across all samples. Either side of the box denote the 25<sup>th</sup> and 75<sup>th</sup> percentiles, and the whiskers correspond to the inter quartile range. Dots outside of this range are considered outliers.

The high correlations observed between beta values calculated using Array and low-pass WGBS provides support for the use of the platform for subgrouping analyses. The concordant beta values that can be derived from low pass raise the possibility that the molecular subgroups of medulloblastoma can be visually identified and correctly classified by existing classification algorithms trained using microarray feature sets. The high correlation identified in this study at low pass is consistent with other studies, where correlation values  $>0.9$  are obtained for WGBS samples sequenced at 10x and above (Titus *et al.*, 2016). Achieving a cross-platform correlation that is any higher would be very difficult, particularly for low-pass data (i.e., 10x or below). For example, a CpG locus with a true methylation value of 65% would be difficult to recapitulate using WGBS; a methylation value of only 60 or 70%, at best, could be achieved using 10x data, whereas any beta value estimate using array may be more accurate given that values are calculated via fluorescence intensity-based scale (Zou *et al.*, 2018). Such a scenario highlights the drawbacks of using low-pass data, but the high PCC obtained at coverages as low as 5x in this study shows that may be less problematic for downstream analysis that often involves the analysis of multiple CpGs and is not dependent on the investigation of any single CpG.

### 3.4.3.2 – Filter recommendations/summary

Inter-platform correlation for imputed data is significantly lower in both cohorts. ICGC data (30x) exhibits an average PCC fall of 0.01 for all samples, falling from 0.97 to 0.96 respectively (paired t-test,  $p = 0.02$ ). NMB samples (10x) display greater falls, falling to 0.90 (SD = 0.02,  $p = < 0.001$ ) when applying a 5x threshold and 0.93 (SD = 0.02,  $p = < 0.001$ ) when using a 3x threshold respectively. This disparity emphasises the far greater numbers of missing values present when applying a higher coverage filter for low-pass data (5x = 11,730,289 / 3x = 5,166,995) and highlights that whilst imputation is useful, it is inherently not as accurate as a true, sequenced methylation value (Figure 3.12). Although imputation models perform significantly better and non-missing correlation values are significantly higher when using a higher coverage threshold, the improved correlation offered when using a lower threshold offers support for applying such a filter when dealing with low pass data, as it indicates that the 60% of missing CpGs recovered are better correlated to array based estimates than imputed values. Significant PCC improvements are observed when using a lower coverage threshold down to 6x ( $p = < 0.001$ ), further demonstrating the benefits of retaining more originally sequenced CpGs (Figure 3.12).



**Figure 3.12** – Cross-platform correlation analysis for downsampled low-pass WGBS data.

**Key:** The black line denotes the mean value across all samples. Either side of the box denote the 25th and 75th percentiles, and the whiskers correspond to the inter quartile range. Dots outside of this range are considered outliers.

In summary, it was determined that applying a low coverage filter of 3x would be the recommended approach taken during pipeline optimisation for low-pass WGBS data during this study. Although imputation models perform better and non-missing values correlate, the vast amount of missing data lost when applying a higher filter causes these gains to be subsequently lost after imputation. Ultimately, it was judged that the issue of data missingness is the most significant factor when dealing with low-pass WGBS data and the significant number of CpGs recovered (over 60%) when using a low-pass filter places less of a burden on the chosen imputation model (larger proportion of missing data to predict), resulting in significantly improved concordance between the WGBS and array platforms. Increased concordance between the two platforms increases the likelihood that robust molecular subgrouping calls can be obtained using low-pass WGBS sample data.

---

### 3.4.4 – Conclusions/chapter summary

The optimised pre-processing pipeline presented in this chapter successfully demonstrates that high quality, non-missing methylation data can be obtained from samples sequenced using low-pass WGBS. Sequenced data was high in quality, providing sample reads capable of being used for various NGS-based analyses, such as variant calling and high-resolution aneuploidy detection. A side-effect of filtering reads is that a proportion of read depth is subsequently sacrificed, resulting in the final coverage of reads being lower than that of the intended sequencing depth. Externally acquired data did not suffer from this issue, as data was already pre-processed. Proportional downsampling of both cohorts was performed using different methods due to the different format that each dataset was acquired in. This led to differences in downsampled data, with decreased missingness and increased coverage observed in pre-processed sample data that was now downsampled at the read level. Methylation data processing filtered unreliable CpGs effectively, providing over 25 million CpGs for downstream analyses. The issue of CpG missingness was highlighted at this stage for low-pass samples, where over 11 million CpGs were missing on average per sample and just 60,000 common CpGs were shared between 36 samples at 10x, a number that would only decrease as cohort size increases. Over 60% of missing CpGs could be recovered if using a lower coverage filter but would present the risk of including potentially unreliable methylation data. However, cross-platform correlation between WGBS and matched array data was very high, with a significant drop in correlation observed in lower depth samples and again when using a lower coverage threshold during processing.

Without imputation, any analytical approach using such a small set of common CpGs would be subject to severe bias, with just a random fraction of the methylome available for investigation. Imputation was therefore imperative, and it was demonstrated that performing XGBoost algorithm-based imputation via BoostMe that accurate, bimodally distributed methylation data could be obtained, with excellent model performance from low-pass WGBS clinical sample data. Performance significantly fell with decreasing coverage, but nevertheless remained strong for sample data at depths as low as 4x, rendering the use of imputation a necessity if one wishes to perform any analysis using this platform at low pass. Unsurprisingly, imputed data correlates significantly less than true values, at which stage the utilisation of a lower coverage filter during processing presents significant benefits. The increased number of true CpGs recovered by using a lower filter result in significantly greater cross-platform correlations for low-pass data. Ultimately, although trained imputation models perform significantly better using higher coverage features, the vastly increased amount of missing data results in poorer concordance between WGBS and array calculated beta values, offsetting differences in imputation model performance. Such lower coverage filters (i.e., 3x) have seldom been used for WGBS (Laufer *et al.*, 2021), and consequently both standard and lower coverage filters were here considered for application to low-pass data. Going forward, the inclusion of an effective imputation strategy should form an essential component of WGBS processing pipelines, especially for low-pass data. Its



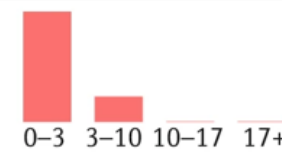
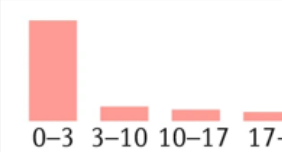
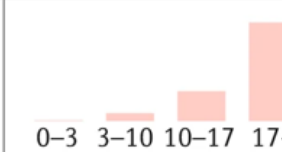
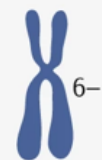
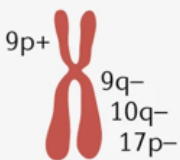

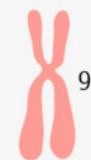
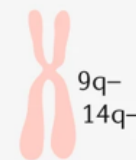
successful implementation allows the full potential of low-pass WGBS to be realised, with tens of millions of CpGs available for downstream analyses such as molecular subgrouping.



## **Chapter 4 – Molecular subgrouping of medulloblastoma via low-depth whole genome bisulfite sequencing**

## 4.1 – Introduction

Expression and methylation profiling studies of MB led to the identification of discrete molecular subgroups present within the disease (Northcott *et al.*, 2011; Taylor *et al.*, 2012; Schwalbe *et al.*, 2013). These subgroups: WNT, SHH, Group 3, and Group 4, each display distinct molecular and clinical phenotypes including age at onset and importantly, patient prognosis (Figure 4.1a-b). Patients with WNT tumours are associated with favourable survival, typically over 90% for childhood cases (Clifford *et al.*, 2006; Ellison *et al.*, 2005). SHH tumours are heterogenous in outcome, which is dependent on age, tumour histology and *TP53* mutation status (Garcia-Lopez *et al.*, 2021). Group 3 is a subgroup characterised by its higher rates of metastases and poor prognosis (Section 1.7.4), with *MYC* amplified patients in this subgroup performing badly (Sharma *et al.*, 2019). The picture is less clear for Group 4 patients (accounting for ~35% of all patients) which is the least biologically understood. This association between clinical outcome and molecular subgroup culminated in their incorporation of molecular subgroup into the WHO ‘Classification of Tumours of the Central Nervous System’ (Louis *et al.*, 2016). Their definition has transformed how the disease is studied and, importantly, how it is managed in the clinic, with the molecular subgroup that a patient belongs to a key consideration for clinicians when deciding on appropriate treatment strategies. For example, the increased survival exhibited by WNT patients has led to clinical trials being implemented that test the effect of decreased treatment intensity in this group, with the aim of alleviating many of the neuro-cognitive deficits that are associated with conventional MB treatment (ClinicalTrials.gov Identifier: NCT01878617). SHH inhibitors are in turn being investigated for use with SHH subgroup patients (ClinicalTrials.gov Identifier: NCT01878617). Since 2016, more recent studies have further identified additional subtypes present within SHH, Group 3, and Group 4, signalling a move to more precise classification and clinical management of MB in the future (Louis *et al.*, 2021).

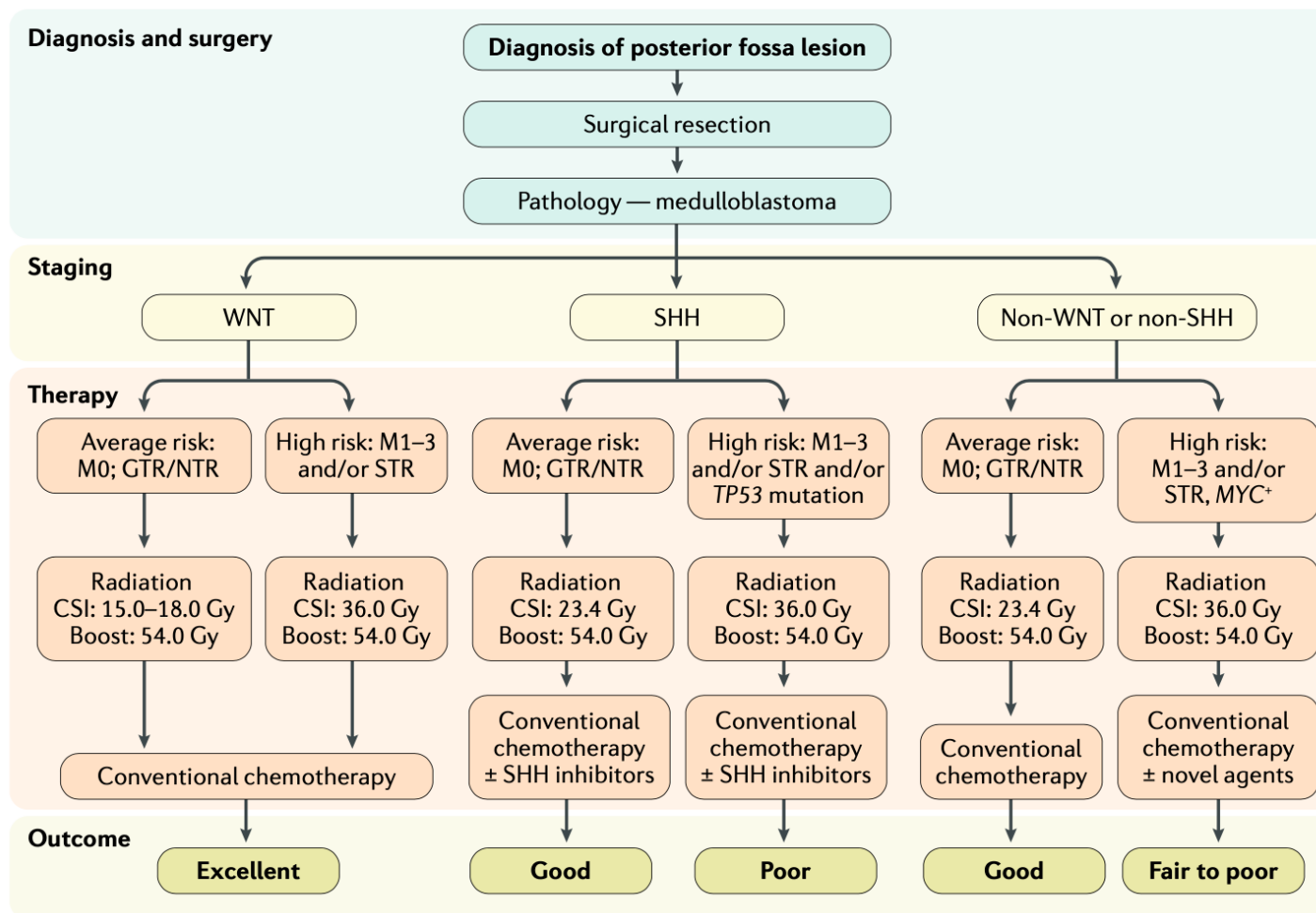
Subgroup		WNT	SHH			
Subtype			$\alpha$	$\beta$	$\gamma$	$\delta$
Demographics	Frequency (%)	100	29	16	21	34
	Age (bar height corresponds with percentage)					
	Gender (%)	45 ♂ 55 ♀	63 ♂ 37 ♀	47 ♂ 53 ♀	55 ♂ 45 ♀	69 ♂ 31 ♀
Clinical features	Histology	Classic	Classic > desmoplastic > LCA	Desmoplastic > classic	Desmoplastic > MBEN > classic	Classic > desmoplastic
	Metastasis (%)	12	20	33	9	9
	5-year OS (%)	98	70	67	88	89
Molecular features	Cytogenetics					
	Driver events	CTNNB1, DDX3X or SMARCA4 mutation	<ul style="list-style-type: none"> <li>• MYCN or GLI2 amplification</li> <li>• TP53 mutation</li> <li>• PTCH1 mutation (less)</li> </ul>	<ul style="list-style-type: none"> <li>• PTCH1 or KMT2D mutation</li> <li>• SUFU mutation/deletion</li> <li>• PTEN deletion</li> </ul>	<ul style="list-style-type: none"> <li>• PTCH1, SMO or BCOR mutation</li> <li>• PTEN deletion</li> </ul>	<ul style="list-style-type: none"> <li>• PTCH1 mutation</li> <li>• TERT promoter mutation</li> </ul>

**Figure 4.1a** – Demographic and molecular features of the current consensus molecular subgroups and subtypes of medulloblastoma (Hovestadt *et al.*, 2019).

Subgroup		Group 3							Group 4
Subtype		I	II	III	IV	V	VI	VII	VIII
Demographics	Frequency (%)	4	13	9	10	8	9	22	25
	Age (bar height corresponds with percentage)								
	Gender (%)	60 ♂ 40 ♀	77 ♂ 23 ♀	78 ♂ 22 ♀	68 ♂ 32 ♀	71 ♂ 29 ♀	67 ♂ 33 ♀	66 ♂ 34 ♀	75 ♂ 25 ♀
Clinical features	Histology	Classic > desmoplastic	LCA, classic	Classic > LCA	Classic	Classic	Classic	Classic	Classic
	Metastasis (%)	35	57	56	58	62	45	45	50
	5-year OS (%)	77	50	43	80	59	81	85	81
Molecular features	Cytogenetics	Balanced	8+ 1q+ i17q (less)	7+ 10q- i17q	14+ 7+ no i17q 8- 10- 11- 16-	7+ 16q- i17q	7+ 8- 11- i17q	7+ 8- i17q (less)	i17q
	Driver events	<ul style="list-style-type: none"> <li>• <i>GFI1</i> and <i>GFI1B</i> activation</li> <li>• <i>OTX2</i> amplification</li> </ul>	<ul style="list-style-type: none"> <li>• <i>MYC</i> amplification</li> <li>• <i>GFI1</i> and <i>GFI1B</i> activation</li> <li>• <i>KBTBD4</i>, <i>SMARCA4</i>, <i>CTDNEP1</i> or <i>KMT2D</i> mutation</li> </ul>	<ul style="list-style-type: none"> <li>• <i>MYC</i> amplification (less)</li> </ul>	<ul style="list-style-type: none"> <li>• No common driver events</li> </ul>	<ul style="list-style-type: none"> <li>• <i>MYC</i> or <i>MYCN</i> amplification</li> </ul>	<ul style="list-style-type: none"> <li>• <i>PRDM6</i> activation</li> <li>• <i>MYCN</i> amplification (less)</li> </ul>	<ul style="list-style-type: none"> <li>• <i>KBTBD4</i> mutation</li> </ul>	<ul style="list-style-type: none"> <li>• <i>PRDM6</i> activation</li> <li>• <i>KDM6A</i>, <i>ZMYM3</i> or <i>KMT2C</i> mutation</li> </ul>

**Figure 4.1b** – Demographic and molecular features of the current consensus molecular subgroups and subtypes of medulloblastoma (Hovestadt *et al.*, 2019).

After subgroups of MB were identified, it was imperative that clinically applicable assays were developed that could diagnose subgroup with high accuracy. Although current WHO clinical guidelines recognise WNT, SHH and collectively place groups 3 and 4 as non WNT/ non SHH molecular variants (Figure 4.2), distinguishing between Group 3 and Group 4 is a clinical requirement for most institutions and is beneficial for patient stratification in research and clinical trials. Initially, transcriptomic assays were developed such as the NanoString where a gene set of 22 genes were used to determine the MB subgroups (Schwalbe *et al.*, 2011; Ramaswamy *et al.*, 2013). Its clinical application was limited however, primarily due to cost, high misclassification rates, and difficulties regarding implementation in developing countries (Korshunov *et al.*, 2017). A qPCR method based on 21 biomarkers offered a reliable method of subgroup classification, but the high number of genes rendered it impractical for use clinically (Kunder *et al.*, 2013). Contemporaneously, it became clear that DNA methylation profiling was the most reliable means of subgroup classification (Schwalbe *et al.*, 2013; Hovestadt *et al.*, 2013). The four primary molecular subgroups of medulloblastoma were originally identified via gene expression, but methylation profiling is a superior alternative due to ability to robustly identify subgroups from FFPE tissue derivatives, coupled with increased accuracy above other routine methods such as mutation screening (which is hampered by the paucity of recurrent mutations in MB) and immunohistochemistry. DNA methylation is also the platform that was utilised to identify further clinically relevant heterogeneity within the medulloblastoma subgroups 3 and 4, where 8 subtypes were identified (Sharma *et al.*, 2019). Subgroup prediction via methylation microarray can subsequently guide a clinician down the correct diagnostic path and offers the potential for further targeted molecular testing, if appropriate. For example, a tumour identified as WNT can be further submitted for beta-catenin mutation analysis or an SHH tumour submitted for germline variant analysis and/or *TP53* mutation status.



**Figure 4.2** - Current risk-adapted algorithm for medulloblastoma (Northcott *et al.*, 2019). Since their identification, the molecular subgroup of a medulloblastoma tumour forms a key primary component of patient risk stratification.

The capability of the methylation array to offer accurate subgroup prediction has cemented the platform as the current gold-standard in paediatric brain tumour diagnostics. DNA methylation is a highly robust and accessible measurement and in recent years has enabled researchers to define new tumour entities and classes. The DKFZ central nervous system tumour classifier can assign a brain tumour into a class with high accuracy, originally trained using 2801 samples from 81 different tumour types (Capper *et al.*, 2018). The classifier was built upon the Illumina 450k and EPIC methylation microarrays which, due to the technical and economical limitations of more comprehensive techniques such as WGBS and RRBS, has remained the technique of choice for researchers since its inception in 2011 (Bibikova *et al.*, 2011). This has ultimately led to the accumulation of many thousands of tumour samples analysed using the platform which enabled the creation of the large reference cohorts that are necessary to create classifiers with good prediction accuracy and performance. Through collaboration with other institutes, the current DKFZ database now comprises over 50,000 tumours analysed by 450k and EPIC array, and has led to the identification of extremely rare, biologically relevant tumour types (Perez & Capper, 2020). The identification of tumour types using DNA methylation is typically done via unsupervised learning techniques, which cluster methylation patterns comprising thousands of CpG sites based on their similarities and differences. As cohort sizes increase, these techniques become more powerful, allowing the identification of rare tumour entities. Combining this analysis with machine learning allows methylation profiles for each defined tumour class to become a diagnostic tool, allowing for incoming diagnostic samples to be assigned into its respective class with an associated probability measure.

The development of the current DKFZ brain tumour classifier remains ongoing, with the continuing addition of 450k and EPIC array samples contributing to further optimisation as recently as 2021. The abundance of methylation microarray sample data and clear utility practically guarantees continued development in the coming years but the previous advantages of the platform, namely its low cost and low input sample requirements have been eroded in recent years in comparison with other techniques. Superior techniques for methylation analysis are now becoming feasible, where methylation sequencing data has demonstrated similar classification performance when integrated with methylation microarray reference sets (Kuschel *et al.*, 2021). This integration is of vital importance if the field is to progress from methylation microarray and take advantage of more advanced technologies, whilst retaining the information gleaned from the many thousands of samples analysed by methylation microarray. As demonstrated in 3.4.3, WGBS sample data generated in this project is highly correlated to matched samples analysed using methylation microarray, raising the possibility that low-pass WGBS can be successfully integrated into array-based classifiers. In addition to integration, the increased polarisation of methylation values derived from the WGBS platform may have an impact on CpG variability, of which the most variable CpGs are typically taken forward for unsupervised analyses and subgroup visualisation. When considering this potential increased variability and the fact that just 3% of all CpGs are assessed by the array, an investigation into the variability of the entire methylome would provide valuable insights into CpG variability outside of the

array probe manifest. Whilst other platforms like RRBS and Nanopore have shown success in CNS tumour classification (van Paemel *et al.*, 2020; Kuschel *et al.*, 2021), they are limited by a severe lack of CpG coverage due to the combined effects of low depth and lack of imputation. Whilst they show clear potential prognostically, the issue of data missingness magnifies as cohort sizes increase due to decreasing numbers of CpGs shared across all samples, which limits its attractiveness for use in discovery. The successful imputation of low-pass WGBS samples achieved in this project (Section 3.4.2) provides the complete data necessary to test the capability of the platform in discerning the molecular subgroups of medulloblastoma. Additionally, comparing how WGBS derived feature sets perform in classification compared to those derived from DNA methylation microarray may demonstrate the potential for the platform in contributing to further subgroup refinement in medulloblastoma and other rare tumour entities.



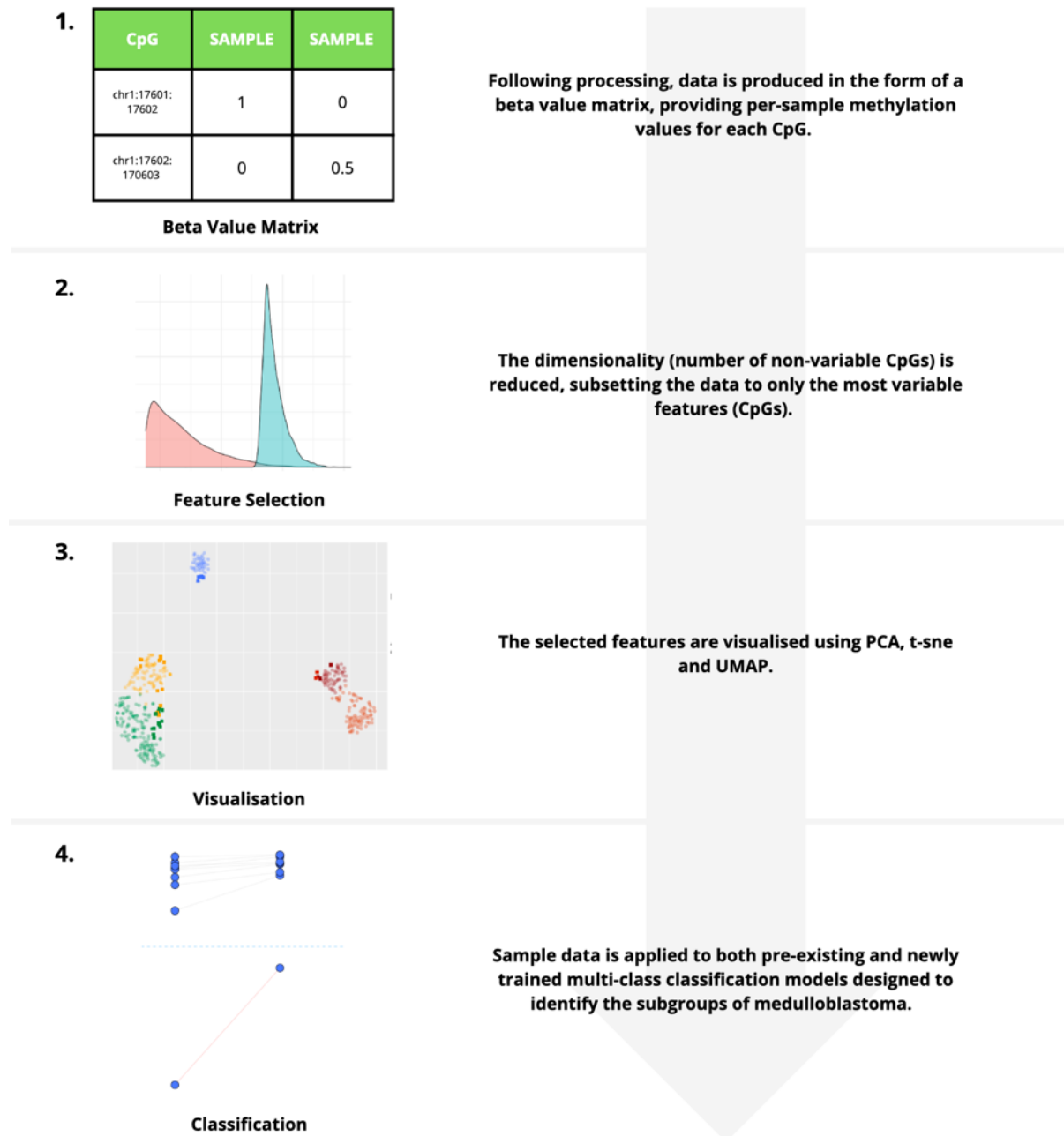
## 4.2 – Aims & objectives

The primary research objective of this chapter was to assess the potential for low-pass WGBS sample data to identify the molecular subgroups of medulloblastoma. The analyses presented in this chapter primarily are concerned with the integration of low-pass WGBS methylation data into existing array-based classifier utilities and investigating the potential performance of classifiers built solely upon WGBS-derived sample training sets. As most unsupervised high-dimensional clustering investigations initially rely on variance filtering to select only the most informative features, the variability of all CpGs in the medulloblastoma methylome of the combined WGBS cohort was surveyed and compared to the probe variability observed in the matched array cohort. Variable CpGs identified were mapped to variable probes identified using the array to determine the degree to which each set of features overlap. This enabled the determination whether there is additional variability that is captured outside of the current DNA methylation microarray probe manifest. In addition, the genomic co-ordinates of the most variable CpGs were compared to investigate any differences in the chromosomal distribution of variable CpGs/probes between the two platforms.

Following this analysis of CpG variability, the dimensionality of the combined WGBS medulloblastoma cohort was reduced to the most variable features ( $n=10,000$ ) and applied to a range of prominent dimensionality reduction/visualisation techniques to investigate whether the MB molecular subgroups were easily identified. This was conducted in parallel with both the matched array cohort and ‘Liftover’ WGBS cohort to observe whether any differences in sample clustering is observed. Finally, WGBS sample data would be integrated with DFKZ brain tumour classifier reference medulloblastoma sample data to investigate how low-pass WGBS sample data clusters amongst non-matched sample data that is currently used for clinical classification (Capper *et al.*, 2018). Following visualisation, WGBS methylation data would be applied to two pre-existing medulloblastoma classifiers whose code was made available for use in this project. Both the subgroup assignment and associated class probability for both MB WGBS cohorts would be collected and compared to that of matched array data to investigate the assessment of imputed low-pass WGBS data with pre-existing classifiers trained using array feature sets. After integration, a comprehensive performance analysis between native classifiers built upon the WGBS and Array medulloblastoma cohorts used in this study was presented. Each classifier would be trained, validated, and respective performance assessed to provide a direct comparison between both platforms when being used to train classifiers with matched sample sets. In addition, matched downsampled WGBS cohorts will also be used to train classifiers to assess the effect of lowering coverage on classification performance.

## 4.3 – Materials & methods

### 4.3.1 – Analysis Overview



**Figure 4.3** – Graphical overview of all analyses conducted within chapter 4. Further details (e.g., software, parameters used, references) are provided in subsequent sections.

### 4.3.2 – Combined cohort description

A combined cohort of 70 primary WGBS medulloblastoma samples (ICGC 30x = 34, NMB 10x = 36) and 8 cerebellar samples (Foetal = 4, Adult = 4), as prepared in section 3.3.2, were used in the following analyses. One medulloblastoma sample, NMB\_390, was a matched FFPE/fresh frozen pair. A corresponding DNA methylation microarray cohort (n=77) was also used in parallel where possible to ensure comparability at each stage of the analysis (Table 4.1). Missing values present in both datasets were imputed as described in section 3.3.9. Downsampled matched WGBS cohorts were also used (10x -> 1x) to test the effect of coverage on any results obtained. To determine the molecular subgroup and subtypes of all primary medulloblastoma samples, sample array data was applied to the DFKZ brain tumour classifier version 11b6 (Capper *et al.*, 2018). Molecular subtype was determined using the Group3/4 classifier as defined by Sharma *et al.*, (2019).

Tumour subgroup composition was chosen carefully for the NMB WGBS cohort, with the majority of samples chosen being classified with high probability in their respective molecular subgroup according to the DFKZ brain tumour classifier. Molecular subtype composition was considered, with a deliberate over-representation of subtype VII samples, due to molecular and clinical evidence that a possible split may be present within this subtype (Sharma *et al.*, 2019).

<b>Characteristic</b>	<b>External Cohort (%)</b>	<b>Internal Cohort (%)</b>	<b>Combined (%)</b>
<b>No. of Samples</b>	<b>34</b>	<b>36</b>	<b>70</b>
<b>Sex</b>			
Male	<b>20</b> (58.8%)	<b>27</b> (75%)	<b>47</b> (67.1%)
Female	<b>14</b> (41.2%)	<b>9</b> (9%)	<b>23</b> (32.9%)
<b>Age at Diagnosis</b>			
< 3	<b>4</b> (11.8%)	<b>3</b> (8.3%)	<b>7</b> (10%)
3-16	<b>29</b> (85.3%)	<b>30</b> (83.3%)	<b>59</b> (84.3%)
>16	<b>1</b> (2.9%)	<b>3</b> (8.3%)	<b>4</b> (5.7%)
<i>Mean</i>	7.66	7.14	7.4
<i>Standard Deviation</i>	4.65	3.89	4.26
<b>Metastatic Stage</b>			
M+	<b>20</b> (58.8%)	<b>23</b> (63.9%)	<b>43</b> (61.4%)
M-	<b>12</b> (35.3%)	<b>13</b> (36.1%)	<b>25</b> (35.7%)
Unknown	<b>2</b> (5.9%)	<b>0</b>	<b>2</b> (11.8%)
<b>Histology</b>			
Classic	<b>20</b> (58.8%)	<b>23</b> (63.9%)	<b>43</b> (61.4%)
Biphasic Classic	<b>0</b>	<b>3</b> (8.3%)	<b>3</b> (4.3%)
Desmoplastic/Nodular	<b>5</b> (14.7%)	<b>3</b> (8.3%)	<b>8</b> (11.4%)
LCA	<b>9</b> (26.5%)	<b>7</b> (19.4%)	<b>16</b> (22.9%)
<b>Gene Amplifications</b>			
MYC amplification	<b>1</b> (2.9%)	<b>4</b> (11.1%)	<b>5</b> (7.1%)
MYCN amplification	<b>2</b> (5.9%)	<b>3</b> (8.3%)	<b>5</b> (7.1%)
<b>Molecular Subgroup<sup>1</sup></b>			
WNT	<b>5</b> (14.7%)	<b>5</b> (13.9%)	<b>10</b> (14.3%)
SHH	<b>6</b> (17.6%)	<b>5</b> (13.9%)	<b>11</b> (15.7%)
Group 3	<b>11</b> (32.4%)	<b>12</b> (33.3%)	<b>23</b> (32.9%)
Group 4	<b>12</b> (35.3%)	<b>14</b> (38.9%)	<b>26</b> (37.1%)
<b>Group 3/4 Molecular Subtype<sup>2</sup></b>			
Total No.	<b>23</b>	<b>26</b>	<b>49</b>
I	<b>1</b> (4.3%)	<b>2</b> (7.7%)	<b>3</b> (6.1%)
II	<b>4</b> (17.4%)	<b>3</b> (11.5%)	<b>7</b> (14.3%)
III	<b>1</b> (4.3%)	<b>3</b> (11.5%)	<b>4</b> (8.2%)
IV	<b>0</b>	<b>3</b> (11.5%)	<b>3</b> (6.1%)
V	<b>0</b>	<b>4</b> (15.4%)	<b>4</b> (8.2%)
VI	<b>5</b> (21.3%)	<b>1</b> (3.8%)	<b>6</b> (12.2%)
VII	<b>5</b> (21.3%)	<b>8</b> (30.8%)	<b>13</b> (26.5%)
VIII	<b>4</b> (17.4%)	<b>2</b> (7.7%)	<b>6</b> (12.2%)
MBNOS	<b>3</b> (8.8%)	<b>0</b>	<b>3</b> (6.1%)

Table 4.1 – Demographics of combined medulloblastoma cohort (n=70)

### 4.3.3 – Computing resources/software

All subsequent analyses described in sections 4.3.3 and 4.3.4 were conducted using the R statistical computing environment version 4.0.3 – “Bunny-Wunnies Freak Out” within RStudio (R Core Team, 2021). Native classifier training steps detailed in section 4.3.5 was conducted using the ROCKET High Performance Computing service (Newcastle University).

### 4.3.4 – Analysing methylation variability

Prior to unsupervised analysis of methylation data, combined cohort beta value matrices for WGBS and Array cohorts were transposed and isolated to the top 100,000, 32,000, 10,000 and 1,000 CpGs/probes as defined by a standard deviation measure calculated for each feature (e.g., row of beta values for all samples). For each feature set, the genomic co-ordinates (hg19) of each feature were recorded, and the chromosomal distribution of each feature set tabulated.

To determine the degree at which the most variable features identified via WGBS overlap with features identified by DNA methylation microarray, each feature set (cast as a data frame RObject) was subsequently converted into GRange objects and checked for overlapping ranges with both the Illumina Human Methylation 450k and EPIC probe manifests. Any CpG identified within the feature set derived from the WGBS cohort that mapped exactly with the genomic co-ordinates specified in each manifest was considered a match. Statistically significant differences in the chromosomal distribution of the top 10,000 features between Array / Liftover / WGBS cohorts were determined via two proportion z-test. As the WGBS platform surveys singular CpGs and each probe assessed by the DNA methylation microarray is intended to encompass a range of ~50 CpGs (owing to the assumption that neighbouring CpGs typically exhibit correlated methylation status) (Illumina, 2021), Feature sets were overlapped with each manifest again with decreased precision, considering each CpG as overlapping with a probe site if it fell within 25, 50 and 125bp either side of a probe site.

### 4.3.5 – Dimensionality reduction/visualisation

Following variance filtering, each feature set for WGBS (including all downsampled datasets) and DNA methylation microarray platforms was visualised using multiple dimensionality reduction techniques – Principal component analysis (PCA), uniform manifold approximation and projection (UMAP) and t-distributed stochastic neighbour embedding (t-sne) (Section 2.3.4). Cohorts were visualised with and without control cerebellar samples and colour assigned according to their respective molecular subgroup and subtype as defined by DNA methylation microarray classification via the DFKZ brain tumour classifier version 11b6.

For PCA, principal components were calculated for each feature set using the `prcomp` function within the `stats v4.0.3` package using default settings (R Core Team, 2021). UMAP was performed using the `umap` function of the `umap` package using a Euclidean metric (Konopka, 2020). The seed was preserved. The neighbour's parameter was adjusted depending on the number of samples used (Table 4.2). T-sne was computed using the `Rtsne` function within the `rtsne` packages, version 0.15 using a seed of 25 (Krijthe, 2015). Default parameters were applied excluding `theta = 0` `max_iter = 100,000`. Perplexity was adjusted depending on the number of samples used (Table 4.2). A second t-sne analysis was performed using settings as defined in the landmark paper by Capper *et al.*, (2018) used to cluster the original DFKZ brain tumour classifier reference cohort. Briefly, an eigenvalue decomposition of all principal components (`no. components = total cohort size - 1`) using the `eigs` function of the `RSpectra` package 0.12 was initially performed. All principal components were then brought forward for t-sne analysis using the same parameters defined above.

Algorithm	Platform	Parameter	All Samples	MB Only (All subgroups)	MB Only (Group 3/4)
t-sne	WGBS Array	<i>perplexity</i>	25	23	16
UMAP	WGBS Array	<i>n_neighbors</i>	38	35	38

**Table 4.2** – Adjusted t-sne and UMAP parameters used during visualization for each dataset. Both perplexity (tsne) and `n_neighbors` (UMAP) parameters have a significant effect on the degree to which each algorithm clusters neighbouring data points.

To investigate how WGBS sample data clusters amongst non-matched DNA methylation microarray data, a final t-sne plot was calculated using the WGBS samples in this study with the medulloblastoma sample component of the DFKZ reference cohort ( $n=426$ ). Samples were acquired using the from the NCBI Gene Expression Omnibus (GEO), accession number GSE109381. The same variable features ( $n = 32,000$ ) were used as defined by Capper *et al.*, (2018) by performing variance filtering on the entire reference cohort ( $n=2801$ ) prior to subsetting to just MB tumour samples. WGBS samples were then combined with the dataset at co-ordinates corresponding to each respective probe site and t-sne calculated using the same parameters as defined previously (Capper *et al.*, 2018).

### **4.3.6 – Applying WGBS sample data to existing DNA methylation microarray classifiers to predict the subgroups of medulloblastoma**

To test the capability of WGBS derived sample data to be classified into the subgroups of medulloblastoma, WGBS beta value matrices were applied to the Methylation Array Classifier (MAC) (Newcastle University, 2021) using two training feature sets ( $n=10,000$ ) that define the four classic molecular subgroups of medulloblastoma and the Schwalbe *et al.*, (2017) seven subgroup classification. Briefly, the classifier is based upon the support vector machine algorithm, validated, and trained using 220 450k samples. WGBS sample data could not be directly applied to the DFKZ brain tumour classifier as the classifier and its associated development code is not publicly available. Whilst array-based classifiers only accept IDAT files (Raw methylated/unmethylated probe intensities) it is beta values that are ultimately used for training models and testing incoming samples. This enables WGBS sample data to easily be integrated into DNA methylation microarray classifiers as beta values are also used. Combined WGBS (at each downsampled level of coverage) and matched array samples were applied to both classifier training sets, and each class and its associated probability was tabulated. Statistical differences in mean classification probability were calculated as stated in 3.3.5.

### **4.3.7 – Performance assessment of validated multi-class models trained using WGBS derived methylation data**

To investigate the potential capability of MB subgroup classification models trained using feature sets derived from the WGBS platform, numerous multi-class classification models using samples sequenced in this study and compared them to like-for-like models trained using matched array samples with array-defined features. For each classifier, a total of 77 samples were used (the matched FFPE sample was removed from WGBS cohort to ensure identical training set size and composition) in a 10-times repeated 7-fold cross-validation. Briefly, the cohort was split into 7 parts (folds), with 6 folds used for model training and 1-fold used for model testing. The classifier is trained 7 times, with each different fold left out one time. This process is subsequently repeated 10 times, with the sample composition in each fold randomly selected for each repeat. The folds are kept identical for all models by using the same randomised seed value prior to dataset splitting. Various combinations of parameters were tested for each model a total of 25 times randomly (Table 4.3). All model training and testing was performed within the caret software package in RStudio (Kuhn, 2021).

To determine performance, we assessed models using the following metrics:

- **Accuracy** – The fraction of correct class label assignments. A score of 1 would denote perfect classifier performance
- **Cohen’s kappa statistic** – An adjusted accuracy measure that accounts for assignments are correct by random chance according to the frequency of each class. Like accuracy, a score of 1 would represent perfect performance after according for chance.
- **Area under receiver operating characteristic curve (AUC)** – A measure of the ability of the model to correctly predict class. To determine this, a curve is plotted denoting the true positive rate against the false positive rate, where the AUC corresponds to the total area underneath the curve. Better performing models display AUC values closer to 1, with 1 being a perfect score.
- **Log-loss** – The negative average of the log of each assignment probability. Calculated values are produced for each sample that class assignments with lower probabilities or incorrect assignments are penalised more heavily, producing greater log loss values. A perfect classifier that assigns each class correctly with maximum probability would be represented with a log loss value of 0, with poorer models producing high log loss values.

A total of seven multi-class classification algorithms capable of assigning class probability were tested: elastic net, random forest, support vector machine with linear kernel, support vector machine with radial kernel, XGBoost with linear booster and XGBoost with tree booster, and logic regression. All models were trained to predict the four primary molecular subgroups of medulloblastoma as well as recognise control cerebellar tissue using the top 10,000 most variable features (WGBS = CpGs, Array = probes, Liftover = CpG sites that map to corresponding probe sites on the 450k probe manifest) as defined using a standard deviation measure.



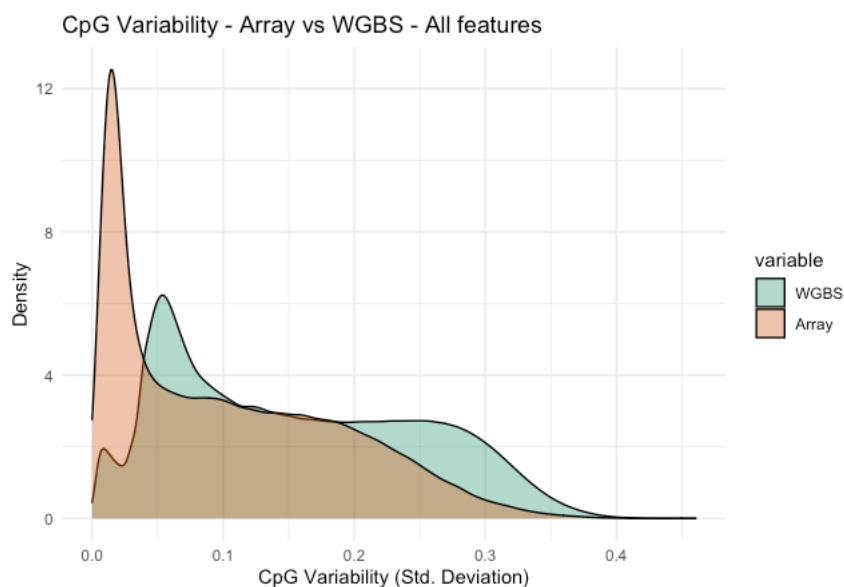
<b>Algorithm</b>	<b>Package</b>	<b>Parameters Tuned</b>
<b>ElasticNet</b>	glmnet	alpha, lambda
<b>Random Forest</b>	parRF	mtry, ntree
<b>SVM (Linear kernel)</b>	e1071	cost
<b>SVM (Radial kernel)</b>	kernlab	sigma, C
<b>XGBoost (Linear Booster)</b>	xgboost	nrounds, lambda, alpha, eta
<b>XGBoost (Tree Booster)</b>	xgboost	nrounds, max_depth, eta, gamma, colsample_bytree, min_child_weight, subsample
<b>Logic Regression</b>	LogicReg	treesize, ntrees

**Table 4.3** – Chosen models for performance testing. All tuneable parameters within the caret software framework were tuned 25 times using randomly selected values.

## 4.4 – Results & discussion

### 4.4.1 – CpGs of greater variability are present within the medulloblastoma methylome that are not assessed by the DNA methylation microarray platform

Analysing CpG methylation values in our sequencing cohort compared to the probe methylation values in our matched DNA methylation microarray cohort revealed multiple differences between both platforms. Significantly larger numbers of CpGs capture greater variability than probes assessed by the DNA methylation microarray platform (Figure 4.4). A standard deviation measure of 0.25 is typically applied when concerned with dimensionality reduction in medulloblastoma methylation datasets, at which point the increased variability captured by WGBS becomes clear. A total of 3,990,549 CpGs sequenced using WGBS were retained when applying a standard deviation filter of 0.25 compared to just 27,211 probes sequenced using the array for matched samples. The top 10,000 most variably methylated probes display standard deviation values ranging from 0.17-0.43 compared to 0.40-0.46 and 0.38-0.45 for the top 10,000 most variably methylated CpGs WGBS samples when applying a depth filter of 3 or 5 respectively. There are multiple technical explanations that can possibly explain the increased variance observed. Beta values display increased polarisation when analysed using WGBS, ranging from 0-1 compared to probe intensity beta values which typically range from 0.1-0.9, inherently lowering the possible standard deviation value that can be calculated from methylation data analysed using the array platform. Beta values calculated using low-depth WGBS may also contribute to increased variability, with sequencing errors contributing to larger changes in any methylation value that is calculated for any respective CpG site. However, any such effect is likely to be minimal, given the high correlation of methylation data between both platforms (Section 3.4.3).



**Figure 4.4** – Density plot of standard deviation value distribution of all features for WGBS (green, n = 26,549,033 CpGs) and Array (red, n = 422,027 probes).

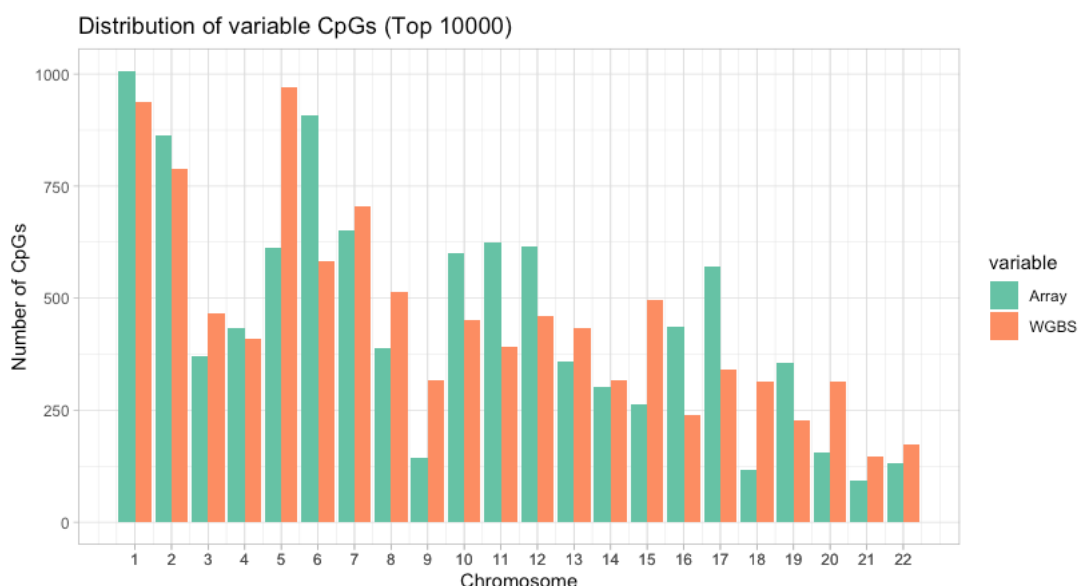
Aside from technical differences, such a stark increase in variability can also be explained by the sheer number of CpGs that are not measured by the array, with just 3% of CpGs assessed by the methylation EPIC platform (Zou *et al.*, 2018). The DNA methylation microarray typically contains a single probe at each position of interest, relying on the assumption that neighbouring CpGs exhibit correlated methylation (Eckhardt *et al.*, 2006). The manufacturer states that each probe is typically representative of a range of 50 CpGs within any given site of interest (Illumina, 2021). The array platform deliberately targets regions of genetic importance, namely CpG islands, genes, promoter, and enhancer regions, and therefore overlooks the large swathes of intergenic and repetitive regions of the genome. Such regions have been shown to display global demethylation in cancer, with many hypomethylated regions shared among different tumour types (Baylin & Jones, 2011). Despite WGBS being performed in MB previously, a comprehensive analysis of genome-wide methylation was not reported (Hovestadt *et al.*, 2014). To investigate whether any of the increased variability observed was contributed by CpGs not covered by the methylation microarray probe manifest, the top 10,000 most variable CpGs assessed by the methylation microarray platforms (both 450k and EPIC) were overlapped using their respective genomic location. When looking for exact CpG/probe matches, just 2.9% and 4.4% of features overlap with the 450k and EPIC probe sites respectively (Table 4.4). To account for local correlations in CpG methylation status, the range either side of the CpG of interest was increased to 25bp, 50bp and 125bp to be classed as overlapping with any probe site. Unsurprisingly, the number of sites determined to be overlapping increases, with 36% and 42% of probe sites found within 125bp either side of any probe site assessed by the 450k and EPIC platforms respectively (Table 4.4). This significant proportion is likely by design, owing to the deliberate selection of regions of genomic interest during probe selection of the methylation microarray platform. However, over 50% of the most variable CpGs are not assessed by the methylation microarray, indicating that there are many CpGs that capture significant variance that are simply not covered by the array even when applying a range of 125bp either side of a probe site.

<b>CpG distance from probe site</b>	<b>No. of overlapping features - 450k Manifest</b>	<b>No. of overlapping features - EPIC Manifest</b>
<b>Exact</b>	291	439
<b>25bp</b>	1,036	1,270
<b>50bp</b>	1,702	2,000
<b>125bp</b>	3,631	4,214

**Table 4.4** – Number of top 10,000 most variable CpGs identified via WGBS that overlap with probes assessed by the Illumina methylation 450k and EPIC microarrays.

## 4.4.2 – WGBS derived variable CpGs are differentially distributed across chromosomes compared to the array platform

Following the identification of a large proportion of variable CpGs in the medulloblastoma methylome, that are not assessed by the DNA methylation microarray probe manifest, the chromosomal distribution the top 10,000 most variable CpGs compared to the most variable probes for matching array samples (Figure 4.5) was investigated. The proportion of features changes significantly within 20 out of 22 chromosomes (not including sex chromosomes) when analysing the most variable CpGs sequenced via WGBS compared to probes analysed via microarray (Two proportion z-test where  $p < 0.05$ ) (see appendix – 8.1.5). In addition, the methylation microarray probe manifest does not assess all chromosomes equally by design. For example, chromosome 9 is assessed by just 2% of all probes in the 450k manifest, where 1.43% of variable probe sites are found in the array cohort. The proportion of variable CpGs found in WGBS samples is significantly higher, with 3.18% of the top 10,000 most variably methylated CpGs found in this chromosome ( $p = < 0.005$ ). Out of 10,000 features, the greatest proportion of variable features assessed by the microarray are located on chromosome 1 (10.07%) compared to 9.38% of features when using WGBS ( $p = 0.024$ ). Since chromosome 1 is the largest chromosome, the presence of significant proportions of variable features is expected, but out of all chromosomes analysed, it does not contain the greatest proportion of variable CpGs when sequencing the medulloblastoma methylome via WGBS. Out of the top 10,000 most variable CpG sites, 9.7% are located on chromosome 5 ( $p < 0.005$ ) and over 10% of the top 1000 CpGs are located on chromosome 7 ( $p < 0.005$ ).

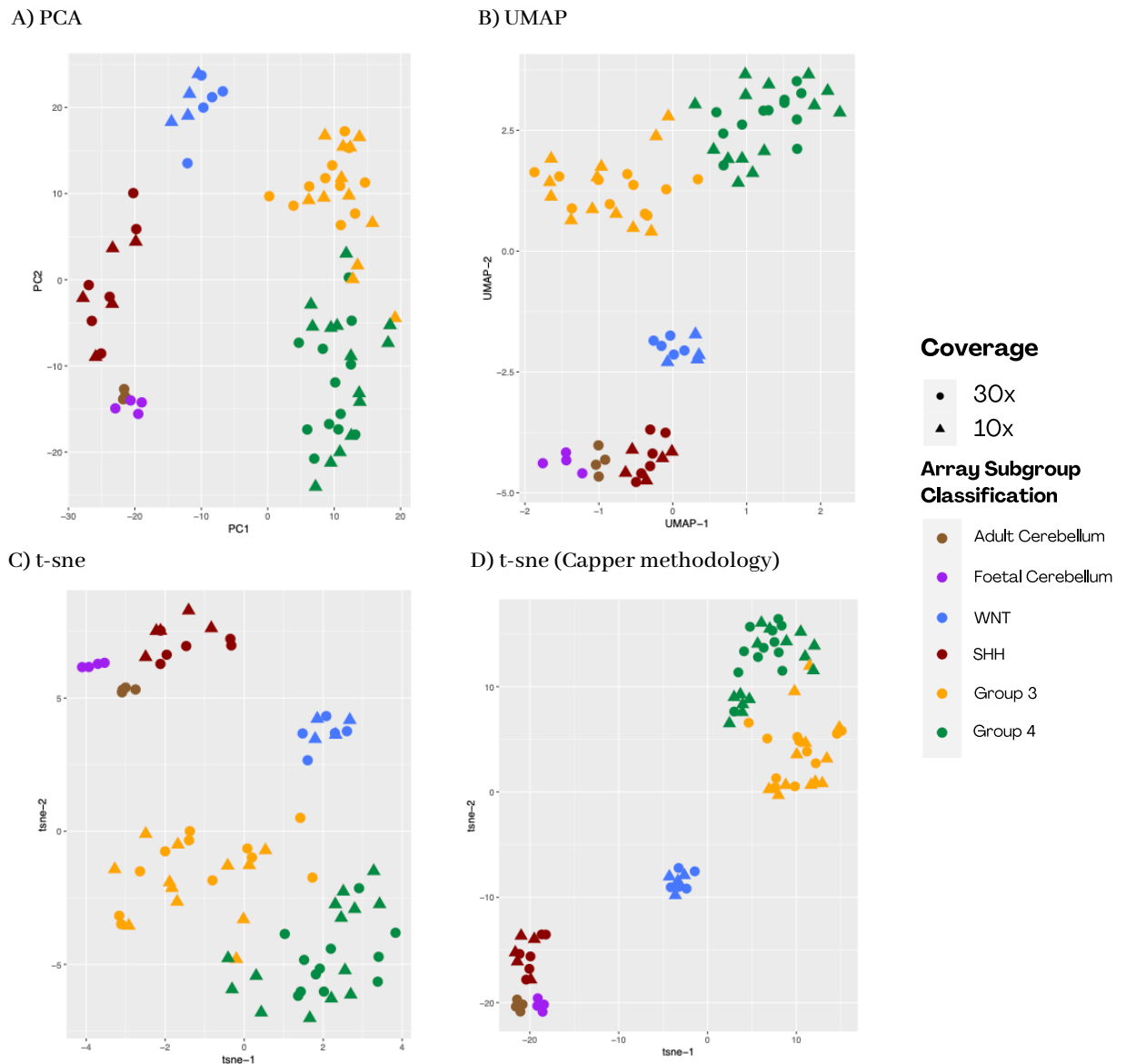


**Figure 4.5** – Chromosomal distribution of the top 10,000 most variable features derived from both WGBS and Array medulloblastoma cohorts

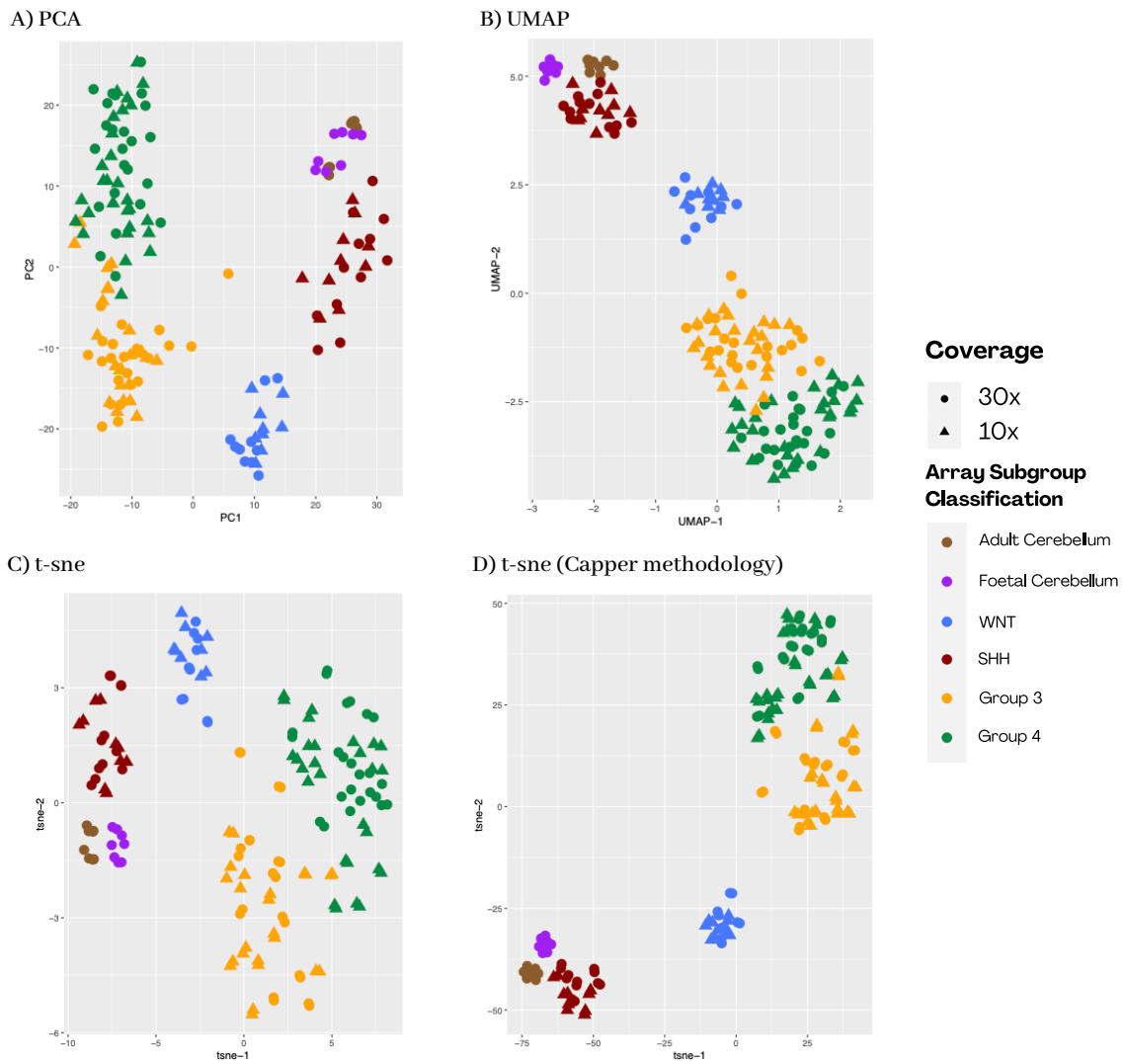
It is difficult to comment any further on why CpGs are distributed in this manner when analysing the methylome as no further related analyses were conducted during this project. However, it became clear that there is substantial, variable methylation occurring at CpG loci located in InterCGI regions (described as CpGs located at least 4kb away from CpG islands) located outside of probe sites assessed by the DNA methylation microarray that may be of biological interest (Figure 4.6). InterCGI regions are typically methylated, and large amounts of differential methylation occurs within these regions (McCabe *et al.*, 2009) in cancer cells and during cell differentiation. InterCGI regions are typically intergenic, located far from CpG islands which are located within promoters and are known to harbour regulatory elements, although research in this area remains lacking (Nelson, Hersh & Carroll, 2004). The discovery of increased CpG variability within the medulloblastoma methylome therefore warrants further investigation, given the potential for the methylation patterns of each subgroup of medulloblastoma to be more comprehensively defined.

### **4.4.3 – Molecular subgroups of medulloblastoma are recapitulated when visualising cohorts sequenced via WGBS**

Following variance filtering, various dimensionality reduction techniques were used to attempt to visualise the molecular subgroups of medulloblastoma using the most variable CpGs/probe sites in our WGBS/Array medulloblastoma cohorts. Dimensionality reduction techniques such as PCA, t-sne and UMAP (Section 2.3.4) are often used when dealing with high-dimensional data (McInnes, Healy & Melville, 2020; van der Maaten & Hinton, 2008). In this study, PCA, t-sne and UMAP were applied to the top 10,000 most variable features to our combined WGBS and Array medulloblastoma cohorts. As expected for the DNA methylation array cohort, the molecular subgroups of medulloblastoma are visually distinct regardless of the technique applied (Figure 4.7). Both NMB and ICGC samples cluster amongst each other, indicating that any batch effects from each cohort are minimal and that appropriate pre-processing was applied to the NMB cohort. WNT tumours cluster separately, as do SHH tumours. SHH tumours cluster near that of cerebellar tissue, in line with the findings that tumours of this subgroup originate from cerebellar granule neuron progenitor cells (Tamayo-Orrego & Charron, 2019). Unsurprisingly, both Group 3 and 4 tumours cluster more closely together, with some overlap observed between samples of each group. Clustering both Array and WGBS ‘Liftover’ data together shows that matched samples were preserved as sample pairs within their respective subgroup (Figure 4.8). This demonstrates that the high correlations between samples assessed by WGBS and array is being translated into similar findings when applying identical analyses to the data, which is encouraging when considering the future integration of WGBS sample data into existing array classifier utilities.



**Figure 4.6** – Visualisation of the top 10,000 most variable features derived from our combined medulloblastoma DNA methylation microarray cohort (n=77). Samples denoted as 30x (circles) describe the ICGC cohort, whereas 10x samples (triangles) denote the NMB cohort

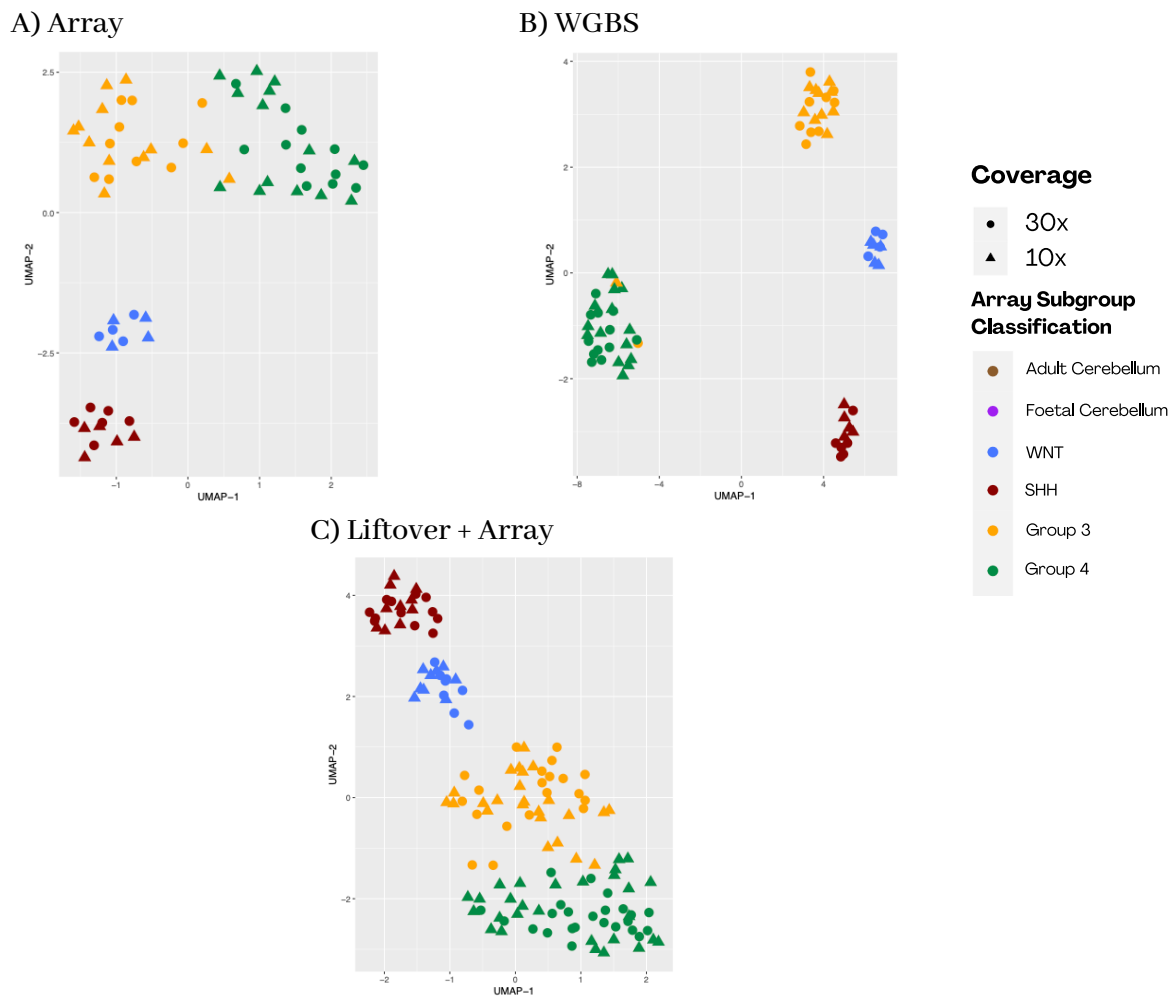


**Figure 4.7** – Visualisation of the top 10,000 most variably methylated CpGs within the medulloblastoma WGBS cohort (n=77), subset to CpGs that map directly to probe sites located on the Illumina DNA Methylation 450k microarray. Samples denoted as 30x (circles) describe the ICGC cohort, whereas 10x (triangles) describe the low-pass NMB cohort.

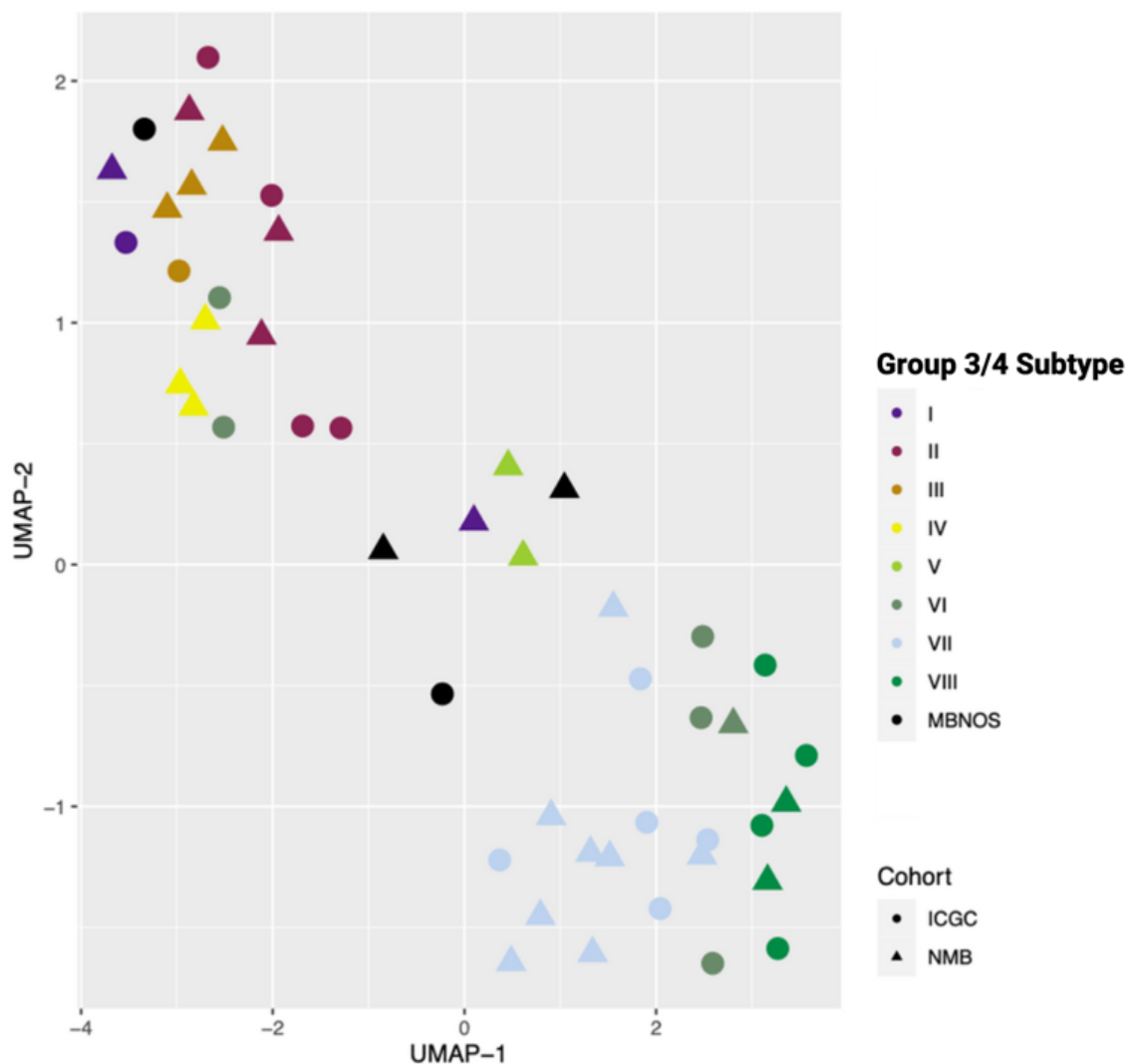
When applying the same techniques to the top 10,000 most variably methylated CpGs for the combined WGBS cohort, the molecular subgroups of medulloblastoma are maintained, but with a degree of greater separation observed between the groups (Figure 4.9). The WNT and SHH molecular subgroups appear to cluster more densely, and groups 3 and 4 appear to cluster with slightly increased separation. This is highly encouraging, given that the WGBS cohort used involves samples sequenced at least seven years apart by different institutions at different levels of coverage (Hovestadt *et al.*, 2014). This increased separation observed is likely due to several reasons, not least of which the deliberate selection of samples that are confidently classified with high probability contributing to the clear separation observed. Groups 3 and 4 exhibit clear separation in the WGBS cohort save for 2 samples, one from each cohort (Figure 4.9). These samples, ICGC\_MB5 and NMB\_772, are classified by the DFKZ paediatric brain tumour classifier as a group 3 with intermediate probability (0.86) and a group 3 with very low probability (0.56). For NMB\_772, there does not appear to be any clinical features that are distinctly characteristic of group 3 (i.e., *MYC* amplification), but it was also classified as a subtype V with low probability, which comprises tumours belonging to both groups 3 and 4 (Sharma *et al.*, 2019).

The fact that the subgroups 3 and 4 cluster with such a clear degree of separation when using WGBS-derived CpGs is cause for further investigation into intergenic methylation between the two subgroups, given the importance of defining group 3 and 4 more clearly given the clinical implications of incorrectly classifying tumours. It is recognised however, that the number of samples analysed in this study is modest and are by no means irrefutable evidence that groups 3 and 4 are clearly separate entities when considering the whole methylome. When visualising this cohort by subtype, there does not appear to be any separation between subtypes I and VI, which represents samples belonging to both subtypes (Figure 4.10). Regardless, it is important that additional samples (including ones that are currently difficult to classify) are analysed via WGBS to investigate whether this clear separation continues to persist.



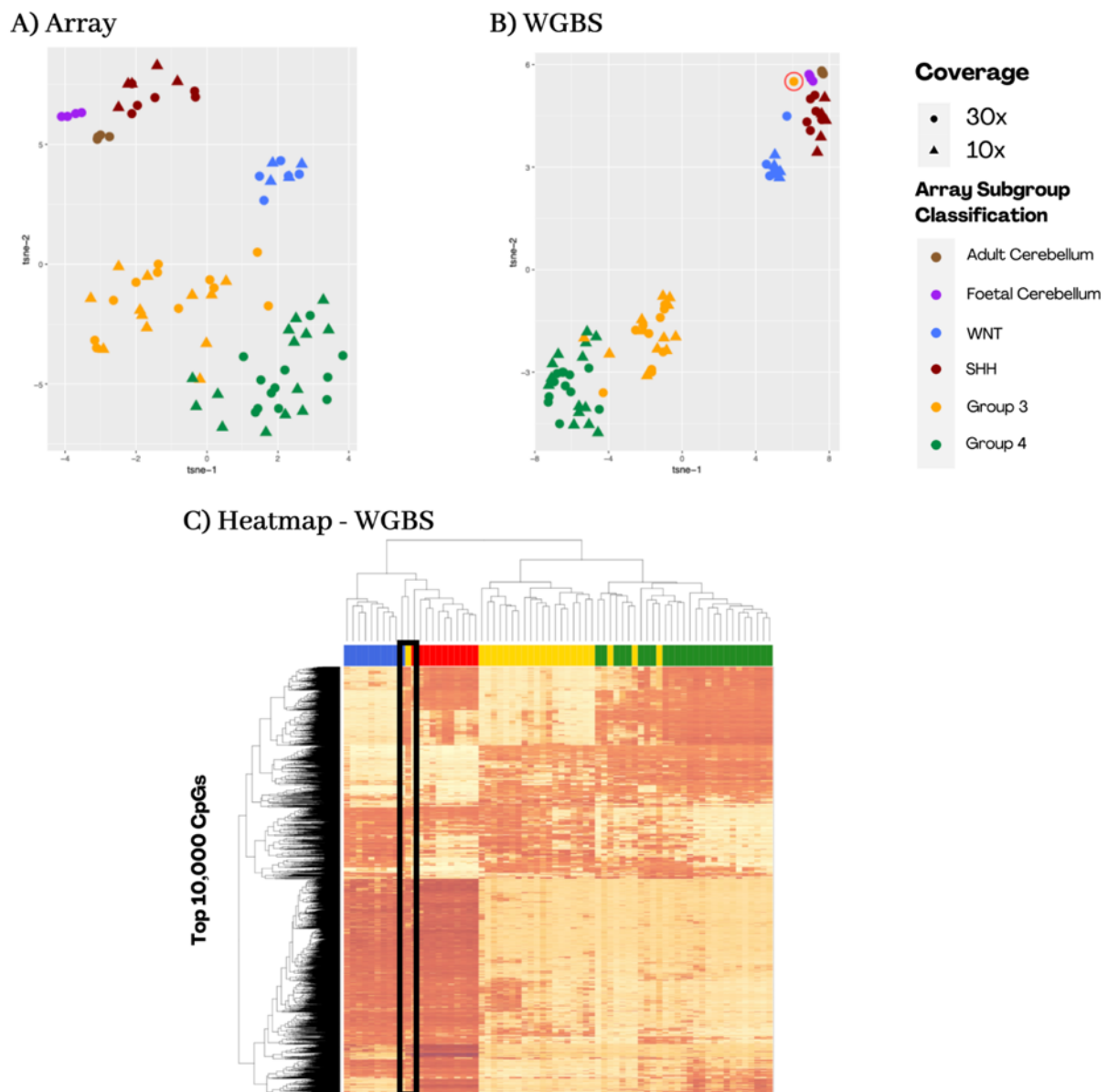


**Figure 4.8** – UMAP clustering of the top 10,000 most variable features as identified within our array ( $n = 69$ ), Liftover ( $n = 138$ ) and WGBS ( $n = 69$ ). The removal of cerebellar samples enables UMAP to focus solely on differences between the subgroups of medulloblastoma. UMAP groups highly similar samples closely together, where the presence of two group 3 samples within the group 4 cluster may be indicative of possible misclassification for these samples.



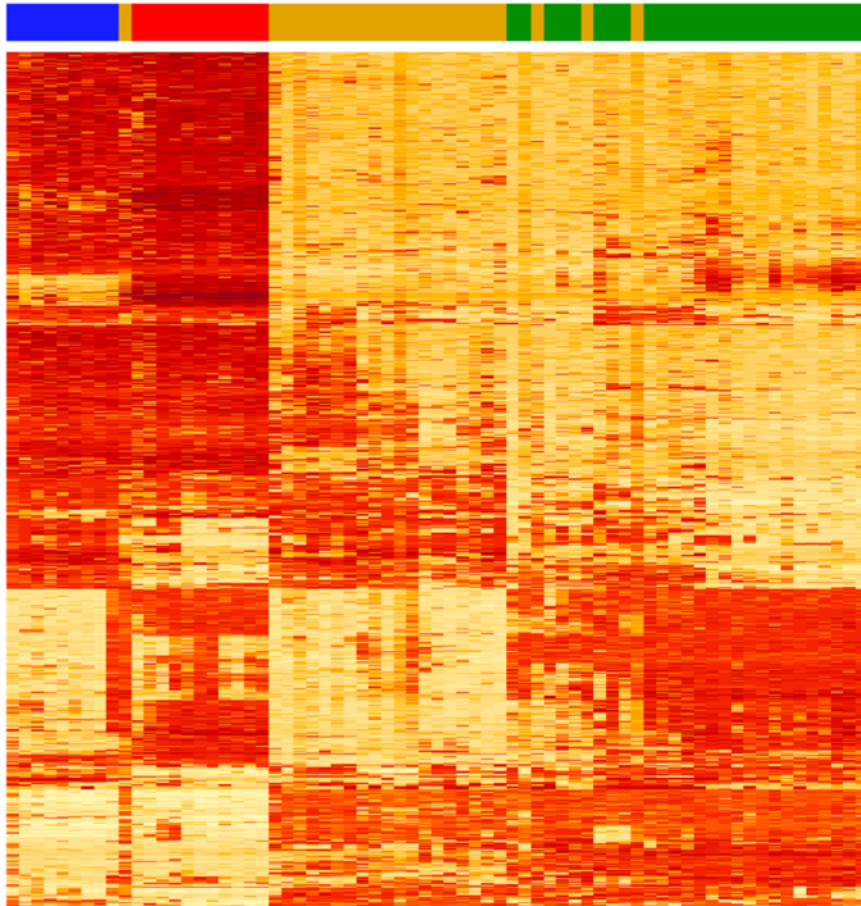
**Figure 4.9** – UMAP clustering of the top 10,000 CpGs identified for the WGBS cohort, restricted to Group 3 and Group 4. Samples are coloured by Group 3 / 4 consensus molecular subtype, defined by Sharma *et al.*, (2019).

The application of dimensionality reduction techniques is incredibly useful for exploratory data analysis, helping to indicate whether features within the dataset may be used successfully to make predictions, given the degree of class separation that is present within a dataset. It can also be used to identify any outliers that are present within the dataset that may not have been clearly seen during prior processing. One such sample was identified as a possible outlier sample; ICGC\_MB17 (Figure 4.10). Further investigation of the methylation levels present within this sample showed high levels of intermediate methylation across all top 10,000 features selected. The reasons for this are currently unclear, as there was no indication that the sample was of poor-quality during QC. Further investigation of this sample and how it behaves during classification is therefore necessary to establish whether its removal is necessary for downstream analysis.

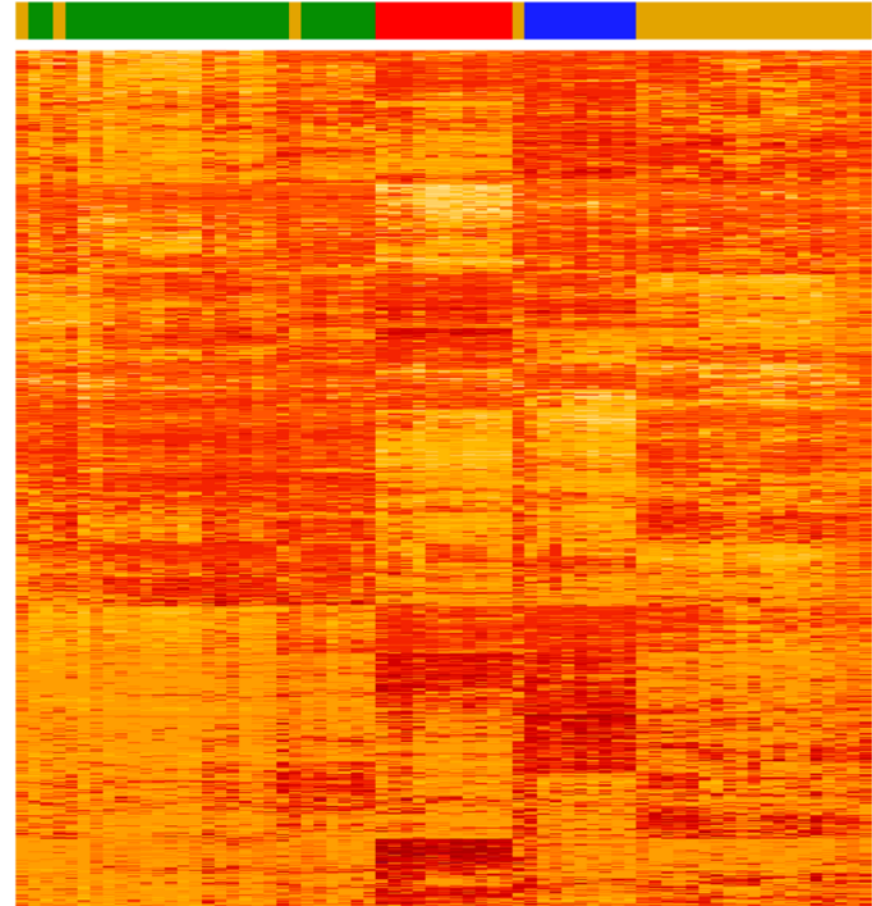


**Figure 4.10** Identification of outlier sample in WGBS cohort (ICGC\_MB17 – outlined in black). A) Clustering of matched array data reveals the sample does not cluster abnormally, B) but is clearly identified when clustering our WGBS samples (denoted with a red circle). C) Plotting methylation levels of all samples via heatmap (denoting beta values of 0-1 by a yellow-red colour intensity scale) reveals a stripe like pattern caused by this sample, characterised by high levels of intermediate methylation present in almost all the top 10,000 features analysed.

# WGBS

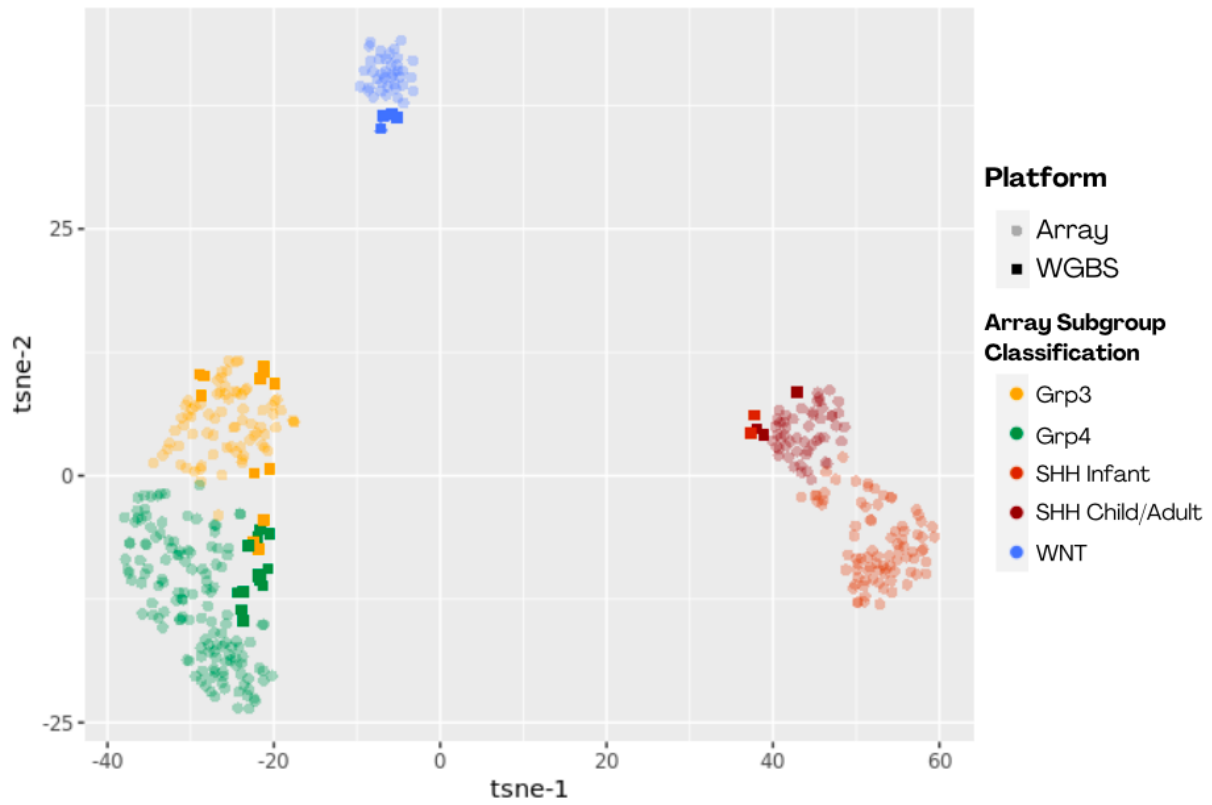


# Array



**Figure 4.11** – Hierarchical clustering of the top 10,000 most variable features in our WGBS and Array medulloblastoma cohorts. Each cell in the heatmap corresponds to a feature (CpG / probe), denoting beta value using a colour intensity scale ranging from yellow (beta = 0) to red (beta = 1).

To investigate the degree to which WGBS sample-derived beta values cluster with non-matching array samples, we clustered our combined WGBS cohort with the reference medulloblastoma cohort that formed part of the initial DFKZ paediatric brain tumour classifier cohort ( $n = 426$ ) (Capper *et al.*, 2018). Following variance filtering of the entire cohort to ensure the same feature set was used, we integrated our WGBS MB sample CpG beta values at all 32,000 matching probe sites and clustered using t-sne (Figure 4.12). Encouragingly, WGBS sample data clusters by its respective subgroup, demonstrating that imputed WGBS sample data clusters well with array sample data sourced and analysed from different institutions. The differences in platform were evident, however, where WGBS samples were typically located on the periphery of their respective primary subgroup clusters. This is most pronounced in the SHH cluster where WGBS MB samples display their own Infant/Child-Adult subtype split on the distal edges of the SHH array cluster. T-sne clusters closely related samples nearer together, explaining both the correct clustering displayed by WGBS MB samples and their respective positions on the edges of each cluster as WGBS derived beta values are numerically more like each other rather than other array samples. This behaviour is only apparent when clustering with non-matched sample data, whereas sample pairs cluster more closely when performing t-sne with our matched cohorts. Regardless, such findings are of note, given the clear inter-operability that is seen between sample data derived from both platforms. The fact that MB data from both platforms cluster tightly amongst their respective subgroups indicates that it is possible that array-trained classification tools (like that of the DFKZ paediatric brain tumour classifier) offer potential to classify WGBS sample data, given their similarity to that of samples used in current classifier reference sets in dimension reduction visualisations.



**Figure 4.12** – t-sne: Overlaying WGBS samples onto MNP reference medulloblastoma samples (n = 426).

#### 4.4.4 – Medulloblastoma samples sequenced via low-pass WGBS can successfully be assigned into subgroup using pre-existing DNA methylation microarray classifiers

Regarding classification, the integration of both platforms was hypothesised to be possible given that methylation data from both platforms are highly correlated and can be analysed in the same format (beta values). Provided the issue of missing data is resolved by imputation, WGBS sample data could theoretically be applied to classifier tools following processing, where WGBS sample data (subset to respective array-defined probe sites) effectively mimics that of any other array test sample. Most publicly available classifiers, including the DFKZ brain tumour classifier, accept sample data only in the IDAT format – the raw data format for DNA methylation microarray sample data. Unless the software is publicly available, allowing beta values to be integrated at the processed data input stage, it is impossible to use WGBS methylation values currently. To test the capability of WGBS sample data to be integrated with array-based classifiers, two pre-existing medulloblastoma classifiers were used for which software access was unrestricted. Both classifiers are based upon the support vector machine algorithm and utilise a 10,000-probe reference set to predict both the four molecular subgroups and a previously defined seven-group classification defined by Schwalbe *et al.*, which splits SHH into infant and child/adult subtypes and both groups 3 and 4 in to low/high-risk subtypes respectively (Schwalbe *et al.*, 2017).

Application of NMB WGBS sample data (10x) to both classifiers revealed high concordance with that of matching array samples, with a mean WGBS sample probabilities of 0.967 (SD = 0.057) and 0.904 (SD = 0.133) returned by both 4 and 7 subgroup classifiers respectively (Tables 4.5 & 4.6). For both classifiers, all 10,000 features (probes) within the training set had a subsequent matched CpG within our NMB WGBS cohort, since all missing data was imputed. Matched array data returned mean class assignment probability scores of 0.949 (SD = 0.104) for the 4-group classifier, significantly lower than WGBS sample data (paired t-test,  $p = 0.031$ ). For the seven-subgroup classifier, matched array samples were assigned with a mean probability of 0.913 (SD = 0.113), with no significant difference between platforms ( $p = 0.341$ ). For the 35 matched samples tested, all were concordant for our 4-group classifier, including the single FFPE/fresh-frozen pair. This sample, a WNT tumour, was classified with a probability of 0.886 and 0.922 for FFPE and fresh frozen tissue respectively, highlighting the well-known difference in quality between both DNA sources. For the Schwalbe *et al.* classifier, 33 out of 35 samples were concordant between platforms. Both samples, NMB\_549 and NMB\_181, belonged to the SHH subtype and exhibited significant falls in assignment probability. All the results were obtained using WGBS data for which a 3x coverage filter during pre-processing. When applying sample data processed with a standard filter of 5x, no significant change in class assignment probability for the 4-group classifier was observed ( $p = 0.905$ ), but a significant fall in probability when predicting the Schwalbe *et al.*, subtypes ( $p = 0.005$ ) (see appendix – 8.1.4). A similar difference was

observed when testing the effect of coverage on mean class assignment probability, with no significant changes in probability observed by the 4-group classifier but significant falls with each drop in coverage for the 7-group classifier. ICGC sample data performed equally as well when applying matched, high depth WGBS sample data to both classifiers (Tables 4.5 & 4.6). Mean assignment probabilities of 0.970 (SD = 0.057) and 0.873 (SD = 0.150) were obtained for both 4 and 7-group classifiers respectively. No significant difference was observed between ICGC and NMB cohorts for either classifier, indicating that samples sequenced at 10x perform adequately (t-test,  $p > 0.05$ ). Compared to matched array samples, no significant difference in assignment probability was observed, with mean array assignment probabilities of 0.964 (SD = 0.071) and 0.870 (SD = 0.150) obtained for each classifier. Like that of the NMB cohort, complete concordance for the 4-group classifier was observed, and 3 misclassifications when classifying 7-groups. All 3 samples were also classified with low probability ( $< 0.6$ ) when applying array sample data, making it likely that these samples are inherently difficult to classify using the defined features that form part of this classifiers training set.



Coverage	WGBS - ICGC				WGBS - NMB		Array - ICGC		Array - NMB	
	Filter Applied	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	
30x		10,000	<b>0.970 (0.057)</b>	NA	NA					
10x	5x	10,000	<b>0.968 (0.060)</b>	10,000	<b>0.967 (0.055)</b>					
	3x	10,000	<b>0.970 (0.055)</b>	10,000	<b>0.968 (0.057)</b>					
9x	5x	10,000	<b>0.967 (0.061)</b>	10,000	<b>0.964 (0.069)</b>					
	3x	10,000	<b>0.970 (0.055)</b>	10,000	<b>0.968 (0.055)</b>					
8x	5x	10,000	<b>0.964 (0.067)</b>	10,000	<b>0.958 (0.084)</b>					
	3x	10,000	<b>0.970 (0.055)</b>	10,000	<b>0.969 (0.049)</b>					
7x	5x	10,000	<b>0.960 (0.078)</b>	10,000	<b>0.956 (0.069)</b>					
	3x	10,000	<b>0.969 (0.057)</b>	10,000	<b>0.969 (0.076)</b>	9,643	<b>0.964 (0.071)</b>	9,643	<b>0.949 (0.104)</b>	
6x	5x	10,000	<b>0.950 (0.100)</b>	10,000	<b>0.940 (0.101)</b>					
	3x	10,000	<b>0.968 (0.060)</b>	10,000	<b>0.968 (0.054)</b>					
5x	3x	10,000	<b>0.965 (0.065)</b>	10,000	<b>0.965 (0.068)</b>					
	1x	10,000	<b>0.971 (0.052)</b>	10,000	<b>0.963 (0.076)</b>					
4x	3x	10,000	<b>0.957 (0.083)</b>	10,000	<b>0.960 (0.072)</b>					
	1x	10,000	<b>0.971 (0.051)</b>	10,000	<b>0.965 (0.070)</b>					
3x	1x	10,000	<b>0.970 (0.053)</b>	10,000	<b>0.965 (0.065)</b>					
2x	1x	10,000	<b>0.967 (0.058)</b>	10,000	<b>0.967 (0.052)</b>					
1x	1x	10,000	<b>0.948 (0.093)</b>	10,000	<b>0.953 (0.086)</b>					

**Table 4.5** – Classifier performance (4-group) for combined WGBS and Array cohorts. Each cohort is split into ICGC (n = 34) and NMB (n = 36) respectively, as each dataset was originally sequenced at different levels of depth.

Coverage	WGBS - ICGC				WGBS - NMB		Array - ICGC		Array - NMB	
	Filter Applied	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	
30x		10,000	<b>0.873 (0.150)</b>	NA	NA					
10x	5x	10,000	<b>0.872 (0.150)</b>	10,000	<b>0.873 (0.144)</b>					
	3x	10,000	<b>0.874 (0.149)</b>	10,000	<b>0.905 (0.131)</b>					
9x	5x	10,000	<b>0.871 (0.149)</b>	10,000	<b>0.862 (0.147)</b>					
	3x	10,000	<b>0.874 (0.148)</b>	10,000	<b>0.903 (0.131)</b>					
8x	5x	10,000	<b>0.870 (0.148)</b>	10,000	<b>0.850 (0.144)</b>					
	3x	10,000	<b>0.873 (0.147)</b>	10,000	<b>0.898 (0.136)</b>					
7x	5x	10,000	<b>0.873 (0.137)</b>	10,000	<b>0.818 (0.154)</b>					
	3x	10,000	<b>0.872 (0.148)</b>	10,000	<b>0.890 (0.141)</b>	9,651	<b>0.870 (0.150)</b>	9,651	<b>0.913 (0.115)</b>	
6x	5x	10,000	<b>0.871 (0.136)</b>	10,000	<b>0.763 (0.178)</b>					
	3x	10,000	<b>0.871 (0.148)</b>	10,000	<b>0.880 (0.148)</b>					
5x	3x	10,000	<b>0.870 (0.148)</b>	10,000	<b>0.862 (0.160)</b>					
	1x	10,000	<b>0.872 (0.151)</b>	10,000	<b>0.914 (0.127)</b>					
4x	3x	10,000	<b>0.870 (0.142)</b>	10,000	<b>0.841 (0.154)</b>					
	1x	10,000	<b>0.870 (0.154)</b>	10,000	<b>0.912 (0.129)</b>					
3x	1x	10,000	<b>0.870 (0.152)</b>	10,000	<b>0.908 (0.131)</b>					
2x	1x	10,000	<b>0.868 (0.149)</b>	10,000	<b>0.894 (0.137)</b>					
1x	1x	10,000	<b>0.856 (0.159)</b>	10,000	<b>0.847 (0.164)</b>					

**Table 4.6** – Classifier performance (7-group) for both WGBS and array cohorts.

The results observed in this project demonstrate that medulloblastoma tumour samples sequenced via WGBS can seamlessly be integrated into pre-existing classifier models and predict molecular subgroup with excellent precision. Concordance between the platform and matched DNA methylation microarray samples is very high, with 100% concordance observed when classifying the four molecular subgroups of medulloblastoma and over 90% concordance when applying the Schwalbe *et al.*, (2017) classification model. Both models are trained solely using small numbers (i.e., <500) of array samples, and this success indicates that the current optimised DFKZ brain tumour classifier, which was derived initially from 2801 tumour samples, would in a similar way likely be capable of classifying WGBS sample data with high accuracy. This could be achieved by making their software publicly available, allowing researchers to apply WGBS sample data at their own discretion following processing, or by enabling sample data to be uploaded in the form of a beta value matrix. Whilst this seems simple, it presents some risk as uploaded data is not subject to standardised pre-processing (e.g., no imputation) and may result in incorrect classification calls. Indeed, whilst data was processed effectively and demonstrates that low-pass WGBS MB samples classify with high probability using our methods, it may be more pertinent to further adapt array-based classifier training sets to deal with WGBS sample data for testing. For example, studies testing the capability of RRBS and nanopore sequencing for paediatric brain tumour classification leverage array training sets by binarizing beta values to 1 and 0 to mimic the numerical binomial distribution of methylation sequencing data (van Paemel *et al.*, 2021; Kuschel *et al.*, 2021). However, such an approach disregards the potential significance of hemi-methylation, which has been shown to be an important biomarker in tumorigenesis (Sun *et al.*, 2021; Shao *et al.*, 2009). In addition, WGBS sample data has been shown to be highly correlated to array beta value calculations suggesting that such a binarisation step may not be necessary. Both studies also do not impute data, drastically reducing the number CpGs available for training due to random missingness causing a fall in the number of common CpGs shared as cohort size increases. Whilst this does not significantly hamper prediction capability too much in these studies, data missingness would undoubtedly become a limiting factor as cohort sizes increase. Regardless, to confirm these findings, further research is needed to comprehensively evaluate how WGBS sample data classifies with array-trained models with increased cohort sizes extended to include additional tumour types.

Whilst performance is high in this study, the cohort size is modest and the NMB cohort was carefully selected to include highly classifiable samples. For the small number of difficult to classify samples, this was also typically the case for matched array samples, suggesting that the features selected by DNA methylation microarray may not be suitable for these cases. This potentially could be resolved when considering the increased CpG variability observed outside of the microarray probe manifest, which contributed to increased visual separation between subgroups. This increased variability also offers potential for optimisation of subgroup definition based on whole genome methylome profiles. This may be of particular benefit when dealing with the additional subtypes that are continuing to be identified in medulloblastoma (e.g., subtypes I-VIII, SHH subtypes), where the whole methylome may

enable difficult to classify samples to be classified more confidently. Further research is required in this area regarding WGBS and MB subgroup classification. Whether there is currently an appetite for this is questionable, given the clear benefits of the methylation array platform and continuing development with the platform. Such research would not be able to be conducted until WGBS sequencing is routinely employed by other clinical institutions. Until such a stage is reached, it makes sense to continue to take advantage of the pre-existing array sample data that is available, and the integration of WGBS subgroup classification into pre-existing array classifier models has been shown to be a relatively easy quality of life addition in this study. Such an integration is an important, necessary step if the field is to take advantage of methylation sequencing technologies, ensuring a smooth transition to such platforms in the coming years.

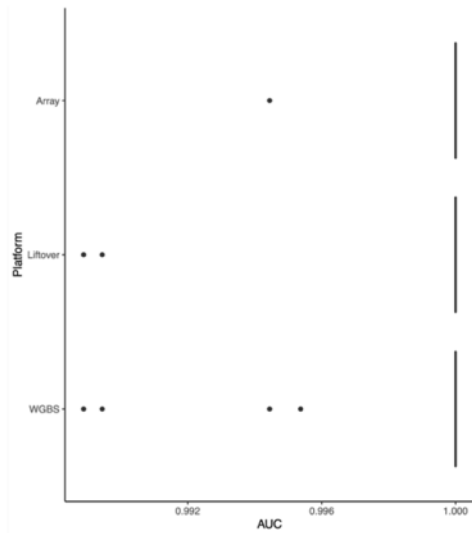
## 4.4.5 – Validated classifier models trained using low-pass WGBS perform equally to array trained models

Given the increased variability found among CpGs not covered by the array platform, it was deemed prudent to survey how classifier models trained to predict the molecular subgroups of medulloblastoma perform when trained solely using WGBS samples rather than samples analysed via DNA methylation microarray. This allows WGBS-derived classifiers to take forward CpGs of increased variability for training, which may enable the creation of models that can predict molecular subgroup with greater accuracy and higher probability than array-based alternatives. To investigate this, seven suitable classification algorithms were applied; capable of assigning multiple classes alongside an associated class probability measure; elastic net, random forest, XGBoost (linear/tree boost algorithms), SVM (linear/radial kernel) and logic regression. Each model was trained using the top 10,000 most variable CpGs/probes for our matched WGBS and Array cohorts (n=77) and validated performance using a 10 times 6-fold cross validation, allowing a direct statistical comparison in performance. For each algorithm, validation performance for all models is presented and a comparison between WGBS and Array trained models is drawn. An overview is given of the effect of downsampling, and coverage filter applied to WGBS data provided. A summary of mean validation model performance is provided in the appendix (8.1.9).

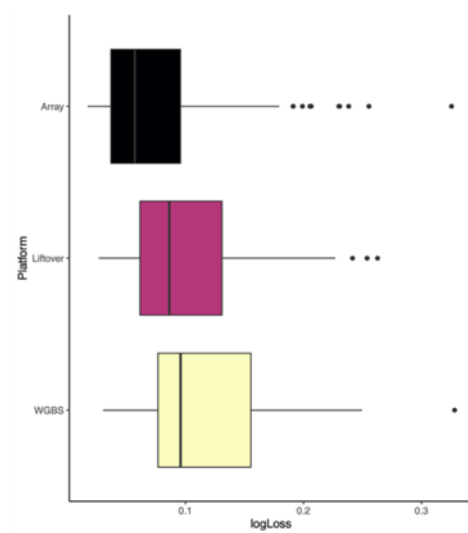
### 4.4.5.1 – Elastic net

Validated performance of elastic net classification models was excellent across both platforms. When training with all samples (n=77), cross-validated WGBS models returned a mean accuracy of 0.967, adjusted Cohen's Kappa statistic (CK) of 0.956, AUC of 0.999 and log loss of 0.117 (Figure 4.13). When considering the effect of lowering coverage, model performance remained stable, with a fall in log loss observed at 2x (Figure 4.14). The final values selected for the WGBS trained model was alpha = 0.4 and lambda = 0.10, whereas for our array model alpha = 0.4 and lambda = 0.010 respectively. No statistical difference in model performance was observed when training models with WGBS sample data filtered using either threshold of 3x or 5x. Array trained models performed similarly, with statistically equal accuracy (0.985, p = 0.060), CK (1, p = 0.054) and AUC (1.00, p = 1) values observed. Mean log loss, a measure of the divergence in class assignment probability compared to the true class probability (closer to 0 indicates better performance), was lower, yielding a mean of 0.084 (p < 0.001). The high performance observed using the elastic net algorithm is consistent with other studies involving medulloblastoma datasets (Weishaupt *et al.*, 2019) and methylation data in general (Zhuang *et al.*, 2012), where its suitability is well established.

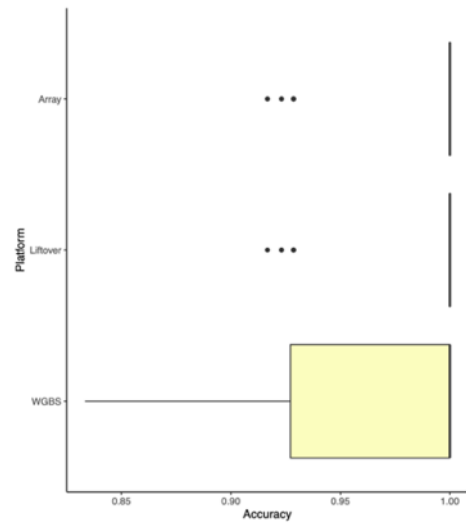
**A) AUC**



**B) Log Loss**



**C) Accuracy**

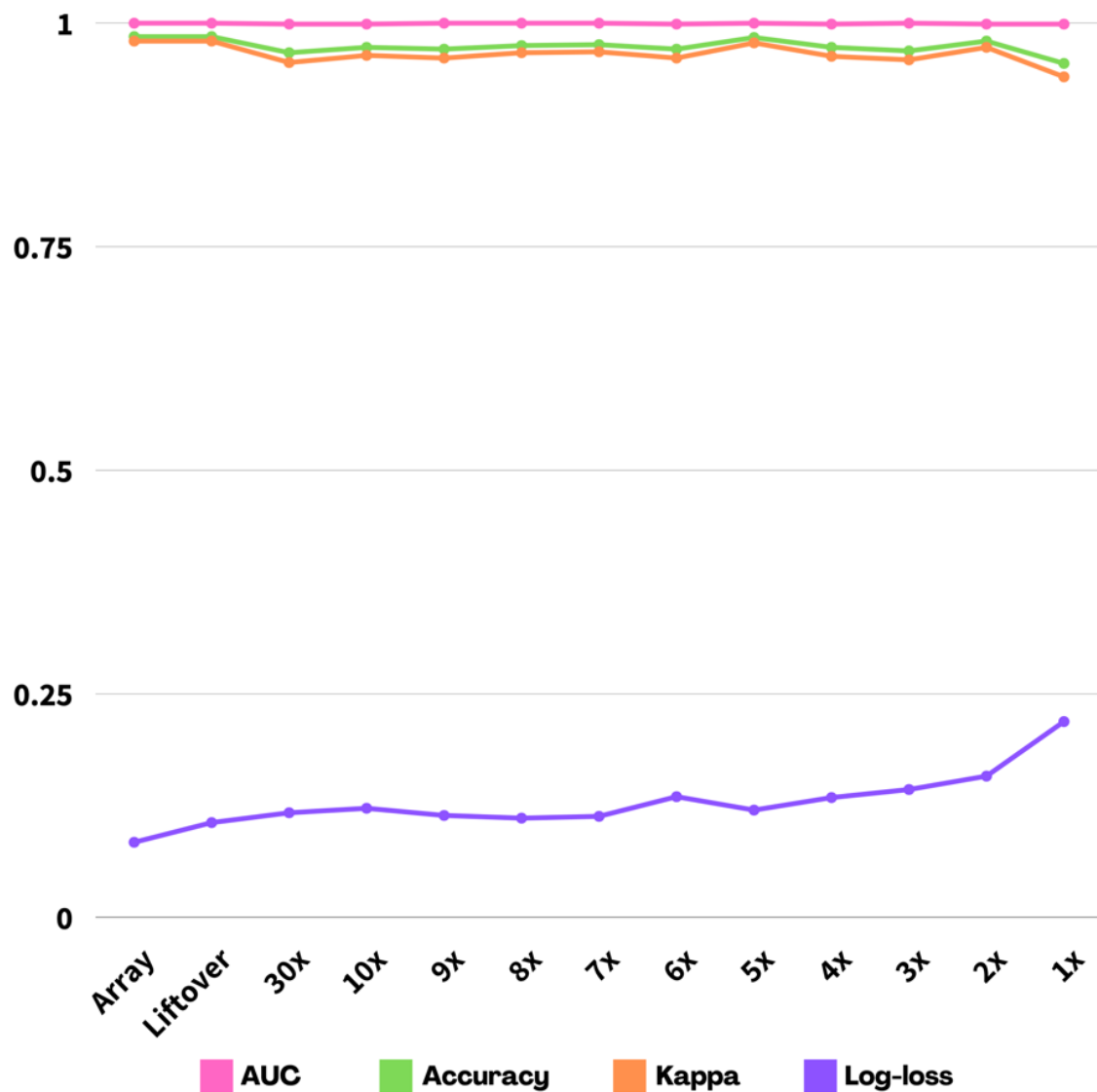


**D) Cohen's Kappa**



**Figure 4.13.** Mean elastic net classifier validation performance for Array, LiftOver and WGBS derived training sets (n = 10,000 features).

**Legend:** The black dot denotes the median value across all validation tests, either side of the box denote the 25th and 75th percentiles, and the whiskers correspond to the inter quartile range. Dots outside this range are considered outliers.

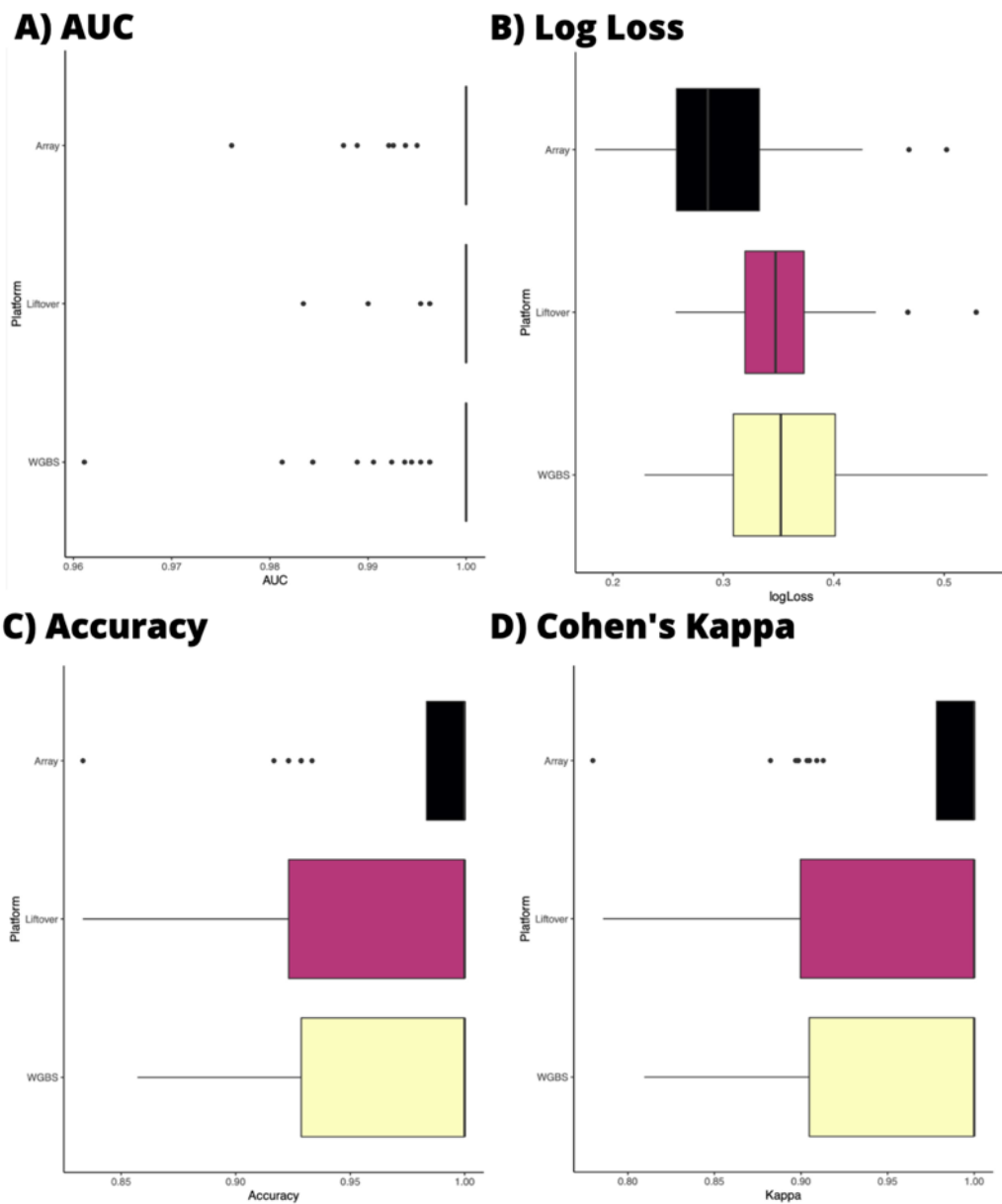


**Figure 4.14.** Mean elastic net classifier performance for array, liftover and each downsampled level of coverage.

**Legend:** Classification models with good performance will yield AUC, accuracy and CK statistics close to 1, with 1 being a perfect score. Log-loss values are a measure of how close all assigned prediction probabilities are to the true class (e.g., Molecular subgroup). Poorly trained classifiers exhibit lower class assignment probabilities and will display higher log-loss values, with 0 being a perfect score.

### 4.4.5.2 – Random forest

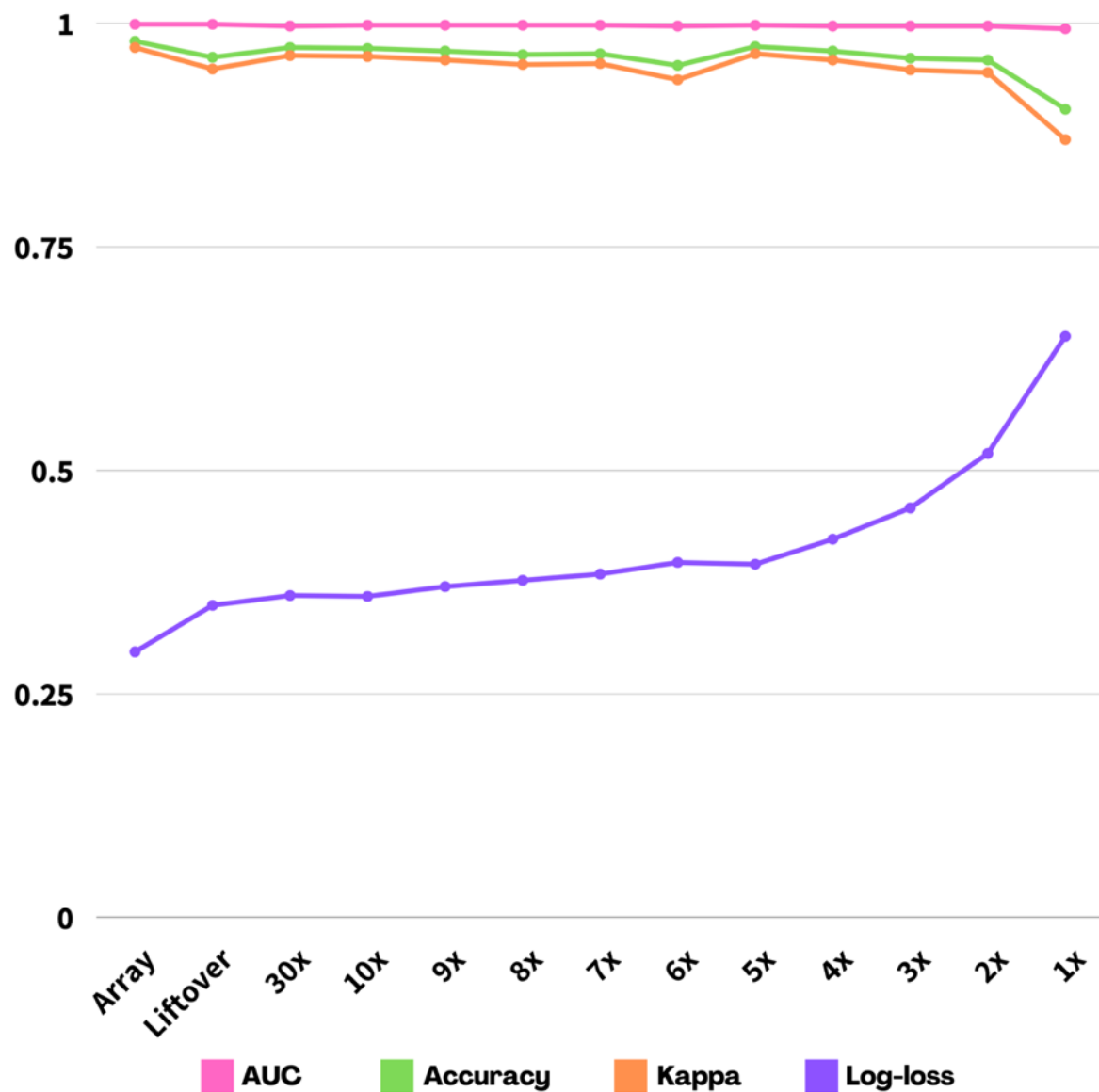
Generally, the random forest classifiers performed equivalently between platforms, with high mean accuracy (WGBS = 0.973, Array = 0.980), mean CK (WGBS = 0.964, Array = 0.973) and AUC (WGBS = 0.997, Array = 0.998) for both platforms respectively ( $p = 1.000$ ) (Figure 4.15). Log loss performance was statistically lower for the array trained model (WGBS = 0.360, Array = 0.297,  $p = < 0.001$ ), indicating greater assignment probabilities for each sample across all parameter combinations tested. For our WGBS cohort, trained models performed best using 1000 trees and  $mtry = 60$  whereas the optimal model trained using array data used 1250 trees and  $mtry = 9999$ . Like elastic net, a drop-off in performance was observed at  $\sim 5x$  (Figure 4.16). When comparing WGBS models trained using sample data pre-processed with a filter of 5x or 3x, the application of a lower filter resulted in significantly better performing models, with greater mean accuracy (High filter = 0.973, Low filter = 0.980), CK (High filter = 0.940, Low filter = 0.964) log loss (High filter = 0.386, Low filter = 0.360) observed. ( $p < 0.001$ ) (see appendix). Similar performance using random forest is encouraging, given the algorithm handles outliers well and is not as susceptible to overfitting.



**Figure 4.15.** Mean random forest classifier validation performance for Array, LiftOver and WGBS derived training sets (n = 10,000 features).

**Legend:** The black line denotes the median value across all validation tests, either side of the box denote the 25th and 75th percentiles, and the whiskers correspond to the inter quartile range. Dots outside this range are considered outliers.



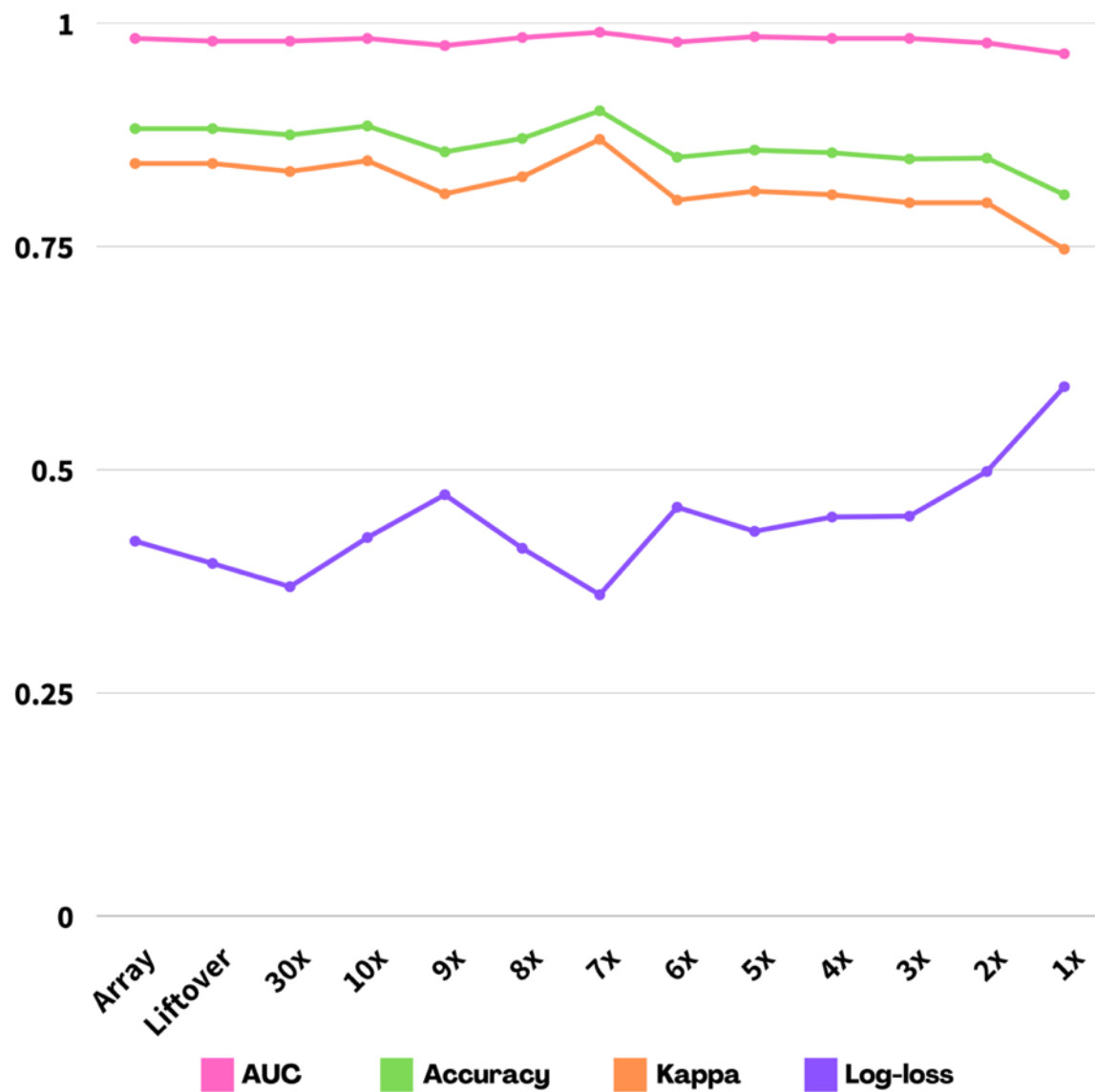


**Figure 4.16.** Mean random forest classifier performance for array, liftover and each downsampled level of coverage.

**Legend:** Classification models with good performance will yield AUC, accuracy and CK statistics close to 1, with 1 being a perfect score. Log-loss values are a measure of how close all assigned prediction probabilities are to the true class (e.g., Molecular subgroup). Poorly trained classifiers exhibit lower class assignment probabilities and will display higher log-loss values, with 0 being a perfect score.

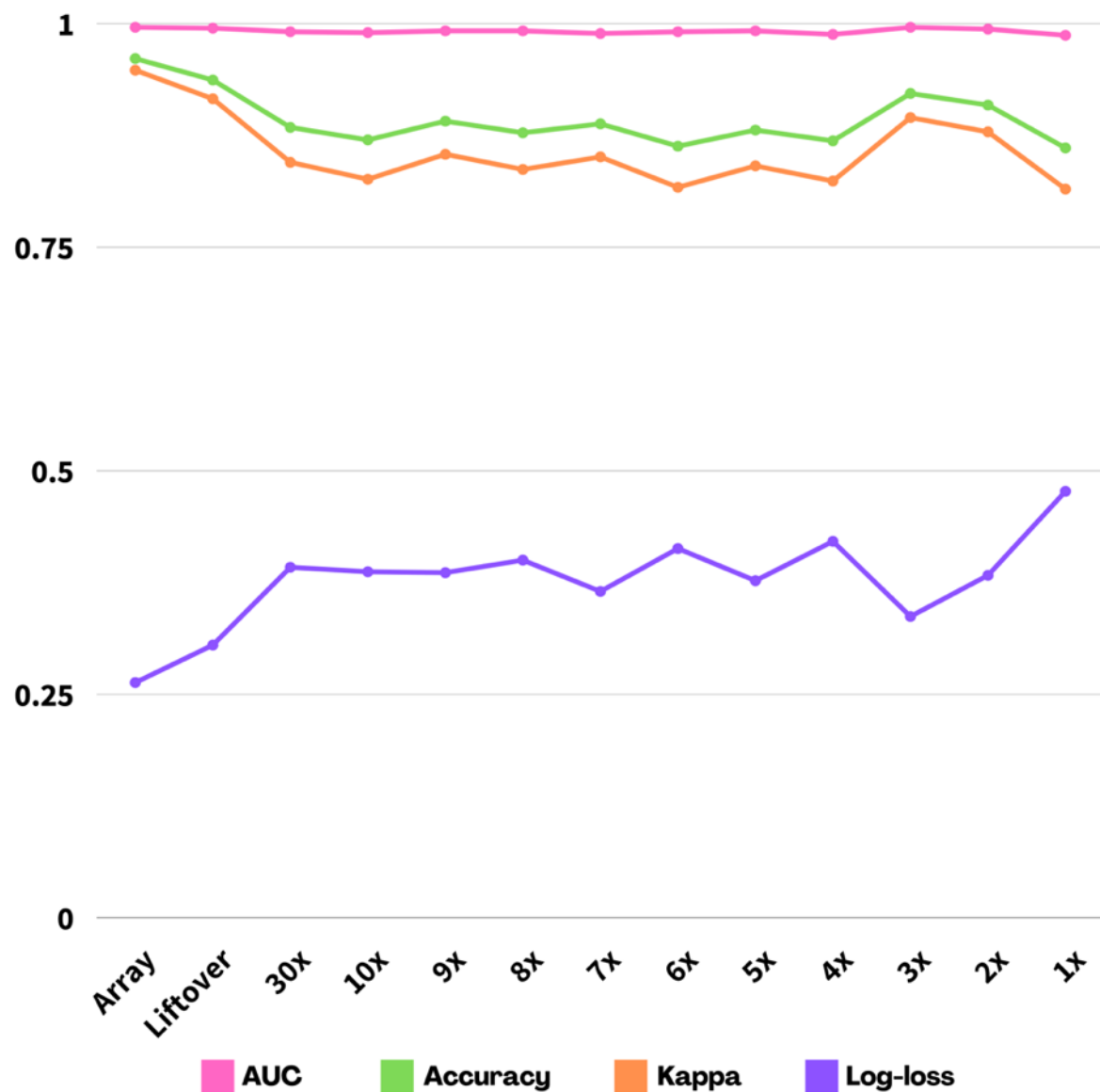
### 4.4.5.3 – Extreme gradient boosting

For our cohorts, linear boosted models performed relatively poorly compared to other algorithms in both WGBS and array sample trained models. No statistical difference was observed in mean model accuracy (WGBS = 0.875, array = 0.882), CK (WGBS = 0.834, array = 0.843) and AUC (WGBS = 0.980, array = 0.983) performance ( $p = 1.000$ ), whereas log loss was statistically lower in WGBS trained models (WGBS = 0.369, array = 0.420,  $p < 0.001$ ). Optimal parameters for our WGBS trained classifier was rounds = 100, lambda = 0.1, alpha = 0 and eta = 0.3. Parameters were identical for our array trained model apart from lambda which was set at 0. Like other models, the usage of WGBS data filtered with a 3x filter resulted in greater model performance, with greater mean accuracy (high filter = 0.818, low filter = 0.875), CK (high filter = 0.759, low filter = 0.834), AUC (high filter = 0.965, low filter = 0.980) and lower log loss values (high filter = 0.508, low filter = 0.369) observed ( $p < 0.02$ ) (see appendix). Tree boosted models performed marginally better than linear boosted models in both platforms. All performance measures were superior in both platforms, with higher mean accuracy (WGBS = 0.906, array = 0.961,  $p < 0.001$ ), CK (WGBS = 0.873, array = 0.948,  $p < 0.001$ ), AUC (WGBS = 0.991, array = 0.999,  $p = 1.000$ ) and lower log loss (WGBS = 0.323, array = 0.263,  $p = 0.029$ ) values observed. Mean log loss was also significantly lower when applying a 3x filter during WGBS processing (high filter = 0.392, low filter = 0.323,  $p < 0.001$ ). Both models exhibited unstable performance as coverage fell, with 1x performing most poorly (Figures 4.17 & 4.18).



**Figure 4.17.** Mean XGBoost (Linear booster) classifier validation performance for Array, LiftOver and WGBS derived. Training sets (n = 10,000 features).

**Legend:** Classification models with good performance will yield AUC, accuracy and CK statistics close to 1, with 1 being a perfect score. Log-loss values are a measure of how close all assigned prediction probabilities are to the true class (e.g., Molecular subgroup). Poorly trained classifiers exhibit lower class assignment probabilities and will display higher log-loss values, with 0 being a perfect score.

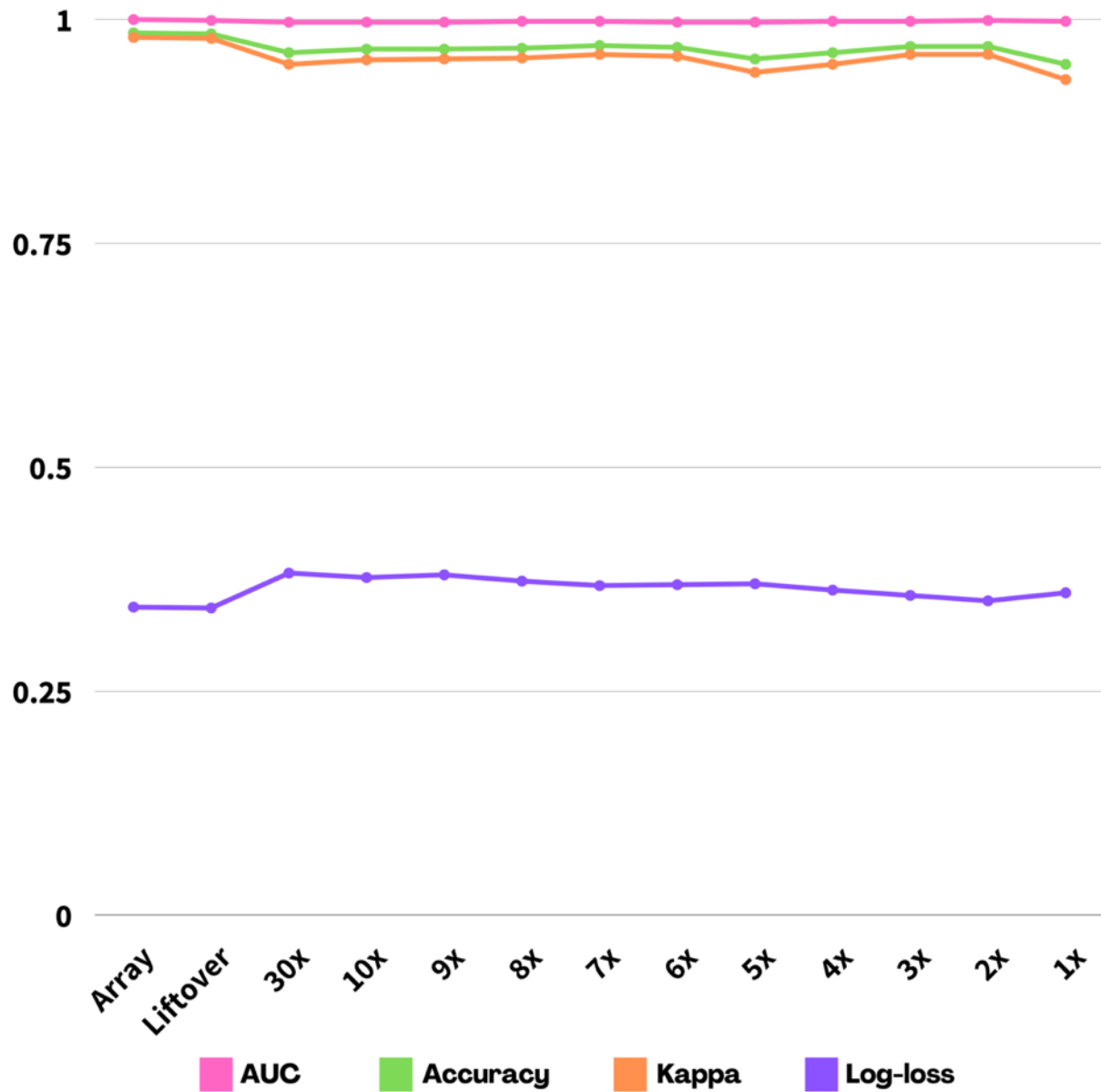


**Figure 4.18.** Mean XGBoost (Tree booster) classifier validation performance for Array, LiftOver and WGBS derived. Training sets (n = 10,000 features).

**Legend:** Classification models with good performance will yield AUC, accuracy and CK statistics close to 1, with 1 being a perfect score. Log-loss values are a measure of how close all assigned prediction probabilities are to the true class (e.g., Molecular subgroup). Poorly trained classifiers exhibit lower class assignment probabilities and will display higher log-loss values, with 0 being a perfect score.

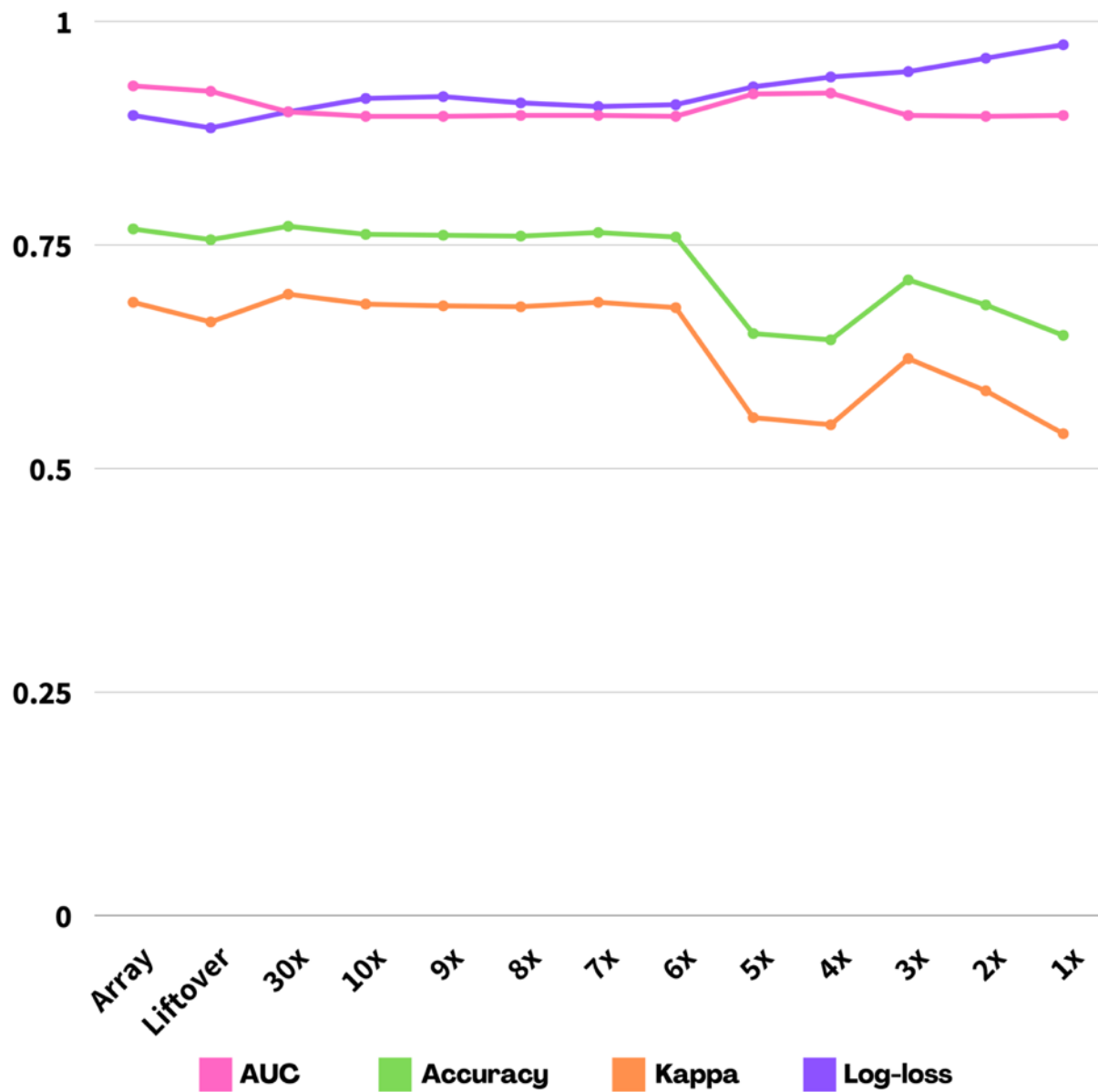
#### 4.4.5.4 – Support vector machine

We trained SVM classifiers using both linear and radial kernels (Figures 4.19 & 4.20). SVM performance using linear kernels was excellent for both platforms, with high mean accuracy (WGBS = 0.963, array = 0.985,  $p = 0.015$ ), CK (WGBS = 0.950, array = 0.980,  $p = 0.005$ ), AUC (WGBS = 0.997, array = 1.000,  $p = 0.085$ ) and low log loss (WGBS = 0.382, array = 0.344,  $p = < 0.001$ ) values reported (Table 4.7). For WGBS samples, our optimal classification model was obtained using a cost value of 128, whereas cost was optimal at 32 for our array model. Accuracy, CK and AUC performance measures remained stable in WGBS models trained using samples processed with a 3 or 5x filter, but log loss was significantly lower in our low filter cohort (high filter = 0.392, low filter = 0.382,  $p = < 0.001$ ). Performance was significantly worse when training SVM classifiers using a radial basis function kernel for both platforms. Accuracy (WGBS = 0.771, array = 0.768,  $p = 1.00$ ), CK (WGBS = 0.695, array = 0.686,  $p = 1.000$ ) and AUC (WGBS = 0.899, array = 0.928,  $p = < 0.001$ ) was low, whereas mean log loss was the highest out of all algorithms tested (WGBS = 0.899, array = 0.895,  $p = 1.000$ ), indicating that samples were predicted with poor probability as well as low accuracy. SVM classifiers are typically better suited to binary classification cases and the chosen kernel is recognised to have a significant impact on performance, which is evidently the case in this project.



**Figure 4.19.** Mean SVM (Linear kernel) classifier validation performance for Array, LiftOver and WGBS derived. Training sets (n = 10,000 features).

**Legend:** Classification models with good performance will yield AUC, accuracy and CK statistics close to 1, with 1 being a perfect score. Log-loss values are a measure of how close all assigned prediction probabilities are to the true class (e.g., Molecular subgroup). Poorly trained classifiers exhibit lower class assignment probabilities and will display higher log-loss values, with 0 being a perfect score.



**Figure 4.20.** Mean XGBoost (Radial kernel) classifier validation performance for Array, LiftOver and WGBS derived. Training sets (n = 10,000 features).

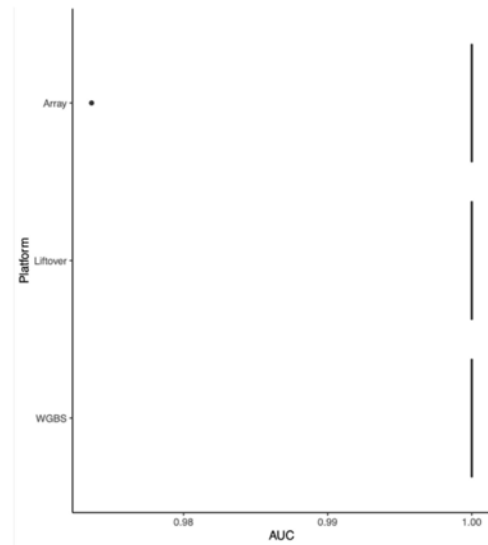
**Legend:** Classification models with good performance will yield AUC, accuracy and CK statistics close to 1, with 1 being a perfect score. Log-loss values are a measure of how close all assigned prediction probabilities are to the true class (e.g., Molecular subgroup). Poorly trained classifiers exhibit lower class assignment probabilities and will display higher log-loss values, with 0 being a perfect score.

### 4.4.5.5 – Logic regression

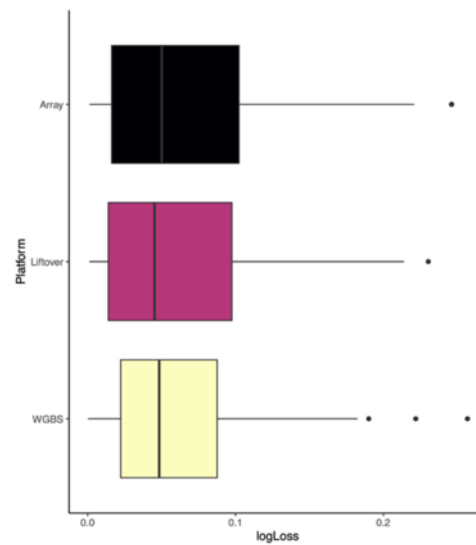
Interestingly, validated logistic regression classifiers performed well in this study for both platforms. Performance was excellent, with high mean accuracy (WGBS = 0.977, array = 0.969,  $p = 0.178$ ), CK (WGBS = 0.969, array = 0.959,  $p = 0.173$ ), AUC (WGBS = 1.000, array = 1.000,  $p = 1.000$ ) and low log loss (WGBS = 0.068, array = 0.070,  $p = 0.009$ ) observed (Figures 4.21). No fall in performance was observed in models trained using downsampled datasets (Figure 4.22), which is possibly explained by the increased proportion of binary values in WGBS data as sequencing depth falls. Classifiers trained using WGBS sample data filtered using a 5x threshold produced models that performed with significantly lower mean log loss values (high filter = 0.043, low filter = 0.068,  $p = < 0.001$ ).



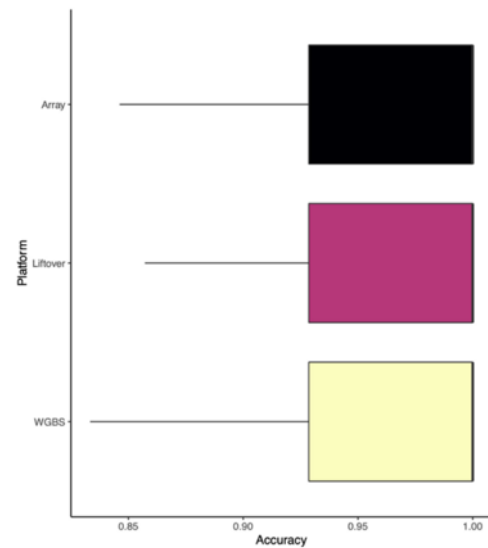
**A) AUC**



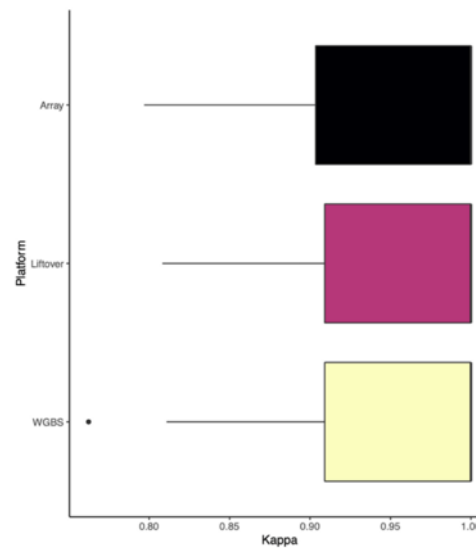
**B) Log Loss**



**C) Accuracy**

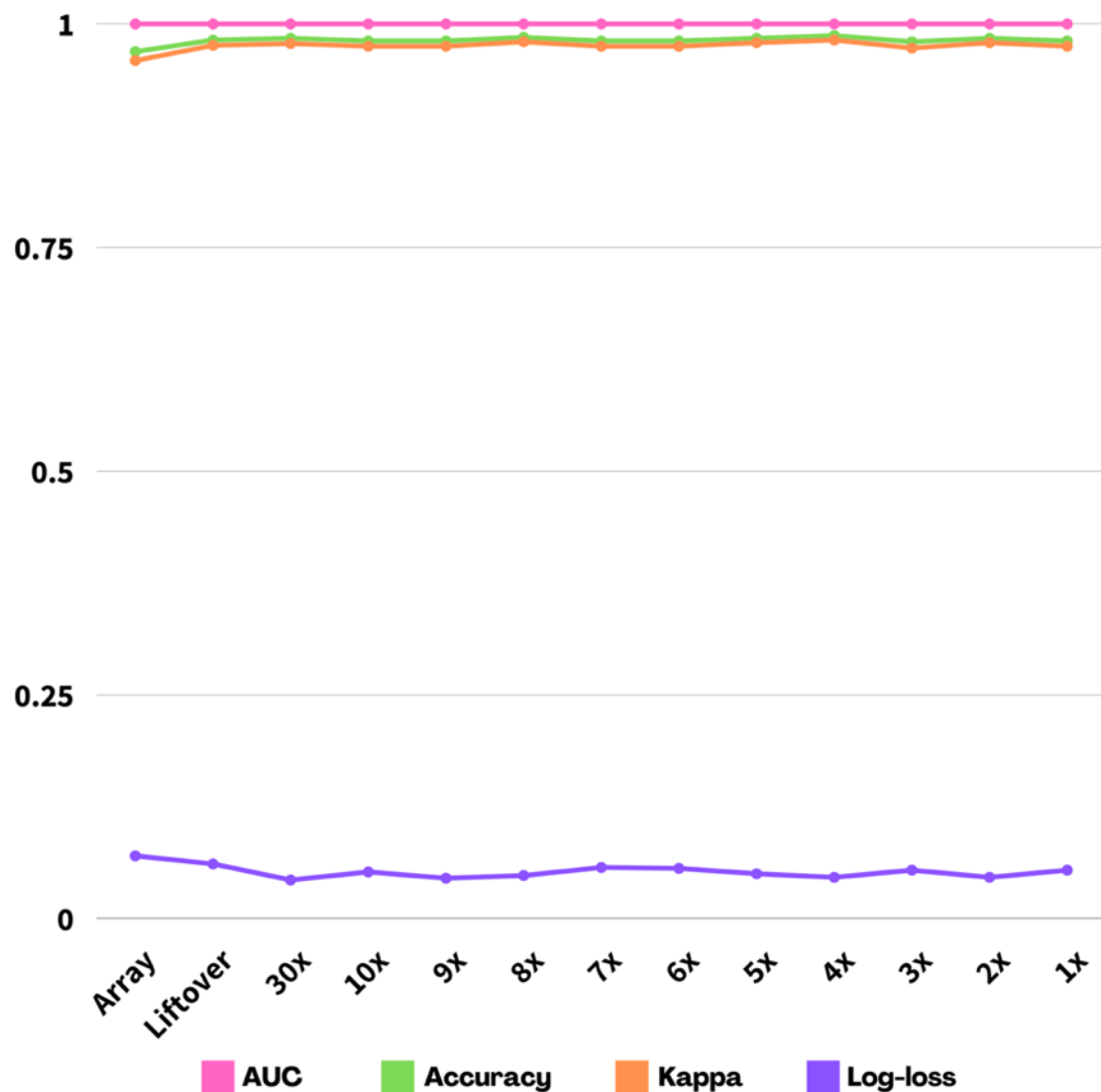


**D) Cohen's Kappa**



**Figure 4.21.** Mean logic regression classifier validation performance for Array, LiftOver and WGBS derived training sets (n = 10,000 features).

**Legend:** The black line denotes the median value across all validation tests, either side of the box denote the 25th and 75th percentiles, and the whiskers correspond to the inter quartile range. Dots outside this range are considered outliers.



**Figure 4.22.** Mean logic regression classifier validation performance for Array, LiftOver and WGBs derived. Training sets (n = 10,000 features).

**Legend:** Classification models with good performance will yield AUC, accuracy and CK statistics close to 1, with 1 being a perfect score. Log-loss values are a measure of how close all assigned prediction probabilities are to the true class (e.g., Molecular subgroup). Poorly trained classifiers exhibit lower class assignment probabilities and will display higher log-loss values, with 0 being a perfect score.

---

#### 4.4.5.6 – Model limitations

While our results show concordance between both WGBS and DNA methylation microarray trained classification models in predicting the molecular subgroups of medulloblastoma, they are subject to some limitations. Due to the modest number of samples in this study, no independent holdout data set was created, preventing an unbiased evaluation of all optimised final models. To address this, we ensured that all models were assessed via k-fold repeated cross-validation, where model performance is assessed numerous times on different isolated partitions of the original training set. Whilst this method is established as a useful way of minimising the risk of model overfitting, the lack of an independent test set ultimately prevents us from assessing whether the model is truly applicable to unseen, incoming samples. In fact, the small cohort size used for training (fewer than 100 samples) is another limitation of our analyses. Ideally, training sets should be as large as possible, particularly when concerned with the classification of multiple heterogeneous classes such as the molecular subgroups of medulloblastoma (Capper *et al.*, 2018).

To ensure comparability, we utilised the same method of feature selection – unsupervised variance filtering of all beta values to the top 10,000 CpGs as defined by a standard deviation measure. This feature selection method has been used multiple times when dealing with methylation data and is shown to capture the variability between the subgroups of medulloblastoma effectively in other studies (Sharma *et al.*, 2019; Maros *et al.*, 2020). However, many of the algorithms tested may be more suited to embedded methods such as the feature importance measure of random forest (Capper *et al.*, 2018) or wrapper methods like recursive feature elimination with the SVM algorithm (Guyon *et al.*, 2002). In fact, the application of unsupervised variance filtering may be quite unsuitable, and comprehensive analysis into the appropriate feature selection algorithm should be performed when focusing on a particular algorithm (Maros *et al.*, 2020). However, for this study we deemed our feature selection method as appropriate, given our intention to investigate the capability of WGBS data to produce comparable performance to that of matched array sample data and not to generate a fully optimised classification model.

Finally, the usage of class labels derived from array data may influence the performance of classifiers trained with samples sequenced by WGBS. Given the increased CpG variability and separation observed between classes when accounting for CpGs missed by the DNA methylation microarray, particularly between that of group 3 and 4 (this resulted in some samples currently identified as group 3 clustering clearly amongst group 4 samples), it is possible that class labels may differ slightly if they were identified using these feature sets. Indeed, the increased CpG variability observed in this study puts forward a strong case for subgroup optimisation using WGBS. Many samples remain difficult to classify via methylation microarray classifiers, and with training sets for the DFKZ classifier reaching >50,000 samples (Perez & Capper, 2020), it is likely that further optimisation for the more common tumour entities such as the subgroups of medulloblastoma would be challenging, given the limited

number of features to select within the array platform. The inclusion of these additional variable features may lead to the identification of further optimised methylation profiles for each respective subgroup and potentially subtype of medulloblastoma. This may have clinical implications for patients, particularly for those with tumours whose subgroup cannot be predicted reliably. However, much of this is hypothetical and would need to be tested as part of a comprehensive study that attempts to redefine the subgroups of medulloblastoma using similar clustering approaches to studies that identified them using the DNA methylation microarray platform.

## 4.5 – Summary & conclusion

The analyses presented in this chapter successfully demonstrate the utility of WGBS derived methylation data as an effective and potentially superior substitute for DNA methylation array tumour data. Variance filtering of the entire methylome in a matched cohort of MB tumour samples revealed significantly greater numbers of CpGs that capture greater variability than that of probe sites assessed by the DNA methylation microarray platform. A total of 3,990,549 CpGs were identified as variable (defined using a standard deviation measure  $> 0.25$ ) compared to just 27,211 probe sites in the matched microarray cohort. Whilst technical differences between the platforms and low sequencing explain some of the increased variability observed, just 36% and 42% of the top 10,000 variable CpGs were identified as overlapping within 125bp of any respective probe site located on the DNA methylation microarray manifest. This indicates that a significant amount of CpG variability found in MB samples in this study are found in intergenic and repetitive regions not assessed by the array. Variability was found to be differentially distributed between both platforms, although it is likely this may be due to the design of the array platform not being proportionally equal to each respective chromosome size by design. Regardless, the clear substantial variable methylation identified provides motivation for the establishment of whole methylome profiles for the subgroups of MB.

Applying various dimensionality reduction techniques to the most variable features revealed that the molecular subgroups of medulloblastoma were visually recapitulated when using methylation in our WGBS MB cohorts. We demonstrate that WGBS sample data clusters among their respective subgroup irrespective of sequencing batch or depth, with near identical clustering patterns observed between WGBS and Array data. We observe potentially greater degrees of separation between groups 3 and 4, likely due to the increased variability inherent to the WGBS-derived feature set and the deliberate selection of highly classifiable samples in our cohort. Two samples, a group 3 with intermediate probability and a low probability group 4 classified as MBNOS cluster amongst group 4 samples to a much greater degree when utilising the WGBS derived feature set. This is nonetheless an encouraging finding, given the two share highly similar methylation profiles when analysed using DNA methylation microarray. Clustering WGBS derived samples to the original DFKZ paediatric brain tumour classifier reference MB cohort corroborated our findings, demonstrating that WGBS MB

sample methylation values cluster among their respective subgroup among non-matched samples analysed from various institutions.

The observed inter-operability between WGBS and Array derived methylation values when applying visualisation techniques translated further when testing the capability of array trained classifier platforms to predict the molecular subgroup of MB tumours sequenced using WGBS. Applying WGBS sample data to historical MB subgroup classifiers (for which we had full development access to) predicted subgroups with high concordance to matched array samples. Mean classification probability was high for both 4 and 7 subgroup classifiers, with mean class prediction probabilities of 0.967 (SD = 0.055) and 0.873 (SD = 0.144) respectively. Matched array data was classified with lower probabilities when considering the 4-group classifier (mean probability = 0.949, SD = 0.104,  $p = 0.031$ ), and equally for the 7-group classifier respectively (mean probability = 0.913, SD = 0.113,  $p = 0.341$ ).

These findings demonstrate that tumour samples sequenced via WGBS can seamlessly be integrated into pre-existing classification models with high precision. This can be achieved if a classifier model and its associated training set is publicly available, as beta values from either sequencing or array sources are compatible. Whilst this is a simple integration for a bioinformatician with intermediate experience, it is subject to risk as no sample data that has not been processed effectively may result in incorrect classification calls. The successful integration of WGBS into pre-existing Array classifiers is an important and beneficial finding for researchers whose work involves paediatric brain tumour classification and is an important step if the field is to ensure a smooth transition from the array platform to methylation sequencing in the coming years. The development of in-house classifier models trained using WGBS sample data to predict the molecular subgroups of medulloblastoma performed excellently for nearly all algorithms tested, performing similarly to models trained using array data. This high concordance indicates that native WGBS trained classifiers will likely perform equally to that of current array trained models once sufficient WGBS sample data has accumulated. Our analyses were not without limitations however, namely due to small cohort size and lack of specialised feature selection. Regardless, all our findings suggest that there is significant variability located in regions not assessed by the DNA methylation microarray platform. This variability translates into increased visual separation when applying dimensionality reduction techniques and concordant classification performance when using array-defined subgroup class labels. It is therefore not an unlikely hypothesis that there is room to further define the methylation profile for each potential subgroup and subtype of medulloblastoma by using the whole methylome, and further study should be encouraged in this area using larger cohorts of samples.

## **Chapter 5 – Generating diagnostic readouts for medulloblastoma via low-depth whole genome bisulfite sequencing**

---

## 5.1 – Introduction

The identification of variable molecular features in MB is vital to current risk stratifications (Section 1.7.7) achieved by using a variety of different platforms in a research setting and is largely dependent on resources and the overarching goals of any given study. Aneuploidy and/or gene copy number measurements can be inferred using a variety of methods, with both microarray and karyotype methodologies predominating for many years. Alongside its application as the current gold-standard for brain tumour/subgroup identification (Capper *et al.*, 2018), DNA methylation microarrays have been widely used to infer copy number variation (CNV) as an additional readout, given its capability to detect large structural variation and the advantage of deriving a copy number profile alongside a methylation subgroup call at no additional cost (Cho *et al.*, 2019). Indeed, a comprehensive CNV profile for each submitted sample is returned as part of the DFKZ paediatric brain tumour classifier analytical pipeline using the ‘conumee’ package, which pools together methylated and unmethylated signal intensities and calculates a ratio against a reference set of diploid samples that have a balanced genome (Hovestadt & Zapatka, 2017). CNV analysis of MB tumours samples analysed via DNA methylation microarray has enabled broad whole and single-arm chromosomal changes to be called with high accuracy. Its utility extends further to larger focal gains/losses where probe sites are numerous. However, the platform is inherently limited by design, with its gene-centric focus limiting the calling of variation in intergenic regions and uneven probe coverage preventing focal changes to be reliably identified where probe density is low. Regardless, the capability to infer CNV from the methylation microarray platform remains incredibly useful, and the integration of genomic and epigenomic analyses presents significant economic benefits to researchers.

DNA methylation microarrays can also be employed for differential methylation analysis. Briefly, differential methylation analysis involves the detection of differentially methylated probes (DMPs) and/or regions (DMRs) between sample groups (e.g., subgroup-specific regions of highly increased/decreased DNA methylation) and is of significant interest in clinical research when concerned with identifying subgroup-specific biomarkers. DMRs are known to occur throughout the genome, and the array is often employed as a cost-effective approach to identify DMRs around genic and well-established regulatory regions. DMRs are increasingly being identified as biomarkers, both diagnostic and prognostic, in many different tumour types including medulloblastoma (Meyer *et al.*, 2021). However, due to the relative paucity of inter-genic coverage, the platform is unable to comprehensively assess DMRs within intergenic regions.

Sequencing approaches are used to identify clinically important mutations in MB, with WGS and WES increasingly being employed in research (Northcott *et al.*, 2017). Whilst the use of whole genome sequencing is a clear gold-standard, the costs of generating sequencing data at high enough depths to reliably call variants (i.e., >30x) are out of reach for most. WES offers a cheaper alternative, where sequencing costs are lowered by focusing only on exonic regions. However, both sample preparation

and bioinformatic processing remains complex, and sacrificing whole genome coverage presents additional questions regarding its utility compared to other limited sequencing platforms. For diagnostic and/or research purposes, a mutation panel covering a range of genes relevant to MB and its associated heritable diseases are typically used to minimise costs (Mattocks *et al.*, 2010). The reliable identification of mutation status in MB remains an integral part of most clinical studies of MB as well as for patient stratification. The determination of *TP53* status in SHH tumours is a crucial prognostic indicator for patients, where tumours with *TP53* mutations are associated with 41% 5-year overall survival compared to 81% for tumours with wildtype *TP53* (Northcott *et al.*, 2019).

The employment of DNA methylation microarray for subgroup assignment and a multi-gene mutation panel are becoming prevalent in medulloblastoma diagnostics in developed nations. The use of the microarray has persisted over many years, which consequently has generated significant collections of tumour methylomes. Over 50,000 paediatric brain tumour samples now exist within the current DFKZ brain tumour classifier, and the clustering of over 1501 Group 3 and 4 samples has enabled the identification of further molecular subtypes within them, each associated with distinct survival profiles and clinically relevant markers of disease (Capper *et al.*, 2018; Sharma *et al.*, 2019). The main reason for the continued use of the array platform has been economic, since methylation sequencing platforms like WGBS have, until recently, required significant amounts of input DNA and library preparation/sequencing costs have been appreciably higher than what is needed for DNA methylation microarrays. However, these barriers to entry have since been lowered, and the cost of generating low-pass data (0.1 - <10x) is now reaching a comparable price to the DNA methylation microarray platform (in 2021: 10x WGBS = ~£430 / Array = £300). The major benefit of methylation sequencing is that all CpGs can be interrogated in the methylome, compared to just 1.5-3% from both Illumina Methylation 450k and EPIC platforms respectively (Zou *et al.*, 2018). Like the array, many additional analyses in addition to DNA methylation measurement can be performed using methylation sequencing assays. Differential methylation-based analyses can be performed with single nucleotide resolution across the entire methylome, far superior to the coverage that array platforms offer. Crucially, intergenic DMRs can be identified when using methylation sequencing, which are known to be associated with tumorigenesis (Rodriguez *et al.*, 2006; Timp *et al.*, 2014). Rather than inferring copy number through summed methylation signal intensities, as is done for the methylation microarray, copy number analysis can be performed at the sequencing read level, offering the potential for CNV to be detected at significantly greater resolution throughout the entire genome. For example, *SNCAIP* duplication, a marker of group 4 medulloblastoma, is a single copy tandem duplication event that is impossible to detect using methylation array due to its size and small number of probes located at that locus and instead requires sequencing-based approaches to detect reliably (Northcott *et al.*, 2012). CNV calling is well established for WES and WGS, and many tools are now available for use with WGBS (Talevich *et al.*, 2016; Raman *et al.*, 2019). The utility of methylation sequencing at low-pass to be used to call chromosomal copy number aberrations has not been extensively explored, and reference-free copy number (i.e., standalone CNV estimation without a diploid control cohort) calling



has yet to be tested in research. Methylation sequencing also offers the possibility of variant calling, with many numerous specialised bisulfite sequencing variants calling tools available (Liu *et al.*, 2021; Merkel *et al.*, 2019).

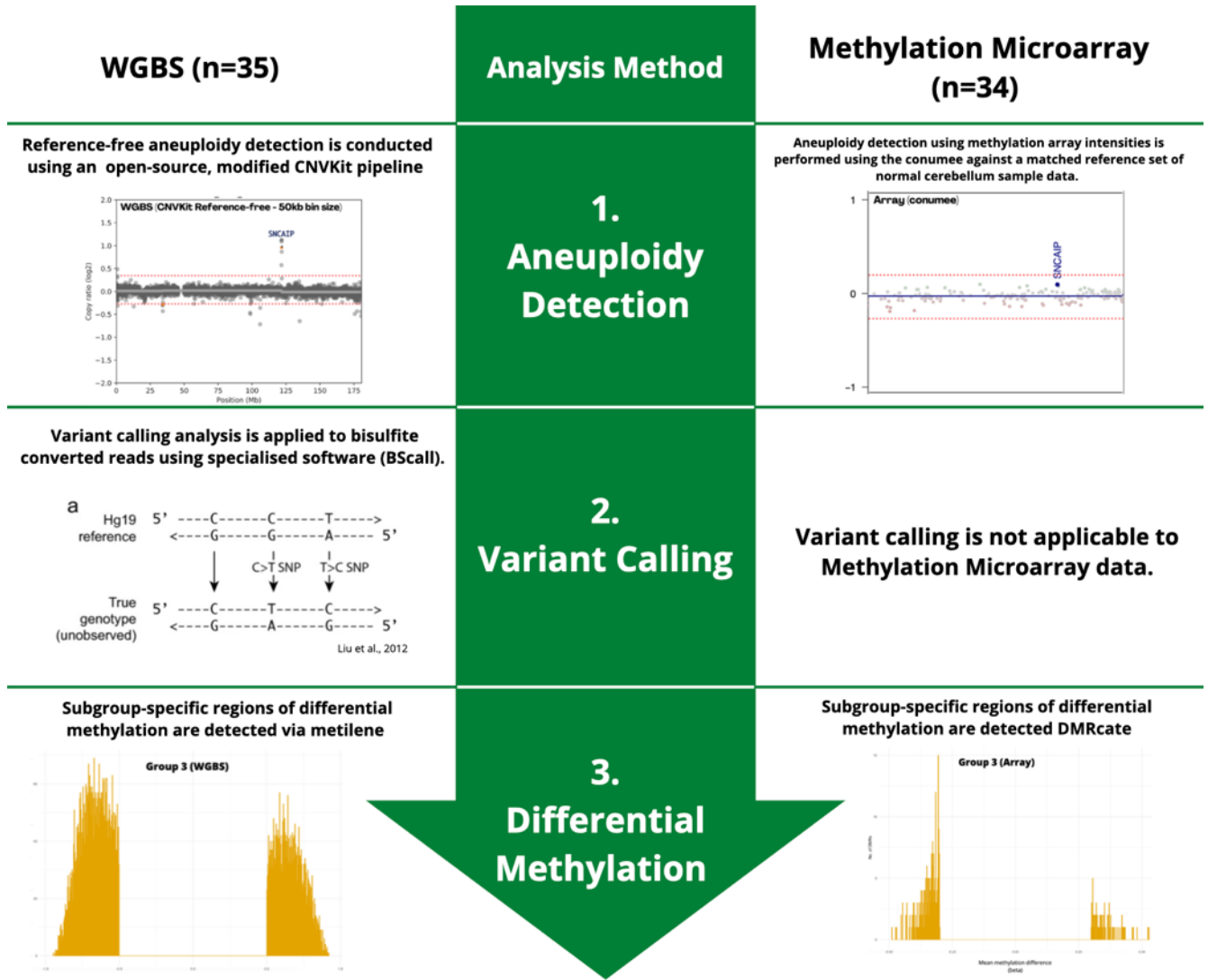
The use of WGBS throughout this project has demonstrated that the platform can be used effectively to generate high quality, full-resolution methylation data for the entire methylome at low pass that is highly correlated to array generated data in matched samples. The detection of increased variability and increased visual separation observed between the molecular subgroups of medulloblastoma when considering CpGs located in regions missed by the array platform provides the motivation to establish whole methylome profiles for each respective subgroup of medulloblastoma (Section 4.4.1). These findings demonstrate that the use of low-pass WGBS is at the very least on an equal footing to array for direct methylation-based analyses, such as methylome-based tumour subgrouping. Moreover, the platform potentially offers a higher-resolution alternative for copy number estimation and could also be used for variant calling, although reliable estimates may be coverage-dependent. Additionally, differential methylation analysis provides the possibility to define DMRs in intergenic regions, offering the potential to identify additional biomarkers specific to each subgroup/subtype. When considering the typical cost of both methylation array coupled with panel sequencing, the cost of low-pass WGBS becomes more comparable (2021 price: 10x WGBS = ~£430, Array + Panel = ~£550). As well as lowering cost, if WGBS can be utilised to derive both methylation and genomic diagnostic data, significantly less strain would be placed on diagnostic laboratories for MB diagnosis, as all essential data could be obtained using a single assay.

## 5.2 – Aims & objectives

The key aim of this chapter was to develop a range of optimised, clinically relevant analyses that form an important part of medulloblastoma diagnostics using low pass WGBS sample data - copy number variation, mutation detection and differential methylation. Using a cohort of internally sequenced samples (n=36), we aimed to develop an optimised, robust copy number calling pipeline for low-pass WGBS samples. As no raw control sample data (available in BAM file format) was available in this study, the optimised pipeline had to be capable of calling CNVs without a paired reference sample set. After a pipeline was established, identified regions of aneuploidy were compared to CNVs called from DNA methylation microarray analysis of matched samples. We also investigated whether reference-free CNV methodologies in WGBS could call focal regions of aneuploidy with higher resolution than that of array-based methods. The possibility of detecting disease-relevant mutations was investigated by developing pipelines for variant calling from low-pass WGBS. Finally, to investigate the landscape of DMRs within subgroups, a comprehensive analysis of DMR distribution and gene ontology enrichment was performed on WGBS data and compared to matched array samples, providing insights into the degree of intergenic subgroup-specific methylation patterns present in MB. The group 3/4 subtype VII had previously been shown to exhibit additional, clinically relevant heterogeneity (Sharma *et al.*, 2019) and the ability of WGBS to better distinguish this heterogeneity was investigated.

## 5.3 – Materials & methods

### 5.3.1 – Analysis Overview



**Figure 5.1** – Graphical overview of all analyses conducted within chapter 5. Further details (e.g., software, parameters used, references) are provided in subsequent sections.

### 5.3.2 – Cohort description

Aneuploidy detection via WGBS required that sample data was provided in the binary alignment map format (BAM). As our ICGC cohort (n = 42) was acquired as a large, pre-processed beta value matrix, these samples were not used during this analysis. Our NMB cohort, consisting of 36 primary medulloblastoma samples was taken forward for CNV analysis. Samples were prepared as described in 3.3.2 and raw data processed to convert fastq files into aligned, deduplicated high quality BAM files as described in 3.3.5. One sample, NMB\_390, was a matched FFPE/fresh frozen pair. Our corresponding matched DNA methylation microarray cohort was also limited to only NMB sample data (n=35) and was processed as described in 3.3.7.2. To test the capability of our proposed CNV methodology in identifying gene amplifications, our NMB cohort was carefully selected to ensure that several samples were present with known, validated amplifications of MYC (n= 9) and MYCN (n= 8) (Table 5.1).

<b>Characteristic</b>	<b>Internal Cohort (%)</b>
<b>No. of Samples</b>	<b>36</b>
<b>Sex</b>	
Male	<b>27 (75%)</b>
Female	<b>9 (25%)</b>
<b>Gene Amplifications</b>	
<u><i>FISH validated</i></u>	
MYC amplification	<b>4 (11.1%)</b>
MYCN amplification	<b>3 (8.3%)</b>
<u><i>MLPA validated</i></u>	
MYC amplification	<b>5 (13.9%)</b>
MYCN amplification	<b>5 (13.9%)</b>

**Table 5.1.** Gene validation data for NMB cohort used in aneuploidy analysis

Differential Methylation based analyses were conducted using methylation values – either beta values or logit-transformed beta values (i.e., M values), allowing analyses to be performed using all samples involved in this study (n = 78). As well as conducting differential methylation analysis in a subgroup specific manner, the methylome landscape of the molecular subtype VII was investigated, since there was clinico-pathological evidence for a further split as defined by NMF (Sharma *et al.*, 2019). The sequencing cohort was deliberately enriched for samples of this subtype (n = 13), enabling the investigation of the differential methylation landscape between these two NMF-defined groups – termed VII\_1 and VII\_2 (Table 5.2).

<b>Sample_ID</b>	<b>Subtype VII split (NMF)</b>
<b>NMB_77</b>	Grp4_LR1_1
<b>NMB_178</b>	Grp4_LR1_2
<b>NMB_185</b>	Grp4_LR1_2
<b>NMB_250</b>	Grp4_LR1_2
<b>NMB_416</b>	Grp4_LR1_2
<b>NMB_529</b>	Grp4_LR1_1
<b>NMB_733</b>	Grp4_LR1_1
<b>NMB_867</b>	Grp4_LR1_2
<b>ICGC_MB19</b>	Grp4_LR1_2
<b>ICGC_MB26</b>	Grp4_LR1_2
<b>ICGC_MB32</b>	Grp4_LR1_1
<b>ICGC_MB40</b>	Grp4_LR1_1
<b>ICGC_MB90</b>	Grp4_LR1_1

**Table 5.2** – NMF assignments of further subtypes present within subtype VII samples in our combined cohort (n=13). Subtype definitions (Grp4\_LR1\_1 and Grp4\_LR1\_2) were obtained from Sharma *et al.*, (2019).

### 5.3.3 – Chromosomal copy number estimation

#### 5.3.3.1 – Computing resources/software

All 36 NMB WGBS samples were analysed in the BAM file format using the ROCKET High Performance Computing service (Newcastle University). The ROCKET hardware infrastructure consists of 123 compute nodes in total, with the largest compute nodes consisting of 56 cores (27.4GB per core) and 8.7TB scratch space. A shared main data file store of 500TB is available to all users. Unless stated otherwise, all software was installed and executed within the Linux CentOS7 operating system. All scripts were written using VIM 8.2. WGBS CNV analysis was conducted using CNVkit v 0.9.9 (Talevich *et al.*, 2016), installed within a virtual Python language environment via Anaconda, an open-source data science analysis toolkit (Anaconda, 2021). Prior to CNV calling, BAM files were sorted into chromosomal order using the ‘sort’ command within SAMtools version 1.9 (Li *et al.*, 2009).

For our 35 methylation microarray samples, CNV analysis was conducted using the R statistical computing environment version 4.0.3 – “Bunny-Wunnies Freak Out” within RStudio (R Core Team, 2021). CNV analysis was conducted using conumee version 1.9.0 (Hovestadt & Zapatka, 2017).

---

### 5.3.3.2 – Establishment of an optimised reference-free CNV calling pipeline for use with low-pass WGBS sample data

As BAM files are used for CNV analysis of data obtained via sequencing-based platforms, our cohort was left without a set of control cerebellum samples to use as a reference set. This required the establishment of a suitable method of performing copy number analysis without reference data. Tools capable for CNV analysis with low-pass data are typically termed as depth of coverage methodologies (DOC). Briefly, DOC tools split sample data into genomic bins to mitigate the issues caused by sequencing gaps and low read depth which would otherwise contribute to inaccurate (or ‘noisy’) segments being called (Vardhanabhuti *et al.*, 2014). After binning, reads are typically normalised against a set of matched reference samples sequenced at the same depth. These act as a suitable guide of diploidy, from which test sample read counts can be normalised against to derive an accurate copy number ratio (increased read depth is indicative of higher copy number, and vice versa). Tools that offer CNV calling without a reference are uncommon, and even more so when using samples sequenced at low depth that have previously been bisulfite converted.

Following a literature search, it was determined that CNVkit had good potential to be used for CNV analysis of bisulfite sequencing data (Talevich *et al.*, 2016). The tool is open source, compatible with numerous sequencing methodologies including whole-exome, target panels and whole genome sequencing, and has a range of bias correction options that can be applied. One such correction is for GC content, which varies between 38-48% across all chromosomes, and its fluctuation is known to affect mapping consistency (Bentley *et al.*, 2008). This correction is optional within CNVkit rather than mandatory, making it more suitable to bisulfite treated DNA than other tools. Crucially, CNVKit allows for aneuploidy to be inferred from test data without control samples, offering the option to construct a ‘generic’ reference derived from the hg19 reference genome, where it automatically assigns a log<sub>2</sub> read depth of 0 to all bins. Theoretically, this should enable the calling of at least large-scale CNVs with relative ease when applying default settings, with the potential for smaller regions of aneuploidy to be detected should the pipeline be optimised. Other tuneable parameters are provided that make the tool more suited for dealing with whole genomes. The option ‘`–no-edge`’ prevents bias corrections within the reference and `fix` commands that deals with the lower depths typically observed around the end regions of exome reads which is also not necessary for whole genome sequencing data. Bin size can also be tuned, which is highly applicable when working with data sequenced at lower depths. To account for repetitive sequences that may interfere with read-depth calculations, CNVKit also provides a list of repetitive regions that are subsequently removed from both generic reference and sample data prior to analysis (Talevich *et al.*, 2016).

To establish the capability of CNVKit in predicting aneuploidy with low-pass WGBS data, we first applied it to one sample, NMB\_17, using default parameters unless stated in Table 5.3. Visual

inspection of identified segments showed high levels of noise present in the data. However, clear large-scale regions of aneuploidy appeared to be identified, providing motivation to optimise parameters further. Upon inspection of the default list of regions that were filtered, it was identified that many problematic regions such as centromeres, telomeres and bridge assembly gaps that are known to affect read mapping were not included in this list. Such regions are difficult to analyse and affect normalisation of test samples to any reference if not removed. Any effects would be compounded further in this case since a generic reference rather than a matched reference was used. The inclusion of centromeric and telomeric regions resulted in bins of ranging levels of ploidy, visually represented as stripes seen at the middle and ends of each chromosome when visually inspecting segments.

Platform	Package	Step	Module	Parameters
WGBS	CNVKit	Creating read bins	<i>autobin</i>	-m wgb's, '--target-min-size 50000'
		Determining bin coverage	<i>coverage</i>	default settings
		Creating flat reference	<i>reference</i>	--no-edge'
		Bias correction	<i>fix</i>	--no-edge'
		Infer copy number segments	<i>segment</i>	--drop-low-coverage', '--drop-outliers 100', '-t 1e-15'
		Assign absolute copy number integer	<i>call</i>	Gain = 0.3 / Amp = 0.7 Loss = -0.4 / Del = -1.1

**Table 5.3** – Summary of optimized CNVkit parameters for aneuploidy estimation of WGBS samples sequenced at low-pass.

To optimise the CN calling, the current CNVKit filter regions list were augmented with additional regions to be filtered, including centromeric, telomeres and other known regions of poor mappability (Table 5.4). Applying CNVKit using these parameters to five samples resulted in much improved segmentation. However, all called segments were characterised by undesirably high standard deviation values for all segments (SD = 0.28) – indicating that bins may be spread across calling thresholds. Whilst this is less problematic for large scale changes (whole chromosome arms), it is a sign of poor-quality segmentation when dealing with focal regions, making it difficult to determine whether smaller segments are artefactual, a false call or a genuine gain/loss. Low sequencing depth was a likely contributing factor to the segment log<sub>2</sub> variability observed across the genomes, where coverage is often variable. Increasing the bin size is a measure often employed to reduce noise in CNV outputs, resulting in more accurate and reliable segmentation. This is at the expense of resolution but is a necessary sacrifice to make when attempting to optimise a reliable method of calling CNVs with low-coverage data. The combination of more stringent filtering and increased window size (from 10kb to 50kb) resulted in much improved segmentation, with a 58% reduction in segment variability across all called segments in the same five samples (SD = 0.12). Validated focal amplifications of MYC and MYCN could clearly be identified, providing confidence that these set of parameters were optimal and could be used to call focal regions of 150kb or larger reliably (Figure 5.2). Following parameter optimisation, CNVKit was applied to all 36 samples in the NMB cohort. Segment co-ordinates were cross-referenced with UCSC Sequence and Annotation data for the hg19 reference genome to

annotate any genes contained within each segment. Segment statistics were also tabulated and bin-level log<sub>2</sub> ratios and segment calls were visualised for all chromosomes via scatter plot.

<b>Filter regions</b>	<b>Filename</b>
Anomalous, unstructured regions identified across NGS experiments	ENCODE_Blacklist.bed
Regions with local alignment issues	GIAB_Align_Problem.bed
CNVKit default regions	access-5k-mappable.hg19.bed
Known assembly gaps within hg19 assembly	hg19_AssemblyGaps
Unusual regions on assembly structure	UCSC_Unusual_Regions.bed
Anomalous, unstructured regions identified across NGS experiments	wgEncodeDacMapabilityConsensusExcludable.bed
Centromeric co-ordinates	centromeres.bed
Telomeric co-ordinates	telomeres.bed
Problematic regions (e.g. satellite/rRNA sequences)	wgEncodeDukeMapabilityRegionsExcludable.bed

**Table 5.4** – List of filter regions incorporated into CNVKit pipeline. Co-ordinate lists were obtained from the ENCOD genome browser (ENCODE, 2022)

### 5.3.3.3 – Validation of *MYC*/*MYCN* amplifications

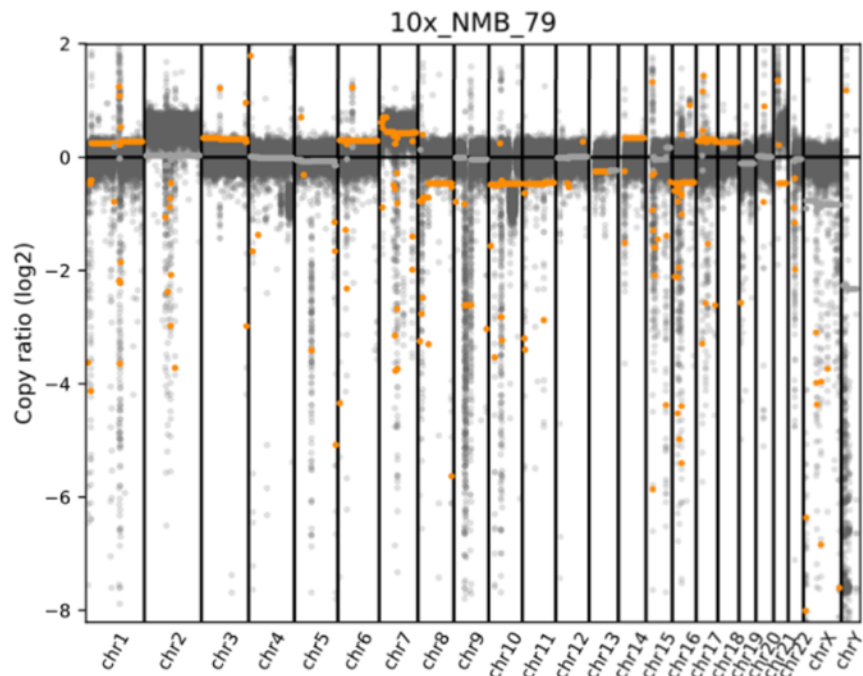
To test the sensitivity of our optimised CNV pipeline in calling clinically relevant gene amplifications of MB, samples were included with previously validated amplifications of *MYC* and *MYCN* (see table 1). Samples with *MYC* and *MYCN* amplifications in our NMB cohort were previously assessed using either fluorescence in situ hybridization (FISH) or via multiplex ligation-dependent probe amplification assay (MLPA) (Table 5.5).

<b>Sample ID</b>	<b>Amplification</b>	<b>Validation Method</b>
NMB_77	<b>MYCN</b>	<b>FISH</b>
NMB_181		<b>FISH</b>
NMB_363		<b>MLPA</b>
NMB_549		<b>FISH</b>
NMB_774		<b>MLPA</b>
NMB_787		<b>MLPA</b>
NMB_867		<b>MLPA</b>
NMB_870		<b>MLPA</b>
NMB_169		<b>MYC</b>
NMB_273	<b>FISH</b>	
NMB_318	<b>MLPA</b>	
NMB_758	<b>FISH</b>	
NMB_769	<b>FISH</b>	
NMB_774	<b>MLPA</b>	
NMB_787	<b>MLPA</b>	
NMB_858	<b>MLPA</b>	
NMB_870	<b>MLPA</b>	

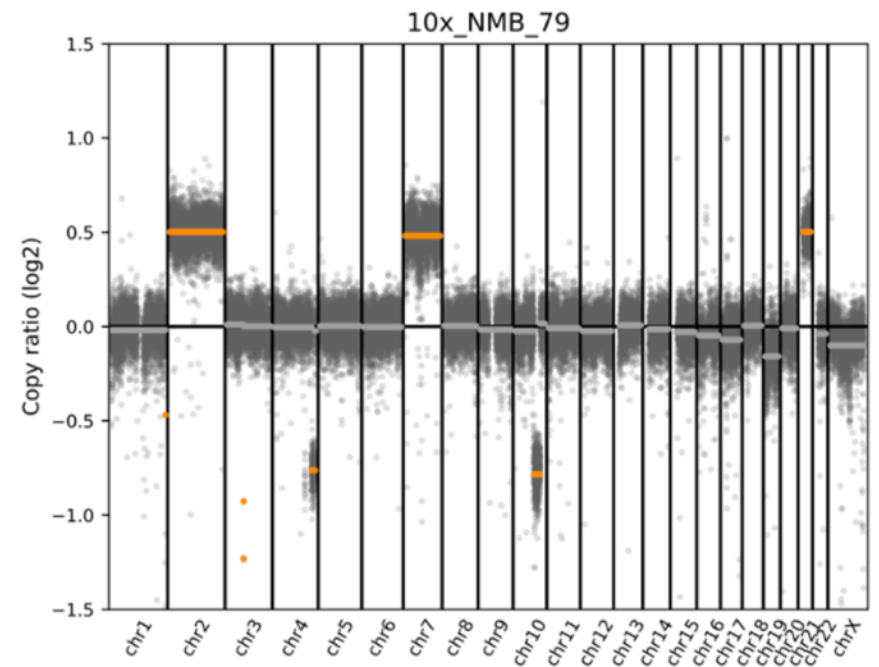
**Table 5.5** – List of samples in NMB cohort with known, validated *MYC*/*MYCN* amplifications



## Default



## Optimised



**Figure 5.2** Optimisation of CNVKit for low-pass WGBS. The inclusion of additional filter regions, threshold adjustment and increased bin size results in effective read binning, segmentation and calling of both large and focal regions of aneuploidy.

---

### 5.3.3.4 – Copy number variation analysis of matched DNA methylation microarray cohort

Aneuploidy detection was conducted on our matched array cohort (n = 35) using the Bioconductor package *conumee* (Hovestadt & Zapatka, 2017) as previously described (Schwalbe *et al.*, 2017). 119 'control' samples were used as a reference for normalisation, obtained from the DFKZ Paediatric Brain Tumour classifier control reference cohort (Capper *et al.*, 2018). Segment co-ordinates were recorded, and statistics were tabulated. All called segments via *conumee* were cross-referenced against those called by our CNVKit pipeline to determine the degree of concordance between both platforms. Segments were deemed as concordant if greater than 50% of their length were overlapping.

### 5.3.4 – Variant calling of low-pass WGBS cohort

To investigate the capability of using low-pass WGBS data to call variants, we used the specialised methylation and sequence variant calling program 'bs\_call' (Merkel *et al.*, 2019). *Bs\_call* forms part of the popular gemBS pipeline for bisulfite sequencing data processing/analysis and has been shown to call variants with high accuracy in high-depth cohorts of samples. For variant calling, 'bs\_call' takes advantage of both reads, applying Bayesian inference to calculate the probability of each possible genotype at any given loci, accounting for bisulfite conversion rate, methylation ratio and sequence error rate. For example, T->C SNPS (shown as T:A in the reference genome) will present as T:G base pairs in bisulfite treated reads. For each specific site, the genotype with the highest maximum likelihood estimate is selected, allowing base content to be assigned accurately. In line with GATK best practices, each variant is assigned various established quality measures including phred scaled genotype quality (GQ), read depth (DP) and mapping quality (MQ).

*Bs\_call* was initially applied to four samples, NMB\_169, NMB\_181, NMB\_436 and NMB\_787. Variants were assessed in the regions of the genome measured by current diagnostic target sequencing panels used in MB, which surveys 60 medulloblastoma-associated genes at 1644 loci (gene list provided in the appendix – 8.1.8). Variants were annotated with Jannovar version 0.36 (Jager *et al.*, 2014) using the hg19 RefSeq gene database (NCBI, 2022). Anticipating a low yield of high-quality variants, a soft filtering approach was used, followed by manual inspection of all filtered variants. Variants were filtered using the following parameters: QUAL >equal 20, DP >equal 2, MQ >equal 40. As expected, the total number of variants called for each sample was low, with a mean depth across all variants of 5.39x; this result led to a decision to abandon variant calling analysis for remaining samples, due to the very high chance of non-confidently called variants called using low-pass WGBS.

### 5.3.5 – Differential methylation analysis

Following aneuploidy detection and variant calling, subgroup specific regions of differential methylation were identified in MB across the entire methylome using WGBS, and the number, distribution, and ontology of differentially methylated regions (DMRs) called compared were compared to DMRs detected using matched DNA methylation microarray samples. The cohort was split into five groups: WNT (n = 10), SHH (n = 11), Group 3 (n = 23), Group 4 (n = 26) and Normal Cerebellum (n = 8). In addition to subgroup specific DMR calling, we also investigated whether differential methylation was present within subtype VII of the current consensus group 3/4 molecular subtypes, where there was evidence of a further split with clinical significance identified via NMF unsupervised clustering of DNA methylation microarray data (Sharma *et al.*, 2019). To identify differential methylation within this subtype, we performed a direct comparison between both VII\_1 and VII\_2 samples subsets (VII\_1 = 6, VII\_2 = 7). Group VII\_1 and VII\_2 samples corresponded to groups denoted as Group4\_LR\_1 and Group4\_LR\_2 as defined by Sharma *et al.*, (2019) (Section 5.3.1).

#### 5.3.5.1 – DMR detection of low-pass WGBS samples via metilene

To identify regions of differential methylation for WGBS sample data, we used metilene (Juhling *et al.*, 2016). Rather than utilise matrix statistics, metilene utilises circular binary segmentation (CBS) to identify differences between mean methylation levels between sample groups. Briefly, the whole genome of samples is split into regions of consecutive methylation values with gaps no larger than 300 nucleotides. For these regions, the mean methylation difference for all CpGs is calculated between sample groups and circular binary segmentation is applied to identify changepoints in methylation. Regions are recursively segmented until a DMR reaches a minimum size of 10 CpGs or the p-value of any given subsegment is larger than the 'parental' segment. The statistical significance of segments is determined via Kolmogorov-Smirnov test. Metilene was chosen due to its demonstrated superior performance compared to other popular tools for calling DMRs with methylation data – MOABS (Sun *et al.*, 2014) and BSmooth (Hansen *et al.*, 2012) by the author (Juhling *et al.*, 2016). In this study, performance was assessed by validating the DMRs detected against known DMRs that correlated with gene expression in the same group 4 samples (n=12) and control cerebellum samples (n = 8) that comprise our ICGC sample cohort (Hovestadt *et al.*, 2014). Out of the 39 DMRs that correlate with gene expression, 29 were identified using metilene, the highest of all tools tested (Juhling *et al.*, 2016). We therefore judged the tool to be highly suitable for analysis, given its demonstrable excellent performance when applied to samples from this study. For each group comparison, we selected DMRs that exhibited a mean difference in methylation of 0.5 or greater.

---

### 5.3.5.2 – DMR detection of array data using DMRcate

DNA methylation microarray subgroup specific DMRs were identified using the Bioconductor packages limma version 3.4.6.0 (Ritchie *et al.*, 2015) and DMRcate version 2.4.1 (Peters *et al.*, 2015). Briefly, limma fits a linear model to all CpG sites for each sample, producing a matrix of M value differences between each sample where each probe site is a row, and each sample is a column. A paired t-test is calculated for each row between sample groups, computing both the mean methylation difference and its associated standard error for each probe. A t-statistic is also calculated, denoting the ratio of the M value compared to its associated standard error. DMRcate computes p-values based upon a Gaussian smoothed model of all t-statistics, enabling significantly differentially methylated probe sites to be identified. Differentially methylated probes are then agglomerated into DMRs, provided they are within 1000bp of each other. This agglomeration continues until the mean methylation difference significantly decreases, producing a final DMR with a differential methylation beta value calculated based on the average methylation difference across all inverse log transformed M values. For each group comparison we considered only DMRs that exhibited a methylation beta difference greater than 0.3.

### 5.3.5.3 – Subgroup-specific DMR annotation and gene ontology analysis

To identify subgroup specific DMRs, two group comparisons for each molecular subgroup were performed. Differentially methylated regions were identified between the subgroup of interest and normal cerebellum, and between the subgroup of interest and the remaining 3 subgroups. Any DMRs that were identified in both comparisons were determined to be subgroup-specific (i.e., DMRs that only occur within the subgroup of interest). To investigate the distribution and characteristics of DMRs specific to each subgroup, the annotatr R package (Cavalcante & Sartor, 2017) was used to annotate DMRs relative to its genomic position within the hg19 reference genome. The annotation distribution of all subgroup specific DMRs was recorded and plotted using ggplot2 (Wickham, 2016). For DMRs that annotated to genic regions, we conducted a gene ontology enrichment analysis using the list of genes affected by differential methylation with a focus on biological processes. Genic regions include regions 1-5kb upstream of the TSS, promoter regions, 5'UTR/3'UTR, exons/introns, and coding sequence regions. Significantly enriched terms were identified by applying an adjusted p-value (calculated via Fishers exact test) filter of 0.05. Remaining terms were then ranked based upon their respective “combined score” – a robust ranking measure that incorporates the calculated p-value with a z-score – which measures the deviation of each genes rank relative to its expected rank within each gene set library (Chen *et al.*, 2013). To further refine the list of over-represented terms, terms were filtered using a combined score cut-off of >15 to identify the most significantly enriched biological processes affected by differential methylation for each subgroup.

## **5.4 – Results & discussion**

### **5.4.1 – Application of an optimised reference-free CNVKit pipeline to low-pass WGBS samples successfully identifies large regions of aneuploidy with excellent concordance to current array methodologies**

Application of our optimised CNV pipeline to low-pass WGBS sample data showed excellent performance regarding the identification of both whole chromosome and single-arm regions of aneuploidy. When comparing array-derived whole and single-arm calls as a reference, our optimised pipeline called these regions with 96.9% sensitivity, with just 7 out of 36 samples not attaining 100% sensitivity (Figure 5.3). CNVs called by both platforms include many chromosomal signatures previously associated with specific molecular subgroups of MB, including isochromosome 17q, monosomy 6, gain of chromosome 7 and loss of chromosome 11.



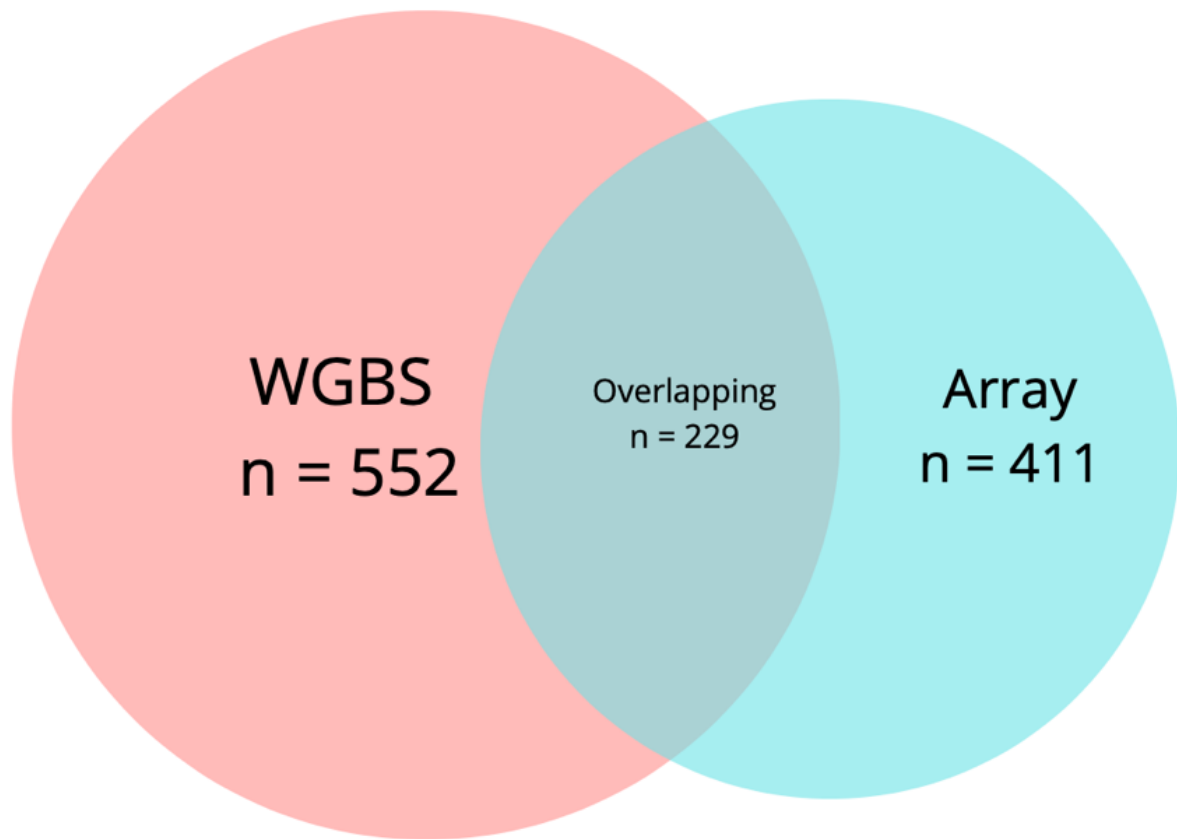
Outside of concordant segments, the WGBS pipeline identified a greater number of whole chromosomal calls which were missed when analysing samples via array. A total of 23 additional whole chromosomal aberrations were identified in the sample cohort compared to paired methylation microarray samples (Figure 5.4). Manual inspection of whole-chromosomal aberrations called by WGBS showed that these segments were high quality (characterised by low segment standard deviation values and upper/lower confidence intervals that did not overlap with its respective calling threshold) across bins contained within non-matched segments called by WGBS that were not identified via array (an example for NMB\_362 is provided in the appendix – 8.1.6). NMB\_17, the sample that exhibited the largest discordance in the number of whole and single-arm chromosome calls compared to its matched array sample, exhibited a mean log<sub>2</sub> copy ratio of 0.4 (SD = 0.11, CI = 0.39 – 0.40) and -0.47 (SD = 0.13, 95% CI = -0.47 – -0.48) for gains and losses respectively. Both upper and lower 95% confidence intervals fall comfortably above and below threshold values for gain and losses, indicating that whole chromosomal segment calls are robust and high in quality. Inspection of array segmentation calls for NMB\_17 showed that chromosomes 8, 10, 11 & 16 losses exhibited lower copy numbers but fell below the threshold of -0.22 used in this study (chr8 = -0.214, chr10 = -0.186, chr11 = -0.155, chr16 = -0.177). Further validation is needed, but it is likely that array-based methods have simply fallen short in calling these regions of loss due to the lower number of positions (probe sites) from which to infer copy number. Sequencing platforms, even at lower depths calculate copy number with bins that span entire chromosomes, providing a more robust and superior means of CNV detection (Coutelier *et al.*, 2021).

A similar number of segments were called between both platforms, with a mean number of segments (41 and 39 called across all samples for WGBS and Array pipelines respectively ( $p = 0.491$  – paired t-test) when focusing solely on called gains and losses (Table 5.6). Whilst segment numbers were similar, segments identified as gain/losses that overlapped between both platforms varied. When considering segments identified from one platform that overlapped with at least 50% of the length of a segment identified by the other platform (multiple segments that overlap with one called by another were counted as singular), it was found that gain/loss calls via WGBS overlapped with 62% of array called gains/losses (Figure 5.4). Whilst relatively low, this degree of overlap is commonly seen when comparing tools used to predict copy number (Kilaru *et al.*, 2019). Differences can be caused by several technical factors, including calling thresholds, segment size and total number of bins per segment. Investigating the average segment size for identified gains/losses between WGBS and array showed that average segment size was similar (WGBS = 39.4mb, Array = 36.4mb), showing that although large-scale changes are predicted with high concordance, differences in segment calling do exist between the WGBS pipeline and conumee and which require further investigation.

Sample_ID	No of called segments		No. of Gains called		No. of Losses called	
	WGBS	Array	WGBS	Array	WGBS	Array
NMB_17	38	28	7	4	11	0
NMB_77	37	36	5	3	7	4
NMB_79	31	34	3	2	4	4
NMB_169	57	25	7	3	8	2
NMB_178	34	46	5	5	6	10
NMB_181	63	59	9	10	18	18
NMB_185	44	40	13	9	8	7
NMB_187	32	27	1	2	5	2
NMB_203	39	44	3	1	6	7
NMB_250	77	31	9	3	22	3
NMB_273	41	50	5	8	11	14
NMB_318	38	49	8	7	7	18
NMB_362	36	38	6	6	7	7
NMB_363	34	38	0	1	7	5
NMB_378	37	40	8	6	11	17
NMB_390	36	31	1	2	6	3
NMB_390_FFPE	58	NA	10	NA	10	NA
NMB_416	42	28	10	4	8	4
NMB_433	32	27	8	5	9	5
NMB_436	30	28	1	1	3	4
NMB_529	31	36	4	7	7	7
NMB_549	86	78	12	15	21	19
NMB_610	27	34	3	5	4	3
NMB_725	38	33	5	3	5	5
NMB_729	29	32	2	1	3	1
NMB_733	45	39	12	14	9	5
NMB_758	46	49	10	10	11	12
NMB_769	35	47	5	9	5	8
NMB_772	39	38	6	7	6	3
NMB_774	42	45	4	3	6	7
NMB_787	32	39	4	7	4	6
NMB_803	28	32	2	2	3	1
NMB_858	30	32	2	3	3	2
NMB_867	48	47	11	6	10	2
NMB_870	28	26	0	1	3	2
NMB_890	40	49	7	7	6	11
Mean	40.556	38.714	5.778	5.200	7.778	6.514
Significance (p-value)	0.491		0.302		0.167	

Table 5.6 – Number of segments called by WGBS and array CNV pipelines for matched NMB samples. Gains and losses are also tabulated separately to check for any statistical bias in calling (assessed via paired t-test).





**Figure 5.4** – Number of segments called by CNVKit (WGBS) and conumee (Array) that were overlapping. Segments were deemed as overlapping if at least 50% of its length was encompassed by a corresponding segment called by the other platform.

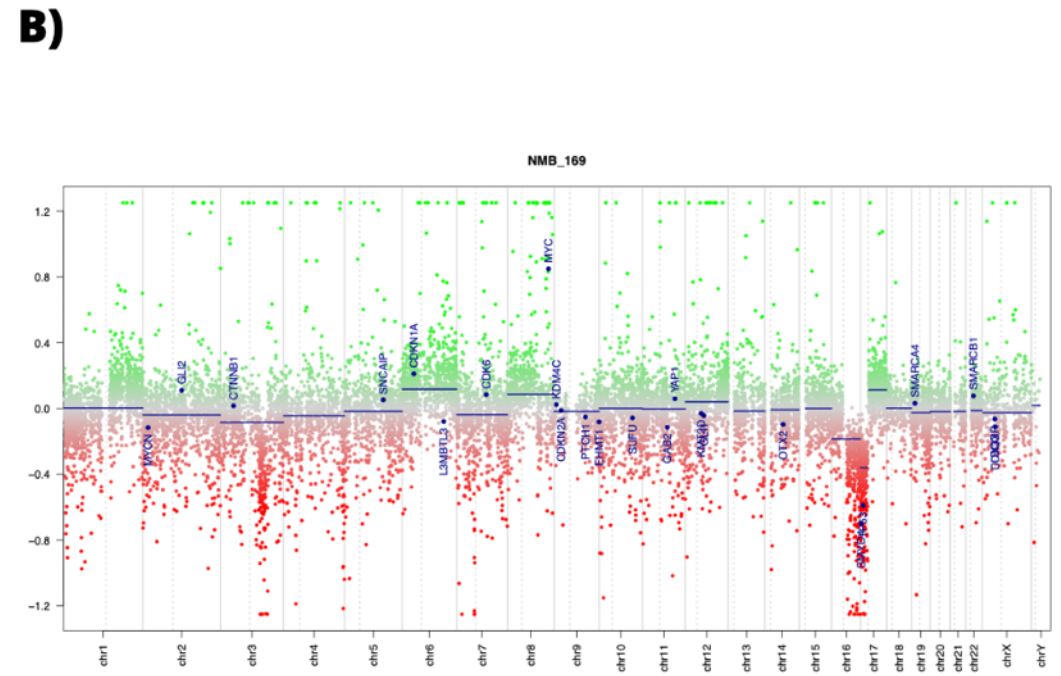
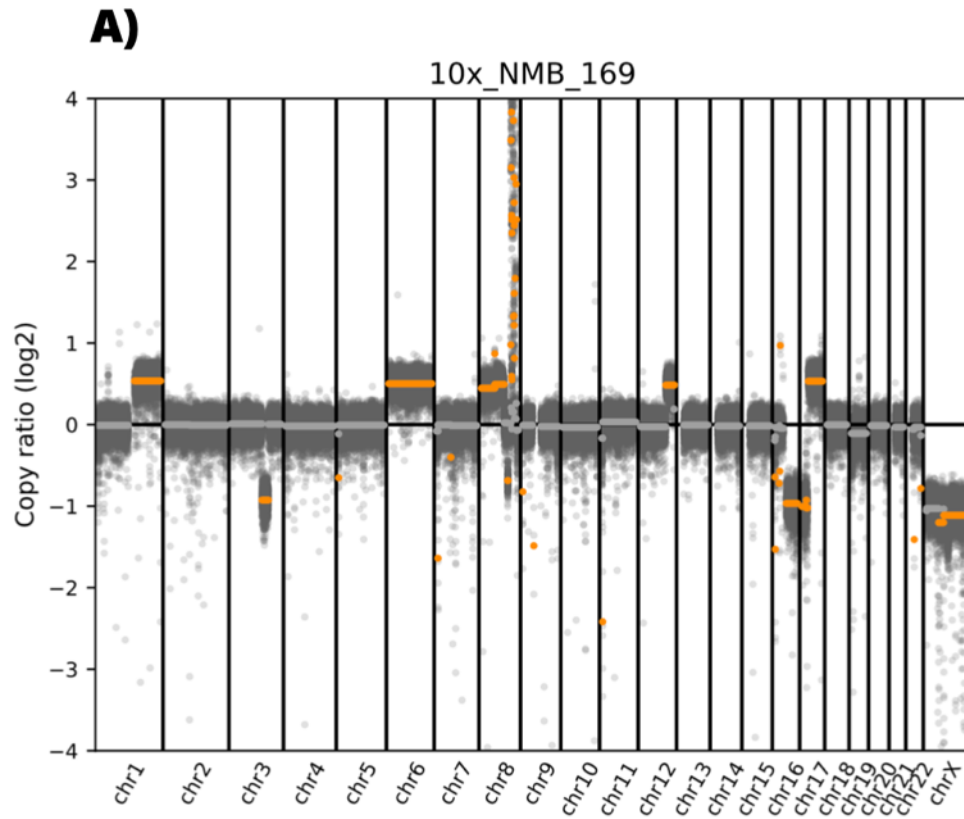
### 5.4.2 – Validated amplifications of MYC and MYCN are identified with improved sensitivity

To test the sensitivity of the optimised CNVKit pipeline to detect focal regions of amplification and deletion, the number of called amplifications of *MYC* (n=9) and *MYCN* (n=8) was compared in samples where the events have been validated with either FISH (*MYC* = 4, *MYCN* = 3) or MLPA (*MYC* = 5, *MYCN* = 5) and further compared to the number of amplifications detected using matched array data via conumee. For DNA methylation microarray, conumee identified 4 out of 9 *MYC* amplifications (sensitivity = 44.44%) and 2 out of 8 *MYCN* amplifications respectively (sensitivity = 25%). Sensitivity in detecting FISH validated amplifications in isolation was markedly higher, where 3 out of 4 *MYC* amplifications (75% sensitivity) and 2 out of 3 *MYCN* amplifications (66% sensitivity) were identified. The *MYC* amplification that was missed appeared to be a convoluted issue regarding segmentation. The individual gene score (a more accurate log<sub>2</sub> measure of the bins that the sole gene of interest encompasses) was high (log<sub>2</sub> = 0.8), but an overall gain of chromosome 8 (log<sub>2</sub> = 0.202) has overshadowed the amplification, resulting in no separate segment being identified. The sensitivity of the reference free CNVKit pipeline was marginally improved, where 5 out of 9 *MYCN*

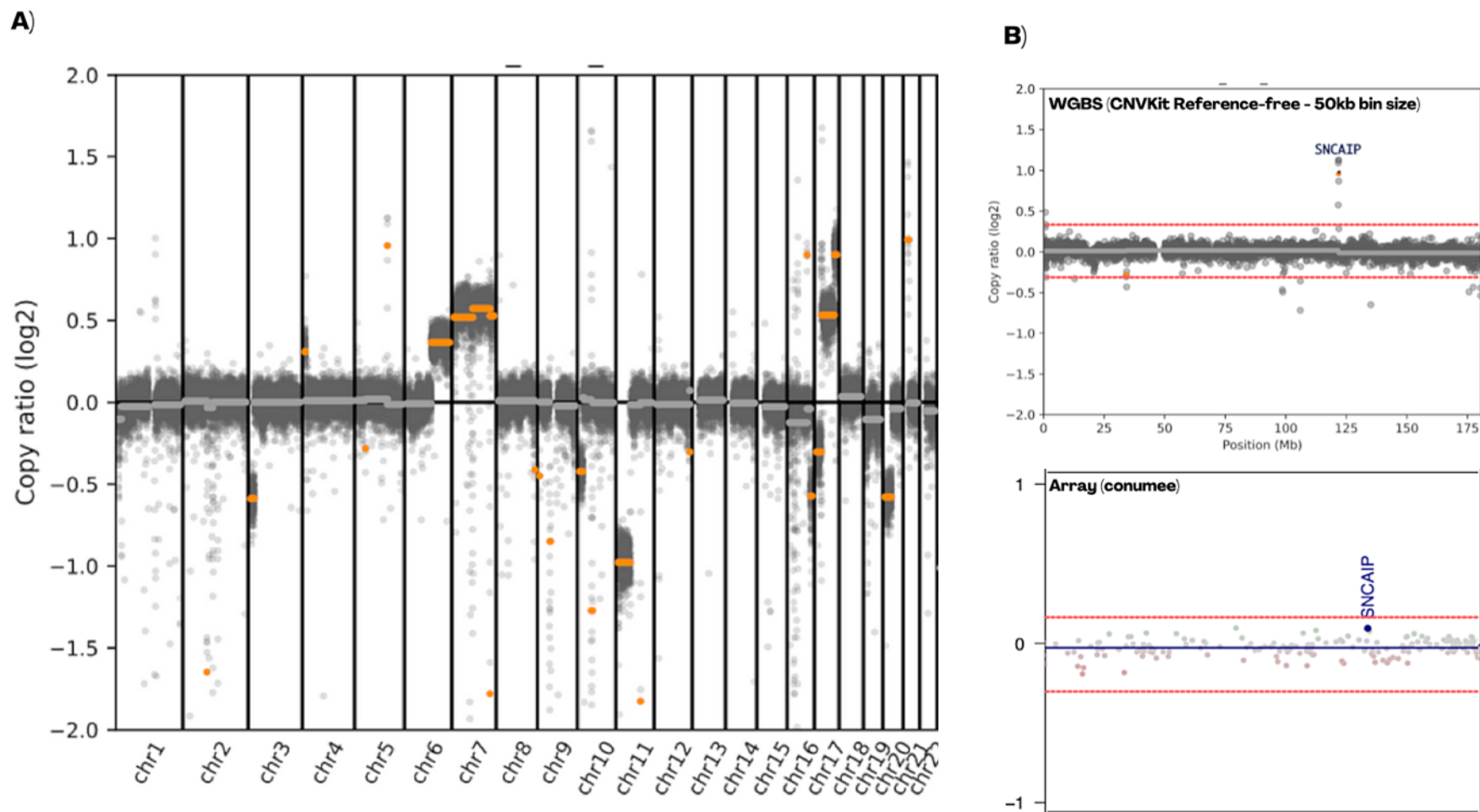
amplifications (sensitivity = 55.56%) and 2 out of 8 *MYCN* amplifications (sensitivity = 25%) were identified respectively. All FISH validated *MYC* amplifications were detected in our WGBS cohort (sensitivity = 100%), whereas 2 out of 3 *MYCN* amplifications were successfully identified (sensitivity = 66%). For NMB\_169, amplification of *MYC* was clearly identified, where amplified segments were successfully called independent of the gain observed across the whole of chromosome 8 (Figure 5.5).

MLPA called amplifications were not detected as reliably by either platform. For DNA methylation microarray, no *MYC* or *MYCN* amplifications were identified. Both NMB\_318 and NMB\_774 had whole arm and whole chromosome gains on 8 identified rather than a *MYC* amplification respectively. For NMB\_318, a similar situation arose to NMB\_169, where the individual gene score was high (0.516), but the overall segment value was called for the entire chromosome arm rather than a raised segmental value for the focally amplified region. The reference free approach for WGBS successfully identified this *MYC* amplification in NMB\_318, where a low-level amplification was successfully identified. MLPA has been shown to identify smaller regions of amplification compared to FISH, offering an explanation as to why both methods in this study were not as sensitive in calling regions validated by MLPA. It is likely that smaller regions could be identified via WGBS had we sequenced at higher depth, but we subsequently had to increase our minimum bin size to 50kb to increase robustness for calling for low-pass WGBS sample data.

Far greater numbers of focal regions were called from WGBS compared to DNA methylation microarray. Out of all segments identified, 51 were called that were smaller than 1mb in size via conumee, whereas 148 were identified using CNVKit. One such additional focal region identified was a *SNCAIP* duplication in sample NMB\_362 (Group 4, Subtype VI,  $\log_2 = 0.96$ ). This event is completely missed by conumee, where a balanced profile is observed (Figure 5.6). As window sizes were identical, the increased number of focal regions identified was likely solely due to the increased coverage offered by WGBS as opposed to a limited number of probe loci by the array. This demonstrates that increased genome coverage offered by WGBS still provides additional power in calling regions of focal amplification or deletion at low pass compared to the array, where the uneven distribution of probes makes it difficult to identify all regions reliably.



**Figure 5.5** - Presence of a MYC amplification that is identified independent of the remaining gain in chromosome 8 by CNVKit. (A) but is subsequently missed during segmentation by conumee (B), where an entire chromosomal segment is called. All dots correspond to bins (minimum size = 50kb), whereas horizontal lines correspond to segments. WGBS segments that differ from diploid copy number are denoted by orange colour.



**Figure 5.6** - Reference-free aneuploidy detection using CNVKit can be used to call focal events of aneuploidy that are missed when using methylation microarray. **A)** Whole genome plots of log<sub>2</sub> copy number ratio via CNVKit for NMB\_362. Identified segments are represented by horizontal lines. Segments deemed as significant are coloured orange. **B)** Chromosome 5 plots of log<sub>2</sub> copy number ratio for NMB\_362 identified via CNVKit (top) and conumee (bottom). Red dotted lines represent thresholds used to define gains and losses.

### 5.4.3 – Analytical limitations of copy number analysis

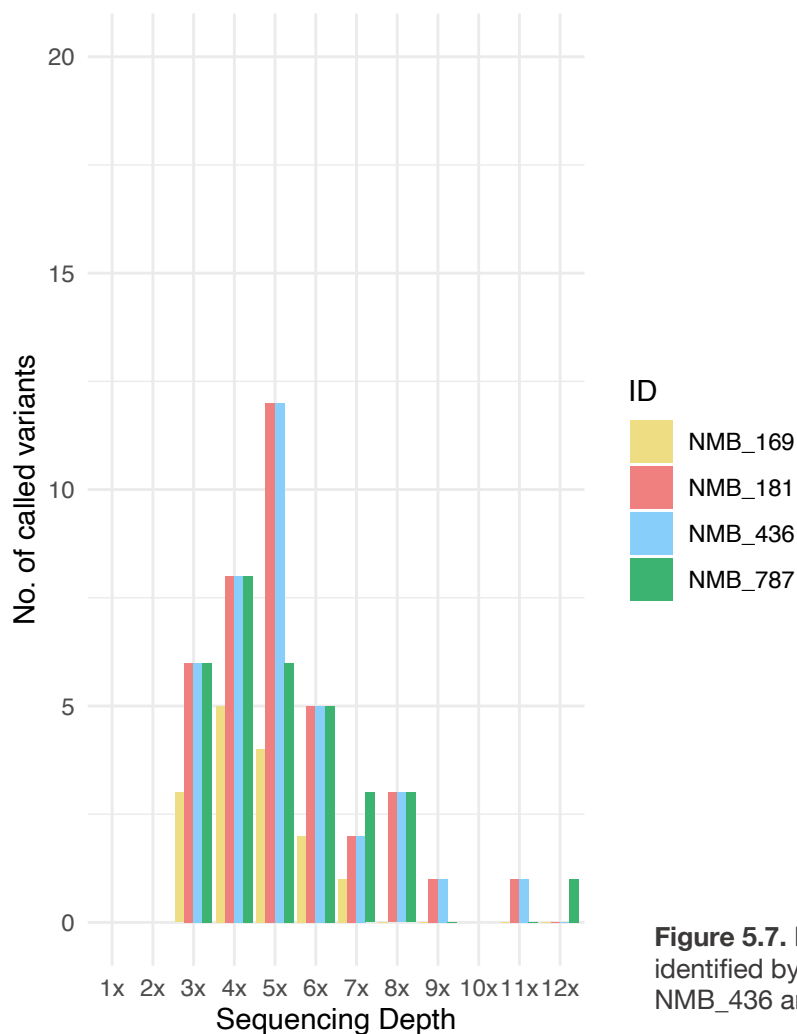
Whilst the data presented in this chapter shows that our optimised reference-free CNV calling pipeline can identify both large and focal regions of aneuploidy with high sensitivity, it is subject to some limitations. Like the issues identified in 4.4.4.6, the modest sample size used in these analyses ( $n = 36$ ) makes it difficult to know whether the pipeline is amenable across a wide range of tumour types. Whilst samples were selected with care to ensure a significant number of samples with validated *MYC/MYCN* amplifications were present, more sample data is needed to comprehensively establish the true sensitivity of the optimised CNVKit pipeline. This is more apparent when considering the additional identification of other clinically relevant CNVs like that of *SNCAIP*, where just one event was recorded in our study. Whilst the sample belongs to group 4 where the event is enriched, validated data for *SNCAIP* in this study is lacking and such information for *SNCAIP* and indeed other genes would be necessary to establish the true capability of our pipeline to detect focal amplifications across the genome.

All DNA methylation microarray samples used in this study were analysed using the Illumina Methylation450k platform, which has since been superseded by the MethylationEPIC array. The EPIC array offers almost twice as much coverage as the 450k platform and has been shown to be marginally superior in detecting regions of aneuploidy compared to its predecessor (Kilaru *et al.*, 2020). It would therefore be prudent that a comparison be made between the optimised pipeline for WGBS and segments identified using the EPIC array to provide a contemporary comparison between both platforms. In 2022, the 450k probe remains platform of choice for classification, and for routine tumour/subgroup assignment, EPIC samples are subset to probes that match to those of its predecessor (~90% of probes are matching), and many analyses that combine data from both platforms are also subject to the same limitation.

Finally, it must be acknowledged that the calling thresholds that were established for our optimised CNVKit pipeline are subjective to the user and decision making regarding CNV thresholds in general can vary between studies. Typically, they are determined by individuals experienced in aneuploidy analysis via manual inspection of segmentation plots for a handful of samples with clearly recognisable events of ploidy (e.g., WNT samples with monosomy 6) and/or validation data regarding chromosomal aberrations (e.g. *MYC* amplification via FISH). Many 'near-miss' events can be seen in segmentation data from both platforms, particularly from our CNVKit pipeline, suggesting that there may be further room for threshold optimisation. However, adjusting thresholds must be done with caution, since it presents the risk of identifying false-positive regions of aneuploidy. It is usually more appropriate to correct near-miss events manually, where such surrounded by segments called as clear gains or losses can be recovered.

### 5.4.4 – Reliable variant calling is beyond the capability of low-pass WGBS

The reliable detection of *TP53* and *CTNNB1* mutations forms a crucial component of MB diagnostics, where all tumours typically undergo mutation assessment via a multi-target NGS panel. Standard NGS panels for medulloblastoma typically seek to identify variant status of ~50-60 genes implicated in MB tumorigenesis and its associated Mendelian disorders. WGBS offered the possibility of uniting both methylation and sequencing-based diagnostics under one platform, and despite the low depth of our NMB cohort, it was investigated whether BS-seq variant calling software could call mutations from low-pass data. Application of 'bs\_call' to our initially sequenced 4 pilot samples (NMB\_169, NMB\_181, NMB\_436, NMB\_787) generated poor quality results, identifying 1,642, 1,746, 2,059 and 1,724 variants for each sample respectively. Anticipating most variants to be low coverage, we applied a soft filtering approach that ignored sequence depth and focused solely on QUAL score. For all samples, this reduced the number of variants significantly, where 15, 27, 38 and 32 variants passed filtering for NMB\_169, NMB\_181, NMB\_436 and NMB\_787 respectively. Manual inspection of all remaining variants revealed that depth was low across all variants (Figure 5.7). Out of 112 passing variants, just 3 were identified above 10x.



**Figure 5.7.** Depth frequency of passing variants identified by bs\_call for NMB\_169, NMB\_181, NMB\_436 and NMB\_787.

Identifying variants at low pass was always expected to be a difficult task. GATK best practices recommend that quality score calibration is performed below these depths, which `bs_call` does not offer – adding further doubt to any called variants in this analysis. Whilst CNV calling pipelines could feasibly be optimised at low-pass due to read agglomeration strategies, such strategies cannot be applied to variant calling, where read depth plays a crucial role in genotype determination. Indeed, the application of a filter greater than 10x may be necessary when dealing with WGBS sequencing data, where bisulfite treatment of DNA can hamper genomic coverage distribution. Despite this, bisulfite seq-specialised software has been shown to call variants reliably, albeit at high sequencing depths (Merkel *et al.*, 2019). Popular BS-seq mutation callers such as Bis-SNP have shown that depths as low as 16x can detect variants with comparable false discovery rates to depths as high as 32x (Liu *et al.*, 2012), but most tools still recommend depths of at least 50x to call variants reliably using WGBS data. Given the potential to incorporate mutation analysis as part of WGBS analytical pipelines, further research is required to establish what minimum level of coverage could be used to call variants reliably with this platform – an important consideration given the economic cost of increasing sequencing coverage.

#### **5.4.5 – Greater numbers of differentially methylated regions are identified from WGBS data compared matched array**

Subgroup specific analysis of our combined array cohort via DMRcate identified 781, 229, 299 and 143 hypo/hypermethylated DMRs that exhibited a methylation change of 30% ( $\Delta > 0.3$ ) for WNT, SHH, Group 3 and Group 4 subgroups respectively. The number of DMRs identified fell significantly when filtering for DMRs that exhibited greater differential methylation (Table 5.7). The characteristics of DMRs called varied between each subgroup of MB (Table 5.8). WNT specific DMRs were most numerous ( $n = 781$ ) and on average the widest, with a mean width of 832bp. Much lower numbers of subgroup specific DMRs were detected within SHH ( $n = 229$ ), Group 3 ( $n = 299$ ) and Group 4 ( $n = 143$ ). The longest DMR identified was present within the SHH subgroup, reaching 7,725bp. Compared to group 4, over twice the amount of specific DMRs were identified within group 3 and DMRs identified within this group spanned greater distances on average.



Subgroup/Subtype	Specific DMRs detected (% change in methylation)		
	30%	40%	50%
<b>WNT</b>	781	267	74
<b>SHH</b>	229	58	9
<b>Grp3</b>	299	43	17
<b>Grp4</b>	143	22	1

Table 5.7 – Total number of subgroup specific DMRs identified by DMRcate (Array).

Subgroup/Subtype	Platform	DMR characteristics		
		Mean width (bp)	Longest DMR (bp)	Most probes/CpGs
<b>WNT</b>	<b>Array</b>	832	6,630	44
	<b>WGBS</b>	931	9,995	269
<b>SHH</b>	<b>Array</b>	645	7,725	33
	<b>WGBS</b>	792	8,124	269
<b>Grp3</b>	<b>Array</b>	568	4,161	25
	<b>WGBS</b>	741	4,776	310
<b>Grp4</b>	<b>Array</b>	400	2,150	13
	<b>WGBS</b>	711	4,528	219

Table 5.8 – Basic statistics of subgroup specific DMRs identified by DMRcate (Array) and metilene (WGBS)

Subgroup/Subtype	Specific DMRs detected (% change in methylation)		
	50%	60%	70%
<b>WNT</b>	23,797	11,622	3,579
<b>SHH</b>	5,707	3,277	1,327
<b>Grp3</b>	7,777	4,316	1,717
<b>Grp4</b>	8,767	5,696	2,602

Table 5.9 – Total number of subgroup specific DMRs identified by metilene (WGBS)

Subgroup	Total No. Identified	Overlap with 450k manifest	Overlap with EPIC manifest
<b>WNT</b>	23,797	6974 (29.31%)	10160 (42.69%)
<b>SHH</b>	5,707	3142 (55.06%)	3762 (65.92%)
<b>Grp3</b>	7,777	3135 (40.31%)	4394 (56.5%)
<b>Grp4</b>	8,767	3035 (34.62%)	4387 (50.04%)

Table 5.10 – Total number of DMRs identified by metilene that overlap with at least one probe located on the Illumina Methylation450k or MethylationEPIC probe manifests.



Far greater numbers of DMRs were identified in our WGBS analysis compared to the array. When seeking DMRs that exhibited a mean methylation difference of 50% or greater (beta value difference of 0.5), 23,797, 5,707, 7,777 and 8,767 subgroup specific DMRs were identified for WNT, SHH, Group 3 and Group 4 respectively. DMRs that exhibited greater methylation differences of 60% and 70% remained well in excess of what was identified when analysing array data (Table 5.9). Compared to our array analysis, mean DMR width increased in all subgroups, as did DMR length. When investigating the degree to which DMRs identified in our WGBS analysis overlapped with DMRs identified via array, low rates of overlap were observed, with 45%, 52%, 15% and 43% for WNT, SHH, Group 3 and Group 4 respectively (Table 5.8).

Various technical explanations can be offered to explain the increased number of DMRs detected via WGBS compared to the array. The capability to analyse the methylome at single-site resolution increases the likelihood that regions of differential methylation can be detected – particularly smaller DMRs below 100bp and regions between or outside of the range of probe sites. To establish the degree to which DMRs identified in our WGBS were located within regions covered by the array probe manifest, we overlapped them with co-ordinates for probe loci from either the Illumina Methylation 450k or EPIC manifests (Table 5.10). The degree of overlap varied between subgroups, where 29%, 55%, 40% and 34% of WNT, SHH, Group 3 and Group 4 specific DMRs overlapped with at least one probe site on the 450k manifest. The rate of overlap increased when overlapping with the EPIC manifest, although not proportionally – where 43%, 66%, 57% and 50% of WNT, SHH, Group 3 and Group 4 specific DMRs were identified as overlapping. These large differences point to significant levels of subgroup specific differential methylation occurring within regions not measured by the array, but also raises additional questions as to why many DMRs detected via array are not also called via WGBS in the same group comparisons. The increased beta value polarisation observed in methylation data derived from WGBS (even more so when dealing with samples sequenced at lower depths) would likely result in any difference in methylation observed to also be larger. This was in part behind the reasoning to increase the methylation difference required in WGBS to 50% for a DMR to be called, and this difference may contribute to the low rates of overlap observed between both platforms. The inverse is also true, where a decreased range in beta value intensities produced by the array is also likely to result in a lower number of DMRs being identified. The fact that different software was used to call DMRs is also likely a contributing factor, although the degree to which this is affecting calls is difficult to determine and requires further study. The DMRcate package is now compatible with WGBS data, and time constraints prevented performing this comparison in this study. Regardless, our data demonstrates significant amounts of DMRs are being detected throughout the methylome within each subgroup of MB, particularly within intergenic regions that are not measured by the design of either the Illumina Methylation 450k or EPIC platforms.

## 5.4.6 – Subgroup-specific differential methylation is occurring throughout the methylome that is missed by the array platform

Genomic annotation of all DMRs called by both platforms revealed subgroup specific patterns in DMR distribution between each platform. Owing to the gene-centric design of the DNA methylation microarray, DMRs identified in our array analysis encompassed primarily coding regions (Table 5.11). Large increases in the number of DMRs overlapping with open sea regions were identified for all subgroups in our WGBS analysis (Table 5.12). The distribution of hypo/hypermethylation within SHH changed depending on the platform used (Figures 5.8 & 5.9). Where DMRs detected via array were predominantly hypomethylated, large numbers of hypermethylated DMRs were detected in our WGBS cohort. Whilst general trends in DMR distribution can be analysed from the data presented in this study, the degree to which annotations overlap must be considered. Due to the size of some DMRs regularly extending beyond 1000bp, multiple biologically relevant regions can be covered by a single DMR. For example, many open sea CpGs are within intronic regions, and the high levels of intronic DMRs detected in each subgroup also overlap with intergenic CpG regions. To provide an example, a paired annotation matrix detailing the degree of overlap observed in SHH subgroup specific DMRs is provided in the appendix – 8.1.7.

<b>Annotation Type</b>	<b>WNT</b>	<b>SHH</b>	<b>Group 3</b>	<b>Group 4</b>
CpG (Open Sea)	347	116	194	97
CpG Islands	354	73	47	22
CpG Shelves	63	31	43	18
CpG Shores	335	83	77	33
FANTOM Enhancer Regions	116	26	24	10
3UTR's	66	12	23	8
5UTR's	143	34	36	15
Exons	314	80	73	39
Intergenic	231	76	91	46
Introns	429	122	153	75
Promoters	205	40	62	23

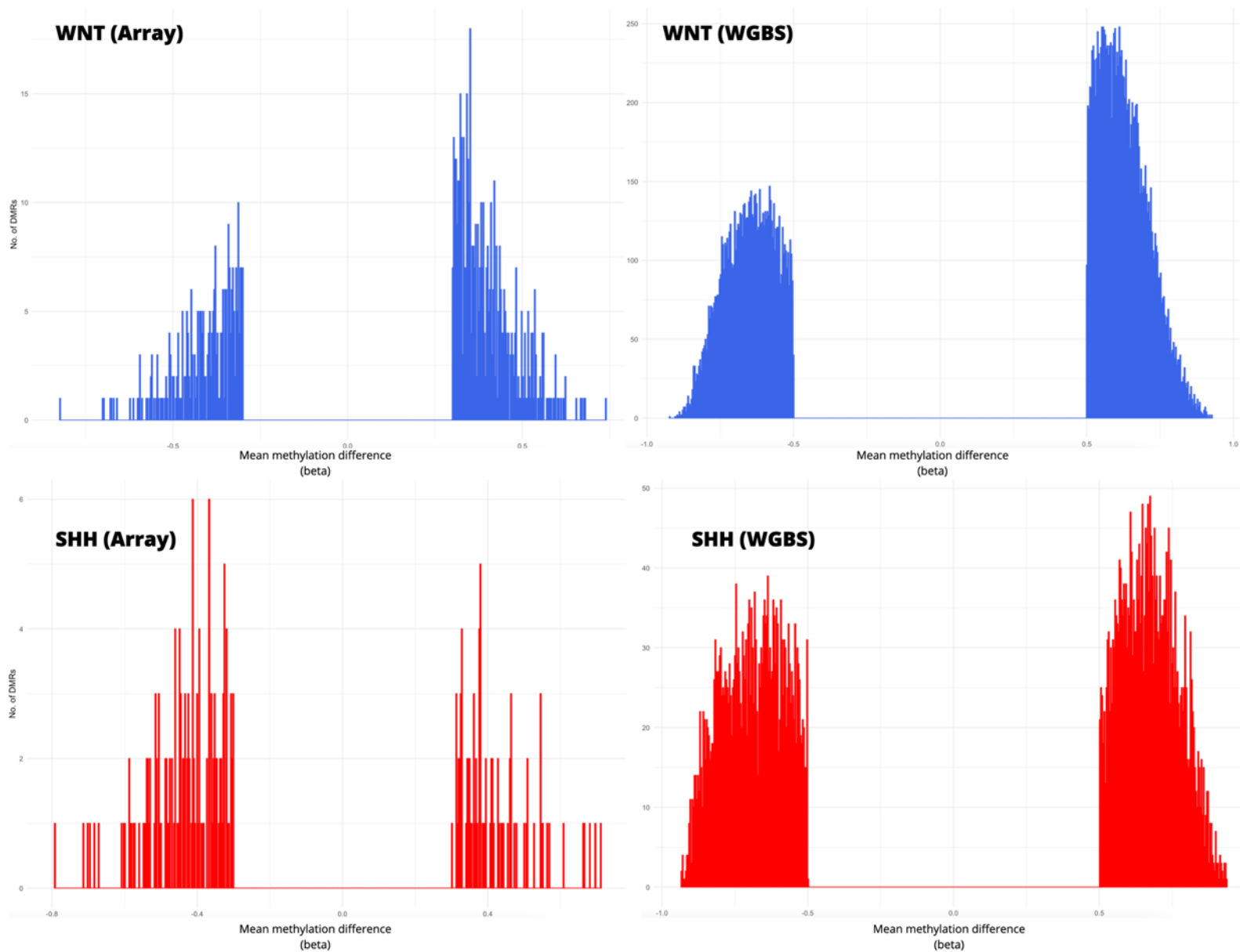
**Table 5.11** – Array subgroup-specific DMR distribution organized by annotation type

<b>Annotation Type</b>	<b>WNT</b>	<b>SHH</b>	<b>Group 3</b>	<b>Group 4</b>
CpG (Open Sea)	18,500	2,990	5,317	6,343
CpG Islands	2,520	1,654	989	732
CpG Shelves	1,583	432	704	918
CpG Shores	3,973	2,139	1,903	1,827
FANTOM Enhancer Regions	1,637	588	622	622
3UTR's	851	387	365	326
5UTR's	1,150	491	501	341
Exons	3,931	1,557	1,694	1,416
Intergenic	10,623	1,918	2,555	3,652
Introns	11,650	3,112	4,550	4,352
Promoters	1,633	725	756	581

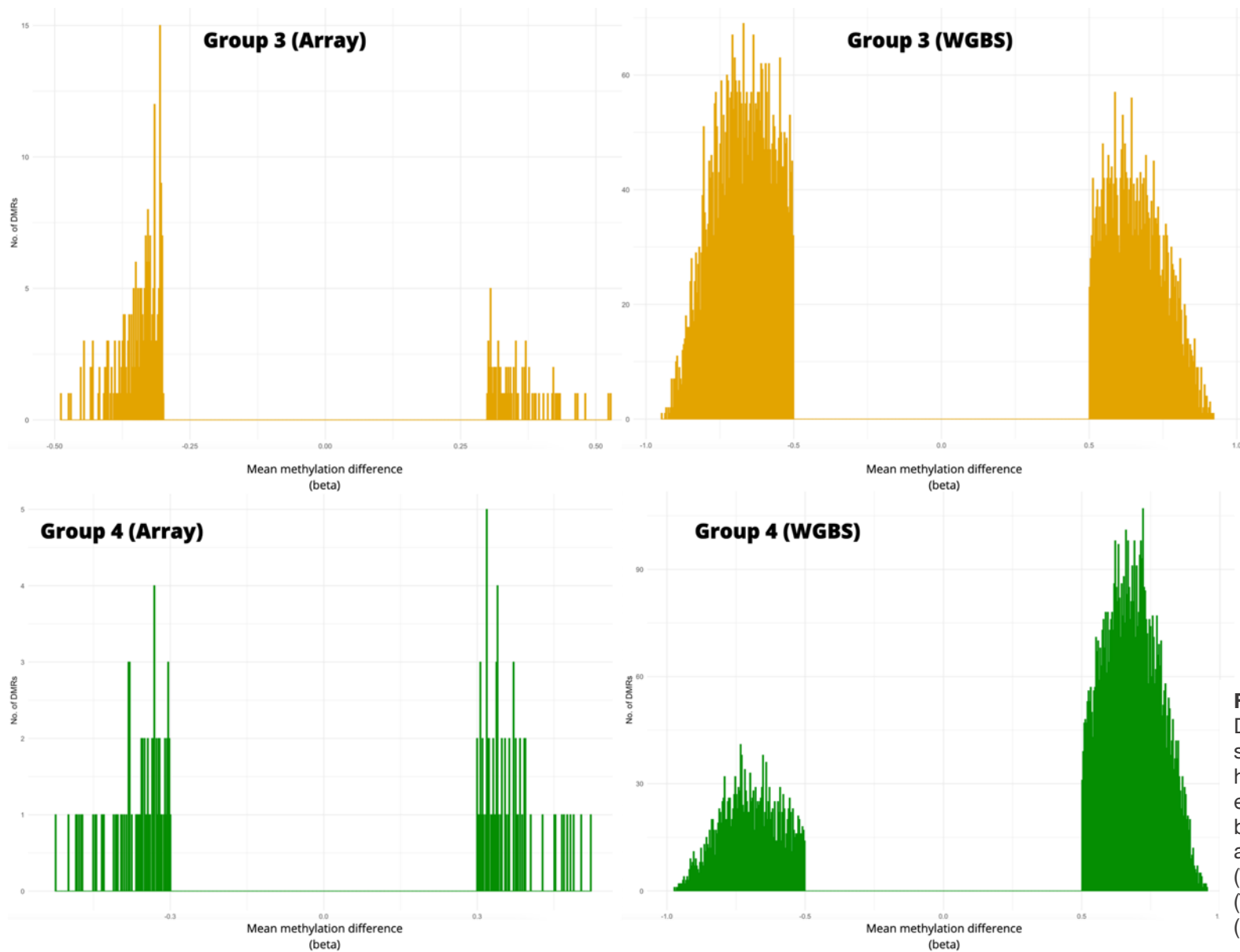
**Table 5.12** – WGBS subgroup-specific DMR distribution organized by annotation type.

The biologically related groups 3 and 4 shared similar DMR distributions, where both groups were characterised by significant numbers of DMRs within open sea regions. We observed differences in the pattern of differential methylation within each subgroup, where group 3 specific DMRs were predominantly hypermethylated in our array analysis, and more hypomethylated DMRs identified in our WGBS analysis (Figures 5.8 & 5.9). Group 4 DMRs exhibited higher levels of hypermethylation in all DMRs compared to group 3 – a trend that became more pronounced when considering DMRs within the whole methylome.

Despite the overlap, what is clear from this data is that significant numbers of subgroup specific DMRs within open sea regions are present within medulloblastoma that are being missed by the array platform. These regions could play a key role in the establishment of comprehensive subgroup profiles within medulloblastoma with potential biological significance. This is particularly relevant for group 4, where the large number of novel DMRs identified in intergenic and open sea regions may hold the key to further establishing the biology of the subgroup, where chromosomal instability and aberrant genetic imprinting are recognised as key traits (Northcott *et al.*, 2012).



**Figure 5.8.** Distribution of subgroup specific hyper/hypomethylated DMRs detected by DMRcate (Array) and metilene (WGBS) in WNT (Top) and SHH (Bottom)



**Figure 5.9.** Distribution of subgroup specific hyper/hypomethylated DMRs detected by DMRcate (Array) and methylene (WGBS) in Group 3 (Top) and Group 4 (Bottom)

### **5.4.7 – Ontology analysis of genes affected by differential methylation identifies greater numbers of enriched subgroup-specific biological processes in WGBS compared to DNA methylation microarray**

For DMRs that overlapped with genes, a gene ontology (GO) enrichment analysis was performed to identify subgroup-specific biological processes that were over-represented. Vastly increased numbers of over-represented biological processes were identified by our WGBS analysis compared to array (see appendix – 8.1.7). WNT enriched terms were the most numerous across both platforms, reflective of its distinct biology and greater number of DMRs detected relative to other subgroups. 13 pathways, including nervous (adjusted p-value = < 0.001, combined score = 112.196) and circulatory system development (adjusted p-value = 0.021, combined score = 40.981) were enriched in gene lists identified in our array analysis. Far greater numbers of enriched pathways were identified in our WGBS analysis, where 61 pathways were identified as being over-represented. Many additional biological processes were identified as being significant, including neuron differentiation (adjusted p-value = 0.003, combined score = 26.226) and WNT signalling processes (adjusted p-value = 0.004, combined score = 27.727). SHH enriched terms occurred in lower numbers, with 2 and 11 enriched processes identified by our Array and WGBS analyses respectively. Lower numbers of enriched terms were identified in groups 3 and 4 (Group 3 = 2, Group 4 = 1) by array, increasing again when analysing gene sets derived from our WGBS analysis. 11 biological processes were identified in group 3, including regulation of cell migration (adjusted p-value = 0.021, combined score = 19.095) – a pathway implicated in tumour metastasis, which is also a characteristic trait of the subgroup. 49 pathways were identified in our group 4 analysis, including microtubule binding (adjusted p-value = 0.033, combined score = 31.777) – an important mechanism that can contribute to chromosomal instability.

The large number of enriched terms identified as significant in our WGBS analysis was likely due to the increased number of DMRs detected using the platform. In our data, this has led to significantly more genes being identified, leading to more terms being identified as enriched. The small cohort size may also affect the power of our findings, but both tools have been shown to handle small group comparisons well by their respective authors (Juhling *et al.*, 2015; Peters *et al.*, 2015). However, increased sample numbers would have enabled us to detect DMRs specific to molecular subtypes rather than subgroups, providing an estimation that is more reflective of the current consensus of the molecular landscape of MB. Whether or not the enriched terms identified are biologically linked to tumorigenesis within each subgroup is beyond the scope of this study but is cause for further investigation given the developmental and neurological nature of most pathways found to be enriched.

### **5.4.8 – NMF-directed splitting of subtype VII is supported by differential methylation analysis**

The deliberate selection of large numbers of subtype VII samples in the sequencing cohort allowed the investigation of the DMR landscape between the two putative subtype VII variants (Sharma *et al.*, 2019). The array analysis identified just one hypermethylated DMR affecting a CpG island within the PPT2 gene. Consistent with the subgroup-directed analysis, significantly greater numbers were identified in the WGBS analysis of subtype VII. 5,981 DMRs were identified, with a total of 84 biological processes identified as enriched following gene ontology analysis (Appendix 8.1.7). Highest scoring terms included mesenchymal stem cell differentiation (adjusted p-value = 0.013, Combined score = 26.127), mesonephros development (adjusted p-value = 0.009, Combined score = 115.167) and neuron fate specification (adjusted p-value = 0.028, Combined score = 134.918). Like the group 4 comparison, large numbers of DMRs overlapped with open sea CpGs, and VII\_1 exhibits increased levels of hypermethylation compared to VII\_2. These findings are conflicting with that of Sharma *et al.*, that identified hypomethylation among the most variable loci between the two groups (Sharma *et al.*, 2019).

The molecular subtypes of medulloblastoma have been established relatively recently. Beyond mutation and aneuploidy associative analyses, the methylome landscape has not yet been established. Clinical support for the split of subtype VII is largely restricted to age, of which the distribution of subtype VII is bimodal (Sharma *et al.*, 2019). The increased hypermethylation observed in VII\_1 may in part be explained by the increased age of patients that fall within this category, where global demethylation occurs as cells age (Johnson *et al.*, 2012). When considering the capability of subtypes to be classified with high probability using DNA methylation array, the detection of high numbers of DMRs within subtype VII across the entire methylome is not surprising, and further research is needed to understand the biological implications that may be associated with the high levels of differential methylation observed within this subtype.

## 5.5 – Summary & conclusion

The data presented in this chapter effectively highlights both the limitations and potential for using low-pass WGBS to obtain clinically relevant readouts in medulloblastoma. Parameter optimisation and the addition of telomeric, centromeric and other recognised regions of low read mappability throughout the genome to the existing filter resulted in highly effective CNV detection in the low-pass WGBS MB cohort, without a matched set of control samples for reference normalisation. Throughout the optimisation process, it became clear that called segments were generally low in quality when using smaller window sizes, characterised by high standard deviation statistics called for all segments ( $SD = 0.28$ ). To overcome this, bin size was increased to 50kb, which resulted in a 58% reduction in segment variability ( $SD = 0.12$ ). This translated into high quality segmentation, which has the capacity to identify whole and single-arm regions of aneuploidy with 96.908% sensitivity compared to those called by conumee in our matched array cohort. In addition, a further 23 additional whole chromosomal aberrations were identified in our WGBS cohort, which upon inspection appeared to fall short of being identified in our array analysis. Whilst minimum window size for both pipelines are identical, WGBS offers a more robust and superior means of copy number segmentation, where bins are present throughout the genome and not spaced unevenly at pre-determined probe sites of interest. The optimised CNVkit pipeline is capable of reliably identifying focal regions of aneuploidy, including amplifications and deletions as small as 150kb-200kb. Validated amplifications of MYC and MYCN were identified with 100% and 66% sensitivity, outperforming the array cohort. Additional focal regions of amplification/deletion were identified compared to the array including SNCAIP within group 4, although our limited validation set, and overall cohort size requires further study to confirm initial findings. Sequencing platforms such as WGS and WES can be used to identify focal amplifications across the genome but do not offer the capability to analyse methylation simultaneously. The demonstrable high performance compared to current array methodologies achieved in this study place WGBS in a uniquely advantageous position in deriving accurate large and focal copy number estimates throughout the whole genome, where the optimised pipeline offers a means to achieve this without the need for reference samples, even at low pass.

To investigate the capability of low-pass WGBS in calling clinically important mutations in MB, the BS-seq specialised variant calling software was applied to 4 pilot samples. The results generated were poor, with the depth of all passing variants too low to be called reliably using established variant calling filters. Calling variants correctly and reliably either requires large cohorts of samples from which to pool coverage from or higher sequencing depth – both of which were beyond the means of this study. Identifying variants is a crucial component of medulloblastoma diagnostics and the current means of identifying them via targeted sequencing panels remains a superior method of achieving this compared to low-pass WGBS. Further research is therefore required to establish a minimum level of coverage suitable for reliable mutation detection using WGBS.



An initial survey of the landscape of differential methylation of the medulloblastoma methylome was performed in a subgroup-specific manner, demonstrating that highly increased numbers of DMRs are identified using WGBS than are currently detected using DNA methylation microarray. DMRs are called at higher resolution with greater differential change, enabling DMRs to be detected that exhibit significant changes in methylation compared to the array. These findings highlight that significant differential methylation is occurring in regions not measured by the array platform. These findings were corroborated when annotating subgroup specific DMRs with genomic location. The genomic distribution and nature of hypo/hypermethylation among DMRs was subgroup-specific, with the greatest number of DMRs identified overlapping with open sea CpGs than any other region. These distributions differ to the array, where DMRs are inherently identified in genic regions, due to the deliberate gene-centric design of the platform. Differential methylation between groups 3 and 4 became more pronounced when considering the DMR landscape of the entire methylome, where subgroup specific DMRs were identified in large numbers and are concentrated among intergenic and open sea regions. Given the biological implications of intergenic methylation and chromosomal instability, this is an interesting finding that provides motivation for the establishment of whole-methylome profiles for subgroups/subtypes of medulloblastoma. Ideally, this would be conducted in a subtype specific manner if cohort size is large enough. It was demonstrated that differences in differential methylation exist throughout the methylome within subtype VII, for which there was evidence for a further split based upon clinical and clustering characteristics when analysing array data (Sharma *et al.*, 2019). 5,981 DMRs were detected between the NMF-directed subsets of subtype VII, with over 84 biological processes identified as enriched following GO analysis. Beyond associations with aneuploidy and variants for each subtype, the methylation landscape of subtypes has not been comprehensively established, and findings suggest that characterising the whole methylome may help further refine the subtype landscape of medulloblastoma. Overall, it was demonstrated that WGBS can be employed to derive superior aneuploidy and DMR detection compared to the array platform. The increased variability located in intergenic regions and increased coverage of all regions of the genome enable events from both analyses to be detected in greater numbers with higher resolution. Importantly, the identification of these events is a crucial component of medulloblastoma research and diagnostics and its demonstrable superiority at low pass will likely increase the appeal of the platform over current microarray platforms going forward given its likely compatibility with array-based classifier models.

## **Chapter 6 – Summary & Discussion**

---

## 6.1 - Introduction

Medulloblastoma (MB) is a grade IV embryonal brain tumour that occurs in the posterior fossa of the cerebellum (Louis *et al.*, 2021). It is the most common malignant brain neoplasm in children, accounting for 65% of embryonal diagnoses (Ostrom *et al.*, 2020). Current treatment strategies for MB comprise surgical resection followed by chemotherapy and craniospinal radiation. Metastasis is a common, identified in 20-30% of cases upon presentation (Zapotocky *et al.*, 2017). Relapse is common, occurring in 30% of patients, at which stage it is near universally fatal (Hill *et al.*, 2020). Survival in standard risk patients (>3 years, gross total resection, non-metastatic) is high, with 5-year survival reaching between 70 and 85% (Gajjar *et al.*, 2006; Oyharcabal-Bourden *et al.*, 2005). High risk patients (< 3 years, sub-total resection and/or metastatic) understandably perform worse, where 5-year survival rates of less than 70% are reported (Jackaki *et al.*, 2012; Gandola *et al.*, 2009). Extensive study of medulloblastoma involving many 100s of samples led to the identification significant heterogeneity underlying the disease. Methylation profiling of MB patients led to the identification of four molecular subgroups; WNT, SHH, Group 3 and Group 4 – based upon distinct methylation, molecular and clinical profiles (Northcott *et al.*, 2011; Taylor *et al.*, 2012). Since 2016, more recent studies with larger sample sets have identified additional heterogeneity within subgroups, with subtypes now identified within SHH, Group 3 and Group 4. The distinct interplay between clinical outcome and molecular subgroup culminated in their recognition by the WHO in 2016, and subtypes are also now recognised to likely play a role in future classifications (Louis *et al.*, 2021).

DNA methylation analysis lies at the heart of subgroup/subtype determination in medulloblastoma and other paediatric brain tumours. It plays a key role in paediatric brain tumour diagnostics, and data from 1000s of samples was used to develop the DFKZ paediatric brain tumour classifier (Capper *et al.*, 2018). The model is built upon the DNA methylation microarray platform, a cost-effective assay that interrogates ~850,000 CpGs, including 99% of genes and 95% of CpG islands (Pidsley *et al.*, 2016). Combining the array with targeted gene sequencing panels enables the identification of most clinical features associated with medulloblastoma (i.e., aneuploidy, mutation, subgroup), and has helped to elevate the array platform to its current 'gold-standard' status across the field of paediatric neuro-oncology. While its diagnostic utility is clear, the array is prone to variable turnaround times due to the specialist laboratory infrastructure required and costs can spiral when not submitting samples in batches (Perez & Capper, 2020). It is also proprietary, presenting the risk of being discontinued. This has happened twice over the course of the platform's lifetime and its creator Illumina now offer a Methyl Capture sequencing kit that offers a superior alternative (Illumina, 2022). Methylation sequencing offers a means of interrogating the methylome at single-base resolution and whole genome bisulfite sequencing (WGBS) offers the potential to interrogate nearly all ~28.3 million CpGs within the human genome on a per-sample basis (Frommer *et al.*, 1992). Despite its conception in the 1990s, adoption of the technique has been limited due to high sample input requirements and economic cost. It has not been until recently that new methodologies have been developed that have

enabled the platform to become a feasible alternative to the array, with the costs of low pass sequencing now reaching comparable levels to the array platform (in 2021: 10x WGBS = ~£430 / Array = £300).

Like the array, WGBS generates methylation data in the form of beta values, allowing it to be used analogously in many analytical pipelines. Theoretically this could apply to classification, where WGBS sample data could be applied to array trained classifiers for subgroup prediction. The platform also offers a superior alternative when concerned with other analyses such as differential methylation and aneuploidy detection, where the single-site resolution and whole genome coverage offered by WGBS can potentially be leveraged to call regions of differential methylation and chromosomal copy number variation with increased resolution. This is incredibly useful for identifying focal aneuploidy or differential methylation events within genes and intergenic regions which are not covered sufficiently the array. WGBS offers the possibility to unite both NGS and methylation analysis within a single platform, offering a robust, independent means of reaching a definitive subgroup diagnosis in medulloblastoma. The platform represents an untapped potential for methylation analysis within MB, where a wealth of extra data of potential biological significance has yet to be comprehensively investigated throughout the medulloblastoma methylome.

Given the drawbacks of the array, increasing economic appeal and superiority of the WGBS platform, the use of methylation sequencing is likely to become more prominent over the next decade. In addition, genome sequencing will become more integrated within diagnostic laboratories of developed nations in the coming years, where methylation sequencing could theoretically play a part in medulloblastoma diagnostics (Genomics England, 2020). For WGBS to be adopted within paediatric oncology, the development of a robust processing methodology suitable for clinical analysis of samples sequenced using WGBS must be established. Whilst some pipelines exist, none are specialised to deal with low-pass data – which for now at least is the most attractive economic option. However, working with data sequenced at low pass presents many issues regarding data missingness that render it unsuitable unless addressed. This project set out to investigate the capability of low-pass WGBS in being used for the clinical analysis of medulloblastoma, and to ultimately determine whether robust molecular subgrouping is a feasible proposition for clinical samples sequenced using the platform. In chapter 3, a robust, optimised processing methodology for clinical samples analysed using low-pass WGBS is presented (Section 3.4.1). It was demonstrated that machine-learning based imputation methodologies can be used to overcome the issue of data missingness associated with sequencing at low-pass (Zou *et al.*, 2018), allowing popular subgrouping analyses like clustering and classification to be performed downstream without issue (Section 3.4.2). Correlation analysis of imputed sample data to matched array samples show high correlation present between CpGs and respective probe sites located on the array platform (section 3.4.3). A comprehensive analysis of the effect of lowering coverage on the performance of each analytical step of our optimised pipeline is

presented, providing valuable insights into the quality of sample data returned as sequencing depth decreases.

In chapter 4, low-pass WGBS sample data generated from chapter 3 was used to investigate its capability to be utilised to detect the molecular subgroups of medulloblastoma. Low-pass WGBS methylation data was integrated with and applied to existing array-based classifiers, demonstrating that robust, accurate subgroups can be assigned to each sample, comparable to probabilities acquired from matched array data (Section 4.4.3). CpG variability was interrogated throughout the whole methylome compared to that observed at probe sites within the array, revealing that significant variability is present among the molecular subgroups of medulloblastoma that is currently missed entirely by array platforms (Section 4.4.1). Visualisation of the most variably methylated sites for both platforms revealed increased separation present among subgroups for the samples tested in our cohort. WGBS data was clustered with non-matched medulloblastoma samples that form part of the current DFKZ reference cohort, showing that subgroup clustering patterns are maintained irrespective of platform or coverage (Section 4.4.2). Finally, an overview of molecular subgroup classifier performance was presented for models trained solely using WGBS data compared to like-for-like models trained using matched array data (Section 4.4.4). Validation performance was equal for most models tested, and performance remained stable until reaching coverages as low as 2x.

In chapter 5, an investigation into the capability of low-pass WGBS to be used to generate clinically relevant readouts that are useful for medulloblastoma diagnostics was presented. An optimised, robust pipeline capable of calling regions of aneuploidy with high sensitivity was presented compared to current established array methodologies, without a reference sample set (Section 5.4.2). The increased coverage afforded by WGBS enabled additional focal regions of aneuploidy to be identified, and that validated amplifications of MYC and MYCN can be identified with superior sensitivity compared to the array platform (Section 5.4.3). We applied variant calling to low-pass WGBS samples, demonstrating that reliable mutations cannot be identified at low-pass (section 5.4.4). Finally, we provided a comprehensive analysis of DMR distribution and gene ontology enrichment using WGBS methodologies compared to the array, providing insights into the degree of subgroup-specific methylation patterns present within MB throughout the genome (Sections 5.4.5 - 5.4.8).

## 6.2 – High quality, non-missing methylation data can be generated for clinical samples sequenced via low-pass whole genome bisulfite sequencing that is highly correlated to matched array sample data

For low-pass methodologies to be adopted, they must be capable of generating high quality data that can be used effectively in downstream analyses. The primary issue concerning low pass sequencing data is data missingness, which significantly hampers the ability to glean useful biological data from the platform. When considering WGBS, the types of data generated are two-fold – aligned reads and methylation ratios for all CpGs. Whilst sequencing analyses like aneuploidy detection have shown demonstrable success at low pass in WGS and WES by pooling reads into bins (Raman *et al.*, 2019), methylation data generated at low pass presents significant issues. Analyses integral to subgrouping studies like classification and clustering do not tolerate missing data, and for WGBS to be integrated into existing classification tools, methylation data must be available for individual CpGs. Data missingness compounds as cohort size increases, where the total number of common CpGs continues to fall significantly. Missing data handling for methylation microarray platforms forms a straightforward component of processing, with a wide range of optimised imputation algorithms available like k-nearest neighbours, random forest, and linear regression. Most approaches are unfeasible when concerned with WGBS data, where the computational pressures of loading very large matrices encompassing tens of millions of CpGs per sample become too high for most computing clusters. Current paediatric oncology research that utilises bisulfite sequencing has avoided the issue entirely, avoiding imputation and instead drastically subsetting sequenced data down to common CpGs (Kuschel *et al.*, 2021; van Paemel *et al.*, 2020).

For this project, we sequenced a representative cohort of 36 primary medulloblastoma samples. In addition, we acquired an external cohort of 34 primary medulloblastoma and 8 control cerebellum samples, publicly available in a pre-processed format (beta values) (Hovestadt *et al.*, 2014). Pre-processing of sequencing data produced high quality sequencing reads, generating an average paired-end sequencing yield of 217,290,129 reads for all samples. Alignment rates very high, averaging 78.75% across all samples, enabling us to reach an average depth of 7.18x following across all samples following methylation data extraction. Externally acquired samples did not suffer from this issue, as data was acquired in a pre-processed format and did not require read filtering (Mean depth = 29.48x). Analysis of extracted methylation data demonstrated key differences in beta values calculated using WGBS sequencing than are via the array. Beta values are increasingly bimodally distributed due to the read-count based calculation method in which beta values are calculated. This effect was more pronounced in lower coverage data, where the number of reads to calculate from decreased. The increased bimodal distribution observed would likely impact methylation variability, a key component of feature selection in subgrouping studies (Northcott *et al.*, 2017; Capper *et al.*, 2018; Sharma *et al.*, 2019).

The issue of data missingness was made apparent following processing of methylation data. Applying the currently recommended filter of 5x to our cohort resulted in just 69,101 common CpGs across all assessed samples, with an average 11,523,459 CpGs missing per sample. This led to the adjusting of the depth filter to 3x, judged to be a necessary step to recover as much methylation data as possible – over 60% of missing CpGs were recovered. Following filtering, we imputed missing methylation values using BoostMe, a machine learning based imputation algorithm built upon the XGBoost algorithm (Zou *et al.*, 2018). Imputation model performance was excellent, with low RMSE (0.15, SD = 0.03), high accuracy (0.93, SD = 0.03), high AUROC (0.97, SD = 0.02) and high AUPRC (0.99, SD = 0.01). Model performance significantly fell with each drop in coverage, with a more pronounced drop in coverage from 4x onwards. Overall, BoostMe was found to be highly amenable to the data generated in this study, where performance has exceeded that achieved by other algorithms dealing with lower rates of missing data in other studies (Zou *et al.*, 2018). A cross-platform correlation analysis between methylation values produced by our processing pipeline and beta values for matched samples analysed via DNA methylation microarray at CpG sites that correspond to probe sites using LiftOver (Titus *et al.*, 2016) was performed. Overall, inter-platform correlation for imputed data was significantly lower (determined via paired t-test) than non-imputed data (PCC = 0.95, SD = 0.014,  $p = 0.02$ ), but recovered significantly when using data filtered using a 3x depth filter (PCC = 0.93, SD = 0.02,  $p < 0.001$ ) rather than 5x (PCC = 0.90, SD = 0.02,  $p < 0.001$ ), emphasising the negative affect that imputing far greater numbers of missing values presents when applying a high coverage filter to low pass data. Regardless, correlation remains high and is consistent with other studies that employ WGBS at 10x depth and above (Titus *et al.*, 2016).

In summary, an optimised pre-processing pipeline capable of generating high quality, non-missing methylation data obtained from clinical tumour samples sequenced using low-pass WGBS is presented. Both sequence and methylation data are high quality, and imputation via BoostMe produces data that correlates well to matched array sample data. Going forward, the inclusion of an imputation strategy is deemed to be an absolute necessity if one wishes to comprehensively analyse bisulfite sequencing data at low pass. Its inclusion allows for the full potential of low-pass WGBS to be realised, making tens of millions of CpGs available for downstream analyses.

### 6.3 – Medulloblastoma sample data sequenced via low-pass WGBS can seamlessly be integrated into pre-existing array trained classification models

DNA methylation is a robust measurement that has shown success in helping researchers define new tumour entities and classes. Methylation analysis has relied upon the DNA methylation array platform, where it has remained the platform of choice for over a decade. This stagnation ultimately resulted in an unintended benefit – the accumulation of many thousands of samples analysed using the same platform, generating large reference sets numbering in the thousands. In paediatric oncology this has culminated in the development of the DFKZ CNS tumour classifier, capable of assigning tumour samples into 81 different tumour types based upon methylation patterns (Capper *et al.*, 2018). This development remains ongoing, and the abundance of sample data and clear utility guarantees continued development in the coming years. With methylation sequencing techniques becoming more prominent, it is vital that methylation data obtained from these platforms can be integrated with array-based classifiers to take advantage of more advanced technologies whilst not forfeiting the classification capabilities achieved by analysing many thousands of array samples. Provided the software is publicly available, processed beta values derived from WGBS samples can be applied to array trained models, enabling classification calls to be assigned to any sample of interest.

To test the ability of array trained models in predicting the molecular subgroup of medulloblastoma samples sequenced via low-pass WGBS, we applied WGBS sample beta values to two pre-existing support vector machine classifiers capable of predicting both the four classic molecular subgroups of medulloblastoma (WNT, SHH, Group 3, Group 4) and a previous 7-group classification (WNT, SHH-Child, SHH-Infant, Group 3 High Risk, Group 3 Low Risk, Group 4 High Risk, Group 4 Low Risk) (Newcastle University, 2021; Schwalbe *et al.*, 2017). High probabilities of assignment in both internal (NMB) and external (ICGC) cohorts was observed. All 36 samples were correctly assigned into subgroup when applied to the 4-group classifier (Mean probability = 0.967, SD = 0.057). Surprisingly, matched array probabilities were significantly lower (Mean probability = 0.949, SD = 0.104, paired t-test:  $p = 0.031$ ). All 34 primary MB samples in the ICGC cohort were also correctly classified (Mean probability = 0.970, SD = 0.057,  $p > 0.05$ ), with no significant difference observed between both cohorts and to matched array samples (Mean probability = 0.964, SD = 0.071,  $p > 0.05$ ), indicating that the effect of sequencing coverage was negligible in our cohort. 7-group classification of WGBS samples was slightly poorer, with NMB samples classing with equal probability to matched array data (WGBS: Mean probability = 0.904, SD = 0.133 / Array: Mean probability = 0.913, SD = 0.113,  $p > 0.05$ ). A total of 33 out of 36 samples were correctly classified, with both misclassifications belonging to the SHH subgroup. ICGC samples also performed equally (WGBS: Mean probability = 0.873, SD = 0.150 / Array: Mean probability = 0.870, SD = 0.150), with no significant difference in probability observed between cohorts or platforms.



These results demonstrated that medulloblastoma tumour samples sequenced via WGBS at low pass can seamlessly be integrated into pre-existing classifier models to predict subgroup with excellent performance. Concordance is very high, indicating that similar models like the current DFKZ brain tumour classifier would also be highly amenable to WGBS sample data if software was publicly available to researchers. Whilst this seems like a simple step, it presents some risk if proper care isn't taking during sample processing. A lack of imputation or incorrect processing may result in incorrect classification, putting forward a case for the development of standardised processing guidelines for WGBS samples prior to their use with the classifier of interest. Previous studies have attempted to calibrate array classifier training sets to be more suited to WGBS methylation data by 'binarizing' beta values – effectively rounding values to 1 and 0 to mimic the bimodal distribution of methylation sequencing data (van Paemel *et al.*, 2021; Kuschel *et al.*, 2021). Whether or not this has had much of an effect is up for debate given our findings and such an approach ignores hemi-methylation, which is known to be an important biomarker in tumorigenesis (Sun *et al.*, 2021; Shao *et al.*, 2009). Overall, our findings demonstrate that WGBS samples sequenced at low pass that are processed using the methodologies in this study can be classified with high probability, an important step if the field is to ensure a smooth transition from the array platform to methylation sequencing in the coming years.

## 6.4 – CpGs of greater variability are present throughout the medulloblastoma methylome that contribute to increased visual separation of medulloblastoma subgroups

Establishing CpG variance across all samples within a cohort is a crucial component of subgroup discovery and classification studies involving methylation data. Filtering non-variable CpGs helps reduce data dimensionality, easing computational pressure whilst retaining information which captures the most variability. In subgrouping studies, these most variable features are focused upon, as it is assumed that the most variable sites capture the separation observed between sample groups. Typically, standard deviation filtering is applied, where it has been shown to be especially effective when dealing with methylation data (Zhuang, Wedschwendter & Teschendorff, 2012). Variance filtering of CpGs/probes in our combined WGBS and array medulloblastoma cohorts revealed multiple differences between platforms. Over 3,990,549 CpGs were identified to be highly variable ( $SD > 0.25$ ) compared to just 27,211 probes. This variability became more pronounced when filtering for the top 10,000 features (WGBS: 0.40 – 0.46 / Array: 0.17-0.43). Whilst an interesting finding, technical differences between the platform help to explain the significantly increased variability that is captured by methylation sequencing. Beta values are polarised towards 0 and 1, whereas probe-based intensities typically only reach 0.1-0.9, limiting any standard deviation value that can be calculated at any locus. This effect likely becomes more pronounced when dealing with methylation sequencing data obtained at low pass, where sequencing errors cause larger changes methylation values. Despite these technical differences, such an increase in variability could not solely be caused by technical differences, but more likely by the fact that all CpGs are being interrogated. The EPIC platform covers just 3% of all CpGs in the genome (Zou *et al.*, 2018), with each probe relying on the assumption that neighbouring CpGs exhibit correlated methylation (Eckhardt *et al.*, 2006). Overlapping the top 10,000 most variable CpGs identified within our WGBS cohort with all probes located on the Illumina Methylation EPIC manifest revealed that just 439 (4.4%) were overlapping. When accounting for a probe width of 50bp, just 1270 (12.7%) were overlapping and when increasing our range further to 250bp (at which point multiple probes begin to overlap) 4,214 (42.1%) CpGs were identified as overlapping.

The manner that most variably methylated CpGs were distributed throughout chromosomes was investigated. Given the fact that array probes are unevenly distributed by design, it was anticipated that CpGs would be distributed proportional to chromosome size. This was not the case. Whilst chromosome 1 captured the greatest number of variable probe sites by the array, significantly less CpGs were identified within this chromosome (Array: 1,007 / WGBS: 938, two proportion z-test = 0.024). Instead, significant numbers of variable CpGs were present within chromosome 5 and over 10% were present within chromosome 7 when narrowing down to the top 1,000 most variable CpGs. While we cannot comment further on why we see uneven distributions of variable methylation within

the medulloblastoma methylome, it is clear that substantial levels of variable methylation is occurring in regions missed by the array.

Finally, the molecular subgroups of medulloblastoma were visualised using PCA, t-sne and UMAP for the top 10,000 CpGs/probes identified in our WGBS and array cohorts. Substructures representing all four molecular subgroups could be identified with distinct separation in both cohorts, with each respective subgroup appearing to cluster more densely and with greater separation in our WGBS cohort. This was an encouraging finding, given that our cohort consisted of two separate datasets sequenced in different laboratories and processed with different methods (Hovestadt *et al.*, 2014). Like that of increased variability, technical reasons that contribute to increased variance likely contribute to the increased separation observed. In addition, samples were deliberately selected that were classified with high probability, which is also a significant contributory factor. However, the fact that groups 3 and 4 cluster with such a clear degree of separation when accounting for variability located throughout the methylome is cause for further investigation given the constant motivation to further define groups 3 and 4 which share similar methylation profiles defined via array. The WGBS cohort was integrated with the medulloblastoma reference cohort that formed part of initial DFKZ paediatric brain tumour classifier cohort (n = 426) (Capper *et al.*, 2018). Data from both platforms clustered well, with subgroups persisting and WGBS samples sitting around cluster fringes of each respective subgroup cluster. The data from chapter 4 demonstrates clear inter-operability between methylation data derived from both platforms, and in part explains the ability for WGBS sample data to be classified confidently into subgroup using array trained models.

---

## 6.5 – Classifiers trained using low pass WGBS samples to predict the molecular subgroups of medulloblastoma perform equally to array-trained models

Methylation sequencing platforms for use in methylation diagnostics has seen increasing prominence in recent years. Both RRBS and low pass nanopore sequencing data has been used alongside array datasets for classifier training with success, even without imputation. Given the high correlation observed between WGBS and array, the increased variability observed throughout the methylome, and more distinct visual clusters observed in WGBS, classifier models trained using WGBS derived feature sets in predicting the molecular subgroups of medulloblastoma was investigated. For both platforms, seven prominent multi-class algorithms were trained; elastic net, random forest, XGBoost (linear & tree boosting), support vector machine (linear & radial kernel) and logic regression, to predict the molecular subgroups of medulloblastoma using our combined cohort of medulloblastoma samples. For each model, the top 10,000 most variable CpGs/probes were used for training and performance was determined using a 10 times 6-fold cross validation strategy.

Validation performance of most models was excellent, with comparable performance observed irrespective of the platform or training sets used. Our logic regression model was the top performer – an adaptive regression algorithm that is highly amenable to binary predictors (e.g., 0 and 1) (Ruczinski *et al.*, 2012). We observed high mean accuracy (WGBS = 0.977, array = 0.969,  $p = 0.178$ ), CK (WGBS = 0.969, array = 0.959,  $p = 0.173$ ), AUC (WGBS = 1.000, array = 1.000,  $p = 1.000$ ) and low log loss (WGSB = 0.068, array = 0.070,  $p = 0.009$ ) between both platforms, with no fall in performance observed in downsampled datasets like that seen in other algorithms that were tested. The use of logic regression in methylation classification has scantily been tested and these findings support further testing of logic regression models in WGBS classification models. However, it must be acknowledged that this analysis was only exploratory, with low sample numbers and a lack of test set available for more comprehensive performance testing. Feature selection was not specialised for each algorithm, which may hold back the performance of some algorithms like random forest and support vector machine models (Guyon *et al.*, 2002; Capper *et al.*, 2018). However, our initial findings are encouraging and indicate that WGBS trained models will likely perform equally to that of array models once WGBS sample data has accumulated.

---

## 6.6 – Low-pass WGBS can be used to generate clinically relevant readouts in medulloblastoma

The identification of molecular features forms a crucial part of diagnosis in medulloblastoma and other paediatric brain tumours (Louis *et al.*, 2021). The combination of methylation class assignment and clinical feature identification forms a critical component in medulloblastoma diagnostics and will likely play an increasingly important role in future WHO classifications. Regions of aneuploidy (copy number variation) like monosomy 6 and MYC amplification are associated with WNT and group 3 respectively, and the determination of TP53 mutation is used by clinicians to stratify SHH patients. The identification of these features typically involves the employment of various platforms. Aneuploidy measurements are typically inferred using methylation microarray and/or karyotyping approaches like FISH, and mutation status is achieved via targeted sequencing panel. Aneuploidy using DNA methylation microarray is a popular option for researchers in paediatric oncology, where a subgroup call can also be obtained at no additional cost. Whole and single-arm chromosomal changes can be called with high accuracy and focal gains/losses can be called where probe sites are numerous. However, the probe centric design inherently limits calling, where a lack of whole genome coverage and uneven probe distribution resulting in many regions of aneuploidy being missed. The calling of differentially methylated regions (DMRs) is also limited, where many regions of differential methylation are missed.

The combination of DNA methylation microarray and multi-gene mutation panel play critical roles in medulloblastoma diagnostics in developed nations. However, sequencing based approaches offer a powerful alternative that has the potential to interrogate the whole genome simultaneously. WGBS is uniquely placed to offer both sequencing and methylation-based analyses, and the cost of generating data at low-pass is becoming economically comparable to array (10x WGBS = ~£430 / Array = £300). The whole genome coverage offered by WGBS theoretically can be used to derive chromosomal copy number estimates with greater resolution than that offered by the array. DMRs could be identified throughout the methylome and with greater resolution thanks to the single-site resolution offered by the platform. Sequencing approaches have successfully been used to identify clinical variants and regions of aneuploidy in MB, with WGBS and WES increasingly being employed in research (Northcott *et al.*, 2017). Variant calling is also possible, with specialised bisulfite sequencing variant calling tools available (Liu *et al.*, 2021; Merkel *et al.*, 2019). Tools for CNV calling are available that are compatible with WGBS (Talevich *et al.*, 2016; Raman *et al.*, 2019), but have yet to be applied to WGBS sample data sequenced at low pass. The same applies for variant calling and differential methylation analysis has not yet been applied comprehensively in medulloblastoma.

In chapter 5, low-pass WGBS sample data was used to attempt to detect CNVs, variants and regions of differential methylation in medulloblastoma. An optimised CNVKit pipeline supplemented with additional filter regions to identify whole and single-arm regions of aneuploidy was presented, with 96.908% sensitivity compared those called by conumee in our matched array cohort. Focal regions of

aneuploidy could be identified as small as 150kb-200kb without a reference cohort of samples for normalisation. Validated regions of *MYC* and *MYCN* with 100% and 66% sensitivity were identified respectively, exceeding that in the matched array cohort. Further additional whole chromosomal aberrations were detected that appeared to be missed by conumee, as well as focal regions including a *SNCAIP* duplication within a group 4 sample, although validation is needed to corroborate our findings. The excellent copy number calling performance demonstrated in chapter 5 places WGBS in an advantageous position compared to the array, enabling accurate copy number estimates to be identified throughout the genome, even at low pass. The capability of low-pass WGBS to be used to call variants using 'bs call' (Merkel *et al.*, 2019) was then investigated. Unfortunately, this was unsuccessful, where coverage was too low to call variants reliably.

Analysing the landscape of differential methylation throughout the entire methylome in a subgroup specific manner revealed vastly increased numbers of DMRs were identified in WGBS compared to array-based analysis. DMRs were larger and called with greater methylation change on average across all molecular subgroups, with the greatest proportion of DMRs found to overlap with open sea CpGs rather than regions covered by the array manifests. These findings agreed with our variable features analysis, highlighting that significant differential methylation is occurring in medulloblastoma that is missed when analysing samples via array. Genomic annotation of DMRs revealed subgroup-specific patterns that differed wildly compared to DMRs identified in our array cohort. Gene ontology analysis of DMRs that overlapped with genic regions identified significantly greater numbers of enriched biological processes that were affected by differential methylation than were in our array analysis.

A focused analysis within the NMF directed split of subtype VII was also performed, for which there was clinical evidence identified when analysing array sample data (Sharma *et al.*, 2019). Over 5,981 DMRs were identified that enriched over 84 biological processes. The same analysis in our array cohort identified just a single DMR that enriched 1 biological process respectively. Our findings put forward a strong case for the usage of low-pass WGBS in being a comparable (and potentially superior) alternative to the DNA methylation microarray platform in identifying differentially methylated regions. The evidence that significant amounts of variability is present within medulloblastoma that is missed by the array is compelling, and the large differences observed between subgroups suggest that they could play a role in providing a more comprehensive, optimised whole-methylome profiles for each respective subgroup/subtype.

## 6.7 – Future work

### 6.7.1 – Developing additional methods of imputation for use with WGBS that are less memory-intensive

Imputation of missing data is a necessary step to circumvent the high degree of missingness that is a key obstacle when working with data sequenced at low pass. Imputation is a memory-intensive operation, with most established techniques for imputing array methylation data becoming computationally unfeasible when concerned with WGBS methylation values, where a significantly higher degree of missingness is present per sample across ~28million CpGs. We demonstrated success using machine-learning based methods, where missing values were predicted with excellent model performance but with the caveat that significant amounts of computing power required to achieve this, even with our modest cohort size (<100 samples). Whilst our method is successful for our cohort ( $n = 78$ ), it would quickly become unfeasible when working with larger cohorts numbering in the 100s or 1000s. To increase accessibility to researchers, the development of imputation techniques that place less of a memory burden upon computing infrastructure is a necessity. Some research is beginning to take place regarding the implementation of imputation algorithms optimised for whole genome methylation data – namely the introduction of sliding-window approaches. Sliding window-based approaches require that only a portion of a matrix is loaded at any one time, significantly reducing memory burden. Such an approach would be directly applicable to WGBS, where there is no need for an entire matrix to be within memory at any time, and rather just a subset of CpG rows, where typically only neighbouring CpGs are needed to impute any specific CpG of interest. A sliding-window k-nearest neighbour approach has now been developed, but its capability to impute WGBS has not been robustly tested in clinical sample data yet (Farrell *et al.*, 2021). The k-nearest neighbours' algorithm is an established imputation technique for use with DNA methylation microarray but is not the most effective method (Di Lena *et al.*, 2020) – the development of other, established methods of imputation that are compatible with WGBS would be an important step forward to increase the accessibility of low-pass WGBS to researchers.

### 6.7.2 – Establishment of user-friendly software for data handling of samples sequenced via low-pass WGBS

The bioinformatic techniques described throughout this document are complex. Various packages across multiple programming languages are utilised and would require at least intermediate to advanced bioinformatics experience to establish and use within a high-performance computing environment. Many software suites that bring together multiple analytical packages are available, even for WGBS (gemBS, BSBolt) which enable them to be more accessible to researchers that are lacking the bioinformatic experience required to process samples and generate meaningful analytical data. Uniting

the analyses presented in this project – sample processing, methylation extraction, subgrouping, copy number calling and differential methylation would present a significant ‘quality-of-life’ improvement for users with limited bioinformatics experience interested in analysing WGBS sample data at low pass.

### **6.7.3 – Enabling WGBS sample compatibility with the DFKZ paediatric brain tumour classifier**

This study has demonstrated that the molecular subgroups of medulloblastoma can successfully be determined using low-pass WGBS. Methylation values are highly correlated to matched sample data analysed via methylation microarray, enabling WGBS sample data to be applied to pre-existing array-trained classification models to obtain molecular subgroup calls with high probability. Classifiers trained to assign subgroup accept data in the form of beta values, which is a universal measurement of methylation used by both the array and methylation sequencing platforms. Providing software for the classification model of interest is publicly available, WGBS samples can easily be applied to models to derive a subgroup prediction. In this study we used two SVM models to derive subgroup but their usage clinically has since been superseded by the DFKZ paediatric brain tumour classifier. We were not able to test the performance of this classifier in assigning subgroup to WGBS sample data as the software is unfortunately not publicly available. Given the data from this study, investigating the compatibility of WGBS sample data with the DFKZ classifier would be an important next step in ensuring a smooth transition towards methylation sequencing in the coming years.

### **6.7.4 – Novel ‘omics profiling of medulloblastoma subtypes via WGBS**

This study has demonstrated the capability of low-pass WGBS to be used to generate accurate measurements of methylation throughout the methylome, supplemented further with generation of CNV and DMR calling throughout the whole genome. Given the increased methylation variability observed throughout the methylome, it would be an important future goal to characterise the methylation profile of the medulloblastoma subtypes throughout the methylome without being restricted to sites covered by the methylation microarray. We show that significant differential methylation is occurring throughout the genome, which contribute to characterising distinct, separable methylation profiles within the subtypes of medulloblastoma. This study was restricted to characterising molecular subgroups due to cohort size but where sample size was sufficient in subtype VII, we show that the proposed NMF-directed split within this subtype can be characterised by distinct patterns of differential methylation. Sequencing of more samples would enable such an analysis to be done, and unification with gene expression datasets would allow any direct correlations between methylation and expression to be identified.



## 6.8 – Implication of Results

The data produced in this project show that WGBS has significant potential to be used as a platform for clinical diagnostics in medulloblastoma in the coming years. Even at lower depths, high-quality methylation values and aneuploidy calls can be generated that are capable of being used for accurate subgroup assignment and risk stratification, with variant calling also shown to be possible at higher depths. These results are promising, demonstrating applicability for WGBS to be used as a platform for clinical diagnostics in medulloblastoma in the coming years. Whilst this project focuses on paediatric medulloblastoma, the applicability of a methylation-based assay for classification and aneuploidy analysis extends to that of other paediatric CNS tumour types, which rely on the generation of the same data to reach accurate subgroup diagnosis and risk stratification within their respective tumour types (Locke *et al.*, 2019). Indeed, WGBS-based assays may also be applicable outside of CNS tumours. Many methylation marker-based assays are available that are designed to detect the presence of multiple different types of cancers. One example is the EPICUP assay, built upon the Illumina Methylation450k platform and is used to identify the primary tissue origin of cancers diagnosed with an unknown origin (Moran *et al.*, 2016). Outside of cancer, genome wide methylation profiling has shown utility as a tool for the molecular diagnosis of neurodevelopmental presentations and congenital anomalies, highlighting the potential for WGBS to be applied among other types of disease (Arif-Eshghi *et al.*, 2019).

Such a scenario where WGBS is used in a diagnostic setting is not unreasonable, where the implementation of sequencing infrastructure within healthcare systems in developed nations is becoming more prominent (Stark *et al.*, 2019). Governments in over 14 developed nations have invested over \$4 billion into the establishment of sequencing infrastructure since 2013, with the UK's NHS Genomics England entering partnership with Illumina to implement DNA sequencing as standard diagnostic practice for rare genetic diseases in 2020 (Genomics England, 2020). With Illumina being the provider of the methylation microarray and WGBS requiring just an additional bisulfite treatment step to take advantage of such infrastructure, it is easy to envision a future where WGBS-based diagnostic assays like that described in this project could be used within such infrastructure in the future. Cost is an important factor in the adoption of assays in a clinical setting and demonstrating that high-quality low-depth methylation data can be generated in this project provides significant economic benefit. With sequencing costs continuing to fall for WGBS and alternative library preparation methods continuing to lower DNA input requirements, it is highly likely that the cost of WGBS may fall below that which is currently offered by Illumina for the methylation microarray (WGBS vs Array costs).

## 6.9 – Concluding remarks

In this study we describe an optimised pipeline for processing, quality control and analysis of clinical sample data sequenced at low pass. Our pipeline includes imputation, which successfully imputes missing values with high correlation compared to matched array samples. Data generated is suitable for subgrouping, where we show that high subgroup assignment probabilities can be acquired by applying WGBS sample data directly to existing classification models trained using array data. Validated classification models trained using WGBS sample data perform equivalently to identical models trained using matched array data, and we show that performance is robust across most models tested at coverages as low as 1x.

We surveyed the entire methylome of medulloblastoma and highlight that methylation that cannot be detected via array platforms is highly variable among molecular subgroups. This subsequently translated to increased subgroup separation when applying visualisation techniques. This translated further when analysing the landscape of differential methylation within medulloblastoma, where we identified significantly increased numbers of differentially methylated regions compared to array-based analyses in the same samples. Finally, we present an optimised copy number variation calling pipeline suitable for use with low-pass WGBS. Large scale chromosomal alterations are detected with near identical sensitivity to array-based methods and additional focal regions are successfully identified without a reference cohort of samples for normalisation.

In summary, we describe a platform-independent assay for molecular subgrouping of medulloblastoma via low-pass WGBS. It performs equivalently to array-based methods at comparable cost and provides a proof-of-concept for its routine adoption in a research and clinical setting using standard WGS technologies. Finally, analysis of the whole methylome enabled the elucidation of existing heterogeneity with greater resolution compared to the array, indicating that additional heterogeneity is present within medulloblastoma that has hitherto been inaccessible.

## **7 – References**

---

ABCAM. (2019). *Histone Modifications: A Guide*. Available at: <https://www.abcam.com/epigenetics/histone-modifications-a-guide>. (Accessed 14th January 2022).

Adams, J.M. and Strasser, A. (2008) 'Is Tumor Growth Sustained by Rare Cancer Stem Cells or Dominant Clones?', *Cancer Research*, 68(11), pp. 4018–4021.

Al-Hajj, M., Wicha, M.S., Benito-Hernandez, A., Morrison, S.J. and Clarke, M.F. (2003) 'Prospective identification of tumorigenic breast cancer cells', *Proceedings of the National Academy of Sciences*, 100(7), pp. 3983–3988.

Altman, N.S. and Altman, N.S. (1991) 'BU-1065MA An Introduction to Kernel and Nearest Neighbor Nonparametric Regression An Introduction to Kernel and Nearest Neighbor Nonparametric Regression'.

Amami, R., Ayed, D. Ben and Ellouze, N. (no date) 'Practical Selection of SVM Supervised Parameters with Different Feature Representations for Vowel Recognition'.

Anaconda Software Distribution. (2020) 'Anaconda Documentation'. Anaconda Inc. Available at: <https://docs.anaconda.com/>.

Andrews S. (no date) *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. Available at: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (Accessed: 7 December 2021).

Archer, T.C., Mahoney, E.L. and Pomeroy, S.L. (2017) 'Medulloblastoma: Molecular Classification-Based Personal Therapeutics.', *Neurotherapeutics: the journal of the American Society for Experimental NeuroTherapeutics*, 14(2), pp. 265–273.

Aref-Eshghi, E., Bend, E. G., Colaiacovo, S., Caudle, M., Chakrabarti, R., Napier, M., Brick, L., Brady, L., Carere, D. A., Levy, M. A., Kerkhof, J., Stuart, A., Saleh, M., Beaudet, A. L., Li, C., Kozenko, M., Karp, N., Prasad, C., Siu, V. M., ... Sadikovic, B. (2019). Diagnostic Utility of Genome-wide DNA Methylation Testing in Genetically Unsolved Individuals with Suspected Hereditary Conditions. *The American Journal of Human Genetics*, 104(4), 685–700. <https://doi.org/10.1016/J.AJHG.2019.03.008>

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., Gibbs, R.A., Green, E.D., Hurles, M.E., Knoppers, B.M., Korbel, J.O., Lander, E.S., Lee, C., Lehrach, H., *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74.

---

Babenko, V.N., Chadaeva, I. V. and Orlov, Y.L. (2017) 'Genomic landscape of CpG rich elements in human', *BMC Evolutionary Biology*, 17(Suppl 1), pp. 1–11.

BAILEY, P. and CUSHING, H. (1925) 'MEDULLOBLASTOMA CEREBELLI', *Archives of Neurology & Psychiatry*, 14(2), p. 192.

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research* 2011 21:3, 21(3), 381–395.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) 'High-Resolution Profiling of Histone Methylations in the Human Genome', *Cell*, 129(4), pp. 823–837.

Baruch, E.N., Youngster, I., Ben-Betzalel, G., Ortenberg, R., Lahat, A., Katz, L., Adler, K., Dick-Necula, D., Raskin, S., Bloch, N., Rotin, D., Anafi, L., Avivi, C., Melnichenko, J., Steinberg-Silman, Y., Mamtani, R., Harati, H., Asher, N., Shapira-Frommer, R., *et al.* (2021) 'Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients', *Science*, 371(6529), pp. 602–609.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., *et al.* (2008) 'Accurate whole human genome sequencing using reversible terminator chemistry', *Nature*, 456(7218), pp. 53–59.

Bhat, S., Coleman, H.G., Yousef, F., Johnston, B.T., McManus, D.T., Gavin, A.T. and Murray, L.J. (2011) 'Risk of Malignant Progression in Barrett's Esophagus Patients: Results from a Large Population-Based Study', *JNCI Journal of the National Cancer Institute*, 103(13), pp. 1049–1057.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., Fan, J.B. and Shen, R. (2011) 'High density DNA methylation array with single CpG site resolution', *Genomics*, 98(4), pp. 288–295.

Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R. and Gunderson, K.L. (2009) 'Genome-wide DNA methylation profiling using Infinium<sup>®</sup> assay', *Epigenomics*, 1(1), pp. 177–200.

Birch, J. and Gil, J. (2020) 'Senescence and the SASP: many therapeutic avenues', *Genes & Development*, 34(23–24), pp. 1565–1576.

Borowska, A. and Józwiak, J. (2016) 'Medulloblastoma: molecular pathways and histopathological classification.', *Archives of medical science: AMS*, 12(3), pp. 659–66.

---

Botezatu, A., Iancu, I. V., Popa, O., Plesa, A., Manda, D., Huica, I., Vladoiu, S., Anton, G. and Badiu, C. (2016) 'Mechanisms of Oncogene Activation', in *New Aspects in Molecular and Cellular Mechanisms of Human Carcinogenesis*. InTech.

Brasme, J.-F., Chalumeau, M., Doz, F., Lacour, B., Valteau-Couanet, D., Gaillard, S., Delalande, O., Aghakhani, N., Sainte-Rose, C., Puget, S. and Grill, J. (2012) 'Interval between onset of symptoms and diagnosis of medulloblastoma in children: distribution and determinants in a population-based study', *European Journal of Pediatrics*, 171(1), pp. 25–32.

Breiman, L. (2001) 'Random Forests', *Machine Learning 2001 45:1*, 45(1), pp. 5–32.

Brinkman, A.B., Simmer, F., Ma, K., Kaan, A., Zhu, J. and Stunnenberg, H.G. (2010) 'Whole-genome DNA methylation profiling using MethylCap-seq', *Methods*, 52(3), pp. 232–236.

Bühren, J., Christoph, A.H., Buslei, R., Albrecht, S., Wiestler, O.D. and Pietsch, T. (2000) 'Expression of the neurotrophin receptor p75NTR in medulloblastomas is correlated with distinct histological and clinical features: evidence for a medulloblastoma subtype derived from the external granule cell layer.', *Journal of neuropathology and experimental neurology*, 59(3), pp. 229–40.

Campen, C.J., Kranick, S.M., Kasner, S.E., Kessler, S.K., Zimmerman, R.A., Lustig, R., Phillips, P.C., Storm, P.B., Smith, S.E., Ichord, R. and Fisher, M.J. (2012) 'Cranial irradiation increases risk of stroke in pediatric brain tumor survivors.', *Stroke*, 43(11), pp. 3035–40.

Cancer Research UK (2022) *Cancer incidence statistics | Cancer Research UK*. Available at: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence#heading=Two> (Accessed: 30 January 2022).

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., Trapnell, C. and Shendure, J. (2019) 'The single-cell transcriptional landscape of mammalian organogenesis', *Nature* 2019 566:7745, 566(7745), pp. 496–502.

Capper, D., Jones, D.T.W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D.E., Kratz, A., Wefers, A.K., Huang, K., Pajtler, K.W., Schweizer, L., Stichel, D., Olar, A., Engel, N.W., Lindenberg, K., *et al.* (2018) 'DNA methylation-based classification of central nervous system tumours', *Nature*, 555(7697), pp. 469–474.

Carvalho Neto, A. de, Gasparetto, E.L., Ono, S.E., Bertoldi, G.A. and Gomes, A.F. (2003) 'Adult cerebellar medulloblastoma: CT and MRI findings in eight cases', *Arquivos de Neuro-Psiquiatria*, 61(2A), pp. 199–203.

Cavalcante, R.G. and Sartor, M.A. (2017) 'annotatr: genomic regions in context', *Bioinformatics*, 33(15), p. 2381.

Cavalli, F.M.G., Remke, M., Rampasek, L., Peacock, J., Shih, D.J.H., Luu, B., Garzia, L., Torchia, J., Nor, C., Morrissy, A.S., Agnihotri, S., Thompson, Y.Y., Kuzan-Fischer, C.M., Farooq, H., Isaev, K., Daniels, C., Cho, B.-K., Kim, S.-K., Wang, K.-C., *et al.* (2017) 'Intertumoral Heterogeneity within Medulloblastoma Subgroups.', *Cancer cell*, 31(6), pp. 737-754.e6.

Chaffer, C.L. and Weinberg, R.A. (2011) 'A Perspective on Cancer Cell Metastasis', *Science*, 331(6024), pp. 1559–1564.

Chang, C.H., Housepian, E.M. and Herbert, C. (1969) 'An Operative Staging System and a Megavoltage Radiotherapeutic Technic for Cerebellar Medulloblastomas1', , 93(6), pp. 1351–1359.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N.R. and Ma'ayan, A. (2013) 'Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool', *BMC Bioinformatics*, 14(1), pp. 1–14.

Chen, H., Liu, H. and Qing, G. (2018) 'Targeting oncogenic Myc as a strategy for cancer treatment', *Signal Transduction and Targeted Therapy*, 3(1), p. 5.

Chen, T. and Guestrin, C. (no date) 'XGBoost: A Scalable Tree Boosting System', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* [Preprint].

Chipumuro, E., Marco, E., Christensen, C.L., Kwiatkowski, N., Zhang, T., Hatheway, C.M., Abraham, B.J., Sharma, B., Yeung, C., Altabef, A., Perez-Atayde, A., Wong, K.-K., Yuan, G.-C., Gray, N.S., Young, R.A. and George, R.E. (2014) 'CDK7 inhibition suppresses super-enhancer-linked oncogenic transcription in MYCN-driven cancer.', *Cell*, 159(5), pp. 1126–1139.

Cho, S., Kim, H.S., Zeiger, M.A., Umbricht, C.B. and Cope, L.M. (2019) 'Measuring DNA Copy Number Variation Using High-Density Methylation Microarrays', *Journal of Computational Biology*, 26(4), p. 295.

Cho, Y.-J., Tsherniak, A., Tamayo, P., Santagata, S., Ligon, A., Greulich, H., Berhoukim, R., Amani, V., Goumnerova, L., Eberhart, C.G., Lau, C.C., Olson, J.M., Gilbertson, R.J., Gajjar, A., Delattre, O., Kool, M., Ligon, K., Meyerson, M., Mesirov, J.P., *et al.* (2011) 'Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome.', *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 29(11), pp. 1424–30.

Choi, J.D. and Lee, J.-S. (2013) 'Interplay between Epigenetics and Genetics in Cancer.', *Genomics & informatics*, 11(4), pp. 164–73.

Choudhry, Z., Rikani, A.A., Choudhry, A.M., Tariq, S., Zakaria, F., Asghar, M.W., Sarfraz, M.K., Haider, K., Shafiq, A.A. and Mobassarrah, N.J. (2014) 'Sonic hedgehog signalling pathway: a complex network.', *Annals of neurosciences*, 21(1), pp. 28–31.

Clevers, H. (2006) 'Wnt/ $\beta$ -Catenin Signaling in Development and Disease', *Cell*, 127(3), pp. 469–480.

Clifford, S.C., Lusher, M.E., Lindsey, J.C., Langdon, J.A., Gilbertson, R.J., Straughton, D. and Ellison, D.W. (2006) 'Wnt/Wingless pathway activation and chromosome 6 loss characterize a distinct molecular sub-group of medulloblastomas associated with a favorable prognosis', *Cell cycle (Georgetown, Tex.)*, 5(22), pp. 2666–2670.

Clinicaltrials.gov. (2017). 'A Phase II Study of Oral LDE225 in Patients With Hedge-Hog (Hh)-Pathway Activated Relapsed Medulloblastoma (MB)'. Available at: <https://www.clinicaltrials.gov/ct2/show/NCT01708174?cond=medulloblastoma&rank=7>. (Accessed 15th January 2022).

Clinicaltrials.gov. (2018). 'Palbociclib Isethionate in Treating Younger Patients With Recurrent, Progressive, or Refractory Central Nervous System Tumors.' Available at: <https://clinicaltrials.gov/ct2/show/NCT02255461>. (Accessed 15th January 2022).

Clinicaltrials.gov. (2018). 'Reduced Craniospinal Radiation Therapy and Chemotherapy in Treating Younger Patients With Newly Diagnosed WNT-Driven Medulloblastoma.' Available at: <https://www.clinicaltrials.gov/ct2/show/NCT02724579?cond=wnt+medulloblastoma&rank=2>. (Accessed 15th January 2022).

Clinicaltrials.gov. (2018). 'Study Assessing the Feasibility of a Surgery and Chemotherapy-Only in Children With Wnt Positive Medulloblastoma'. Available at: <https://www.clinicaltrials.gov/ct2/show/NCT02212574?id=NCT02212574&rank=1&load=cart>. (Accessed 15th January 2022).

Coleman, B., McLendon, R., Kranz, P.G. and Adamson, D.C. (2012) 'Medulloblastoma', *Contemporary Neurosurgery*, 34(3), pp. 1–5

Coutelier, M. *et al.* (2021) 'Combining callers improves the detection of copy number variants from whole-genome sequencing', *European Journal of Human Genetics 2021*. Nature Publishing Group, pp. 1–9.

---



Coutelier, M., Holtgrewe, M., Jäger, M., Flöttman, R., Mensah, M.A., Spielmann, M., Krawitz, P., Horn, D., Beule, D. and Mundlos, S. (2021) 'Combining callers improves the detection of copy number variants from whole-genome sequencing', *European Journal of Human Genetics* 2021, pp. 1–9.

Crary-Dooley, F.K., Tam, M.E., Dunaway, K.W., Hertz-Picciotto, I., Schmidt, R.J. and LaSalle, J.M. (2017) 'A comparison of existing global DNA methylation assays to low-coverage whole-genome bisulfite sequencing for epidemiological studies.', *Epigenetics*, 12(3), pp. 206–214.

Crosier, S., Hicks, D., Schwalbe, E.C., Williamson, D., Leigh Nicholson, S., Smith, A., Engebretsen, S. and Bohlin, J. (2019) 'Statistical predictions with glmnet', *Clinical Epigenetics*, 11(1).

Dawood, S., Austin, L. and Cristofanilli, M. (2014) 'Cancer stem cells: implications for cancer therapy.', *Oncology (Williston Park, N.Y.)*, 28(12), pp. 1101–7, 1110.

de Bont, J.M., Packer, R.J., Michiels, E.M., den Boer, M.L. and Pieters, R. (2008) 'Biological background of pediatric medulloblastoma and ependymoma: a review from a translational research perspective.', *Neuro-oncology*, 10(6), pp. 1040–60.

del Charco, J.O., Bolek, T.W., McCollough, W.M., Maria, B.L., Kedar, A., Braylan, R.C., Mickle, J.P., Buatti, J.M., Mendenhall, N.P. and Marcus, R.B. (1998) 'Medulloblastoma: time-dose relationship based on a 30-year review.', *International journal of radiation oncology, biology, physics*, 42(1), pp. 147–54.

Derakhshan, A., Chen, Z. and Van Waes, C. (2017) 'Therapeutic Small Molecules Target Inhibitor of Apoptosis Proteins in Cancers with Deregulation of Extrinsic and Intrinsic Cell Death Pathways', *Clinical Cancer Research*, 23(6), pp. 1379–1387.

Di Lena, P., Sala, C., Prodi, A. and Nardini, C. (2019) 'Missing value estimation methods for DNA methylation data', *Bioinformatics (Oxford, England)*, 35(19), pp. 3786–3793.

Diede, S.J., Guenthoer, J., Geng, L.N., Mahoney, S.E., Marotta, M., Olson, J.M., Tanaka, H. and Tapscott, S.J. (no date) 'DNA methylation of developmental genes in pediatric medulloblastomas identified by denaturation analysis of methylation differences'.

Djirackor, L., Halldorsson, S., Niehusmann, P., Leske, H., Capper, D., Kuschel, L.P., Pahnke, J., Due-Tønnessen, B.J., Langmoen, I.A., Sandberg, C.J., Euskirchen, P. and Vik-Mo, E.O. (2021) 'Intraoperative DNA methylation classification of brain tumors impacts neurosurgical strategy', *Neuro-Oncology Advances*, 3(1).

DNeasy Blood & Tissue Kits (no date). Available at: <https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/dna-purification/genomic-dna/dneasy-blood-and-tissue-kit/> (Accessed: 7 December 2021).

Doherty, R. and Couldrey, C. (2014) 'Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment', *Frontiers in Genetics*, 5(MAY), pp. 1–1.

Dong, Z., Xie, W., Chen, H., Xu, J., Wang, H., Li, Y., Wang, J., Chen, F., Choy, K.W. and Jiang, H. (2017) 'Copy-Number Variants Detection by Low-Pass Whole-Genome Sequencing', *Current protocols in human genetics*, 94, pp. 8.17.1-8.17.16.

Doxey, D., Bruce, D., Sklar, F., Swift, D. and Shapiro, K. (1999) 'Posterior Fossa Syndrome: Identifiable Risk Factors and Irreversible Complications', *Pediatric Neurosurgery*, 31(3), pp. 131–136.

Dunford, A., Weinstock, D.M., Savova, V., Schumacher, S.E., Cleary, J.P., Yoda, A., Sullivan, T.J., Hess, J.M., Gimelbrant, A.A., Beroukhi, R., Lawrence, M.S., Getz, G. and Lane, A.A. (2017) 'Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias.', *Nature genetics*, 49(1), pp. 10–16.

Dzutsev, A., Badger, J.H., Perez-Chanona, E., Roy, S., Salcedo, R., Smith, C.K. and Trinchieri, G. (2017) 'Microbes and Cancer', *Annual review of immunology*, 35, pp. 199–228.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T.A., Haefliger, C., Horton, R., Howe, K., Jackson, D.K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., *et al.* (2006) 'DNA methylation profiling of human chromosomes 6, 20 and 22', *Nature genetics*, 38(12), pp. 1378–1385.

El Doussouki, M., Gajjar, A. and Chamdine, O. (2019) 'Molecular genetics of medulloblastoma in children: Diagnostic, therapeutic and prognostic implications', *Future Neurology*, 14(1).

El-Kenawi, A.E. and El-Remessy, A.B. (2013) 'Angiogenesis inhibitors in cancer therapy: mechanistic perspective on classification and treatment rationales.', *British journal of pharmacology*, 170(4), pp. 712–29.

Ellison, D.W. (2010) 'Childhood medulloblastoma: novel approaches to the classification of a heterogeneous disease', *Acta Neuropathologica*, 120(3), pp. 305–316.

Ellison, D.W., Dalton, J., Kocak, M., Nicholson, S.L., Fraga, C., Neale, G., Kenney, A.M., Brat, D.J., Perry, A., Yong, W.H., Taylor, R.E., Bailey, S., Clifford, S.C. and Gilbertson, R.J. (2011) 'Medulloblastoma: clinicopathological correlates of SHH, WNT, and non-SHH/WNT molecular subgroups.', *Acta neuropathologica*, 121(3), pp. 381–96.

Ellison, D.W., Kocak, M., Dalton, J., Megahed, H., Lusher, M.E., Ryan, S.L., Zhao, W., Nicholson, S.L., Taylor, R.E., Bailey, S. and Clifford, S.C. (2011) 'Definition of disease-risk stratification groups in childhood medulloblastoma using combined clinical, pathologic, and molecular variables', *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 29(11), pp. 1400–1407.

Ellison, D.W., Onilude, O.E., Lindsey, J.C., Lusher, M.E., Weston, C.L., Taylor, R.E., Pearson, A.D. and Clifford, S.C. (2005) 'beta-Catenin status predicts a favorable outcome in childhood medulloblastoma: the United Kingdom Children's Cancer Study Group Brain Tumour Committee', *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 23(31), pp. 7951–7957.

ENCODE (2017) 'Whole-Genome Bisulfite Sequencing Data Standards and Processing Pipeline – ENCODE'. Available at: <https://www.encodeproject.org/data-standards/wgbs/> (Accessed: 7 December 2021).

Engelbrechtsen, S. and Bohlin, J. (2019) 'Statistical predictions with glmnet', *Clinical Epigenetics*, 11(1).

Engelender, S., Kaminsky, Z., Guo, X., Sharp, A.H., Amaravi, R.K., Kleiderlein, J.J., Margolis, R.L., Troncoso, J.C., Lanahan, A.A., Worley, P.F., Dawson, V.L., Dawson, T.M. and Ross, C.A. (1999) 'Synphilin-1 associates with  $\alpha$ -synuclein and promotes the formation of cytosolic inclusions', *Nature Genetics*, 22(1), pp. 110–114.

Ericson, J., Briscoe, J., Rashbass, P., van Heyningen, V. and Jessell, T.M. (1997) 'Graded sonic hedgehog signaling and the specification of cell fate in the ventral neural tube.', *Cold Spring Harbor symposia on quantitative biology*, 62, pp. 451–66. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/9598380> (Accessed: 15 January 2021).

Euskirchen, P., Bielle, F., Labreche, K., Kloosterman, W.P., Rosenberg, S., Daniau, M., Schmitt, C., Masliah-Planchon, J., Bourdeaut, F., Dehais, C., Marie, Y., Delattre, J.Y. and Idbaih, A. (2017) 'Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing', *Acta neuropathologica*, 134(5), pp. 691–703.

Evan, G. and Littlewood, T. (1998) 'A matter of life and cell death.', *Science (New York, N.Y.)*, 281(5381), pp. 1317–22.

Faget, D. V., Ren, Q. and Stewart, S.A. (2019) 'Unmasking senescence: context-dependent effects of SASP in cancer', *Nature Reviews Cancer* 2019 19:8, 19(8), pp. 439–453. 0156-2.

Farrell, C., Thompson, M., Tosevska, A., Oyetunde, A. and Pellegrini, M. (2021) 'BiSulfite Bolt: A bisulfite sequencing analysis platform', *GigaScience*, 10(5).

Farwell, J.R., Dohrmann, G.J. and Flannery, J.T. (1984) 'Medulloblastoma in childhood: an epidemiological study', *Journal of Neurosurgery*, 61(4), pp. 657–664.

Feng, S., Zhong, Z., Wang, M. and Jacobsen, S.E. (2020) 'Efficient and accurate determination of genome-wide DNA methylation patterns in *Arabidopsis thaliana* with enzymatic methyl sequencing', *Epigenetics and Chromatin*, 13(1), pp. 1–17.

Flavahan, W.A., Gaskell, E. and Bernstein, B.E. (2017) 'Epigenetic plasticity and the hallmarks of cancer.', *Science (New York, N.Y.)*, 357(6348), p. eaal2380.

Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B.W., Hudson, T.J., Fertig, E.J., Greenwood, C.M. and Hansen, K.D. (2014) 'Functional normalization of 450k methylation array data improves replication in large cancer studies', *Genome Biology*, 15(11), p. 503.

Fortin, J.P., Triche, T.J. and Hansen, K.D. (2017) 'Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi', *Bioinformatics*, 33(4), p. 558.

Frappaz, Di., Barritault, M., Montané, L., Laigle-Donadey, F., Chinot, O., Le Rhun, E., Bonneville-Levard, A., Hottinger, A.F., Meyronnet, D., Bidaux, A.S., Garin, G. and Pérol, D. (2021) 'MEVITEM-a phase I/II trial of vismodegib + temozolomide vs temozolomide in patients with recurrent/refractory medulloblastoma with Sonic Hedgehog pathway activation', *Neuro-oncology*, 23(11), pp. 1949–1960.

Friedrich, C., von Bueren, A.O., von Hoff, K., Kwicien, R., Pietsch, T., Warmuth-Metz, M., Hau, P., Deinlein, F., Kuehl, J., Kortmann, R.D. and Rutkowski, S. (2013) 'Treatment of adult nonmetastatic medulloblastoma patients according to the paediatric HIT 2000 protocol: A prospective observational multicentre study', *European Journal of Cancer*, 49(4), pp. 893–903.

Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L. and Paul, C.L. (1992) 'A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.', *Proceedings of the National Academy of Sciences of the United States of America*, 89(5), pp. 1827–31.

Fuks, F., Burgers, W.A., Brehm, A., Hughes-Davies, L. and Kouzarides, T. (2000) 'DNA methyltransferase Dnmt1 associates with histone deacetylase activity', *Nature Genetics*, 24(1), pp. 88–91

Fuks, F., Hurd, P.J., Deplus, R. and Kouzarides, T. (2003) 'The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase.', *Nucleic acids research*, 31(9), pp. 2305–12.

Gajjar, A., Chintagumpala, M., Ashley, D., Kellie, S., Kun, L.E., Merchant, T.E., Woo, S., Wheeler, G., Ahern, V., Krasin, M.J., Fouladi, M., Broniscer, A., Krance, R., Hale, G.A., Stewart, C.F., Dauser, R., Sanford, R.A., Fuller, C., Lau, C., *et al.* (2006) 'Risk-adapted craniospinal radiotherapy followed by high-dose chemotherapy and stem-cell rescue in children with newly diagnosed medulloblastoma (St Jude Medulloblastoma-96): long-term results from a prospective, multicentre trial', *The Lancet. Oncology*, 7(10), pp. 813–820.

Gajjar, A.J. and Robinson, G.W. (2014) 'Medulloblastoma—translating discoveries from the bench to the bedside', *Nature Reviews Clinical Oncology*, 11(12), pp. 714–722.

Gandola, L., Massimino, M., Cefalo, G., Solero, C., Spreafico, F., Pecori, E., Riva, D., Collini, P., Pignoli, E., Giangaspero, F., Luksch, R., Berretta, S., Poggi, G., Biassoni, V., Ferrari, A., Pollo, B., Favre, C., Sardi, I., Terenziani, M., *et al.* (2009) 'Hyperfractionated accelerated radiotherapy in the Milan strategy for metastatic medulloblastoma', *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 27(4), pp. 566–571.

Garcia-Lopez, J., Kumar, R., Smith, K.S. and Northcott, P.A. (2021) 'Deconstructing Sonic Hedgehog Medulloblastoma: Molecular Subtypes, Drivers, and Beyond', *Trends in genetics: TIG*, 37(3), pp. 235–250.

*Genomics England and Illumina partner to deliver whole genome sequencing for England's NHS Genomic Medicine Service | Genomics England* (2020). Available at: <https://www.genomicsengland.co.uk/genomics-england-illumina-partner-nhs-genomic-medicine-service/> (Accessed: 30 January 2022).

Ghasemi, D.R., Sill, M., Okonechnikov, K., Korshunov, A., Yip, S., Schutz, P.W., Scheie, D., Kruse, A., Harter, P.N., Kastelan, M., Wagner, M., Hartmann, C., Benzel, J., Maass, K.K., Khasraw, M., Sträter, R., Thomas, C., Paulus, W., Kratz, C.P., *et al.* (2019) 'MYCN amplification drives an aggressive form of spinal ependymoma', *Acta neuropathologica*, 138(6), pp. 1075–1089.

Giangaspero, F., Perilongo, G., Fondelli, M.P., Brisigotti, M., Carollo, C., Burnelli, R., Burger, P.C. and Garrè, M.L. (1999) 'Medulloblastoma with extensive nodularity: a variant with favorable prognosis', *Journal of Neurosurgery*, 91(6), pp. 971–977.

Gibson, P., Tong, Y., Robinson, G., Thompson, M.C., Currle, D.S., Eden, C., Kranenburg, T.A., Hogg, T., Poppleton, H., Martin, J., Finkelstein, D., Pounds, S., Weiss, A., Patay, Z., Scoggins, M., Ogg, R., Pei, Y., Yang, Z.-J., Brun, S., *et al.* (2010) 'Subtypes of medulloblastoma have distinct developmental origins', *Nature*, 468(7327), pp. 1095–1099.

Gilbertson, R.J. (2004) 'Medulloblastoma: signalling a change in treatment', *The Lancet Oncology*, 5(4), pp. 209–218.

Gilbertson, R.J. and Ellison, D.W. (2008) 'The Origins of Medulloblastoma Subtypes', *Annual Review of Pathology: Mechanisms of Disease*, 3(1), pp. 341–365.

Goodrich, L. V, Milenković, L., Higgins, K.M. and Scott, M.P. (1997) 'Altered neural cell fates and medulloblastoma in mouse patched mutants.', *Science (New York, N.Y.)*, 277(5329), pp. 1109–13.

Griffiths, J.A., Richard, A.C., Bach, K., Lun, A.T.L. and Marioni, J.C. (2018) 'Detection and removal of barcode swapping in single-cell RNA-seq data', *Nature Communications 2018 9:1*, 9(1), pp. 1–6.

Grivennikov, S.I., Greten, F.R. and Karin, M. (no date) 'Leading Edge Review Immunity, Inflammation, and Cancer', *Cell*, 140, pp. 883–899.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) 'Gene Selection for Cancer Classification using Support Vector Machines', *Machine Learning 2002 46:1*, 46(1), pp. 389–422.

Hamilton, S.R., Liu, B., Parsons, R.E., Papadopoulos, N., Jen, J., Powell, S.M., Krush, A.J., Berk, T., Cohen, Z., Tetu, B., Burger, P.C., Wood, P.A., Taqi, F., Booker, S. V., Petersen, G.M., Offerhaus, G.J.A., Tersmette, A.C., Giardiello, F.M., Vogelstein, B., *et al.* (1995) 'The Molecular Basis of Turcot's Syndrome', *New England Journal of Medicine*, 332(13), pp. 839–847.

Hanahan, D. (2022) 'Hallmarks of Cancer: New Dimensions', *Cancer Discovery*, 12(1), pp. 31–46.

Hanahan, D. and Weinberg, R.A. (2000) 'The hallmarks of cancer.', *Cell*, 100(1), pp. 57–70.

Hanahan, D. and Weinberg, R.A. (2011) 'Hallmarks of cancer: the next generation.', *Cell*, 144(5), pp. 646–74.

Hansen, K.D., Langmead, B. and Irizarry, R.A. (2012) 'BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions', *Genome Biology*, 13(10), pp. 1–10.

Hill, R.M., Richardson, S., Schwalbe, E.C., Hicks, D., Lindsey, J.C., Crosier, S., Rafiee, G., Grabovska, Y., Wharton, S.B., Jacques, T.S., Michalski, A., Joshi, A., Pizer, B., Williamson, D., Bailey, S. and Clifford, S.C. (2020) 'Time, pattern, and outcome of medulloblastoma relapse and their association with tumour biology at diagnosis and therapy: a multicentre cohort study', *The Lancet Child and Adolescent Health*, 4(12), pp. 865–874.

Holdhof, D., Johann, P.D., Spohn, M., Bockmayr, M., Safaei, S., Joshi, P., Masliah-Planchon, J., Ho, B., Andrianteranagna, M., Bourdeaut, F., Huang, A., Kool, M., Upadhyaya, S.A., Bendel, A.E., Indenbirken, D., Foulkes, W.D., Bush, J.W., Creytens, D., Kordes, U., *et al.* (2021) 'Atypical teratoid/rhabdoid tumors (ATRTs) with SMARCA4 mutation are molecularly distinct from SMARCB1-deficient cases', *Acta Neuropathologica*, 141(2), pp. 291–301.

Hoppe, R., Kandabarau, S., Fan, P., Winter, S., Abderrahman, B., Cunliffe, H., Schroth, W., Jordan, V.C. and Brauch, H.B. (2020) 'Abstract 144: Genome-wide DNA methylation changes in AI-resistant breast cancer models during E2-treatment', *Cancer Research*, 80(16) pp. 144–144.

Hotelling, H. (1933) 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology*, 24(6), pp. 417–441.

Hovestadt, V., Ayrault, O., Swartling, F.J., Robinson, G.W., Pfister, S.M. and Northcott, P.A. (2019) 'Medulloblastomics revisited: biological and clinical insights from thousands of patients', *Nature Reviews Cancer* 2019 20:1, 20(1), pp. 42–56.

Hovestadt, V., Jones, D.T.W., Picelli, S., Wang, W., Kool, M., Northcott, P.A., Sultan, M., Stachurski, K., Ryzhova, M., Warnatz, H.-J., Ralser, M., Brun, S., Bunt, J., Jäger, N., Kleinheinz, K., Erkek, S., Weber, U.D., Bartholomae, C.C., von Kalle, C., *et al.* (2014) 'Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing', *Nature*, 510(7506), pp. 537–541.

Hovestadt, V., Remke, M., Kool, M., Pietsch, T., Northcott, P.A., Fischer, R., Cavalli, F.M.G., Ramaswamy, V., Zapatka, M., Reifemberger, G., Rutkowski, S., Schick, M., Bewerunge-Hudler, M., Korshunov, A., Lichter, P., Taylor, M.D., Pfister, S.M. and Jones, D.T.W. (2013) 'Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays.', *Acta neuropathologica*, 125(6), pp. 913–6.

Hsu, P.P. and Sabatini, D.M. (2008) 'Cancer Cell Metabolism: Warburg and Beyond', *Cell*, 134(5), pp. 703–707.



Hu, N., Richards, R. and Jensen, R. (2016) 'Role of chromosomal 1p/19q co-deletion on the prognosis of oligodendrogliomas: A systematic review and meta-analysis', *Interdisciplinary Neurosurgery*, 5, pp. 58–63.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R. and Bird, A.P. (2010) 'Orphan CpG islands identify numerous conserved promoters in the mammalian genome.', *PLoS genetics*, 6(9), p. e1001134.

Illumina Adapter Sequences (no date). Available at: <https://emea.support.illumina.com/downloads/illumina-adapter-sequences-document-1000000002694.html> (Accessed: 7 December 2021).

Illumina. (2019). 'High Density Arrays for DNA Analysis.' Available at: <https://www.illumina.com/science/technology/beadarray-technology/infinium-assay.html?langsel=/us/>. (Accessed 16th January 2022).

*International Society of Paediatric Oncology (SIOP) PNET 5 Medulloblastoma - Full Text View - ClinicalTrials.gov* (no date). Available at: <https://clinicaltrials.gov/ct2/show/NCT02066220> (Accessed: 26 January 2022).

Jafri, M.A., Ansari, S.A., Alqahtani, M.H. and Shay, J.W. (2016) 'Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies.', *Genome medicine*, 8(1), p. 69.

Jäger, M., Wang, K., Bauer, S., Smedley, D., Krawitz, P. and Robinson, P.N. (2014) 'Jannovar: A Java Library for Exome Annotation', *Human Mutation*, 35(5), pp. 548–555.

Jakacki, R.I., Burger, P.C., Zhou, T., Holmes, E.J., Kocak, M., Onar, A., Goldwein, J., Mehta, M., Packer, R.J., Tarbell, N., Fitz, C., Vezina, G., Hilden, J. and Pollack, I.F. (2012) 'Outcome of children with metastatic medulloblastoma treated with carboplatin during craniospinal radiotherapy: a Children's Oncology Group Phase I/II study', *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 30(21), pp. 2648–2653

Jang, H.S., Shin, W.J., Lee, J.E. and Do, J.T. (2017) 'CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function', *Genes*, 8(6), pp. 2–20.

Johnson, A.A., Akman, K., Calimport, S.R.G., Wuttke, D., Stolzing, A. and De Magalhães, J.P. (2012) 'The Role of DNA Methylation in Aging, Rejuvenation, and Age-Related Disease', *Rejuvenation Research*, 15(5), p. 483.



Jones, D.T.W., Worst, B.C., *et al.* (2017) 'The whole-genome landscape of medulloblastoma subtypes', *Nature* 2017 547:7663, 547(7663), pp. 311–317.

Jones, R.G. and Thompson, C.B. (2009) 'Tumor suppressors and cell metabolism: a recipe for cancer growth.', *Genes & development*, 23(5), pp. 537–48.

Jühling, F., Kretzmer, H., Bernhart, S.H., Otto, C., Stadler, P.F. and Hoffmann, S. (2015) 'metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data', *Genome Research*, 26(2), pp. 256–262.

Kahn, M. (2014) 'Can we safely target the WNT pathway?', *Nature reviews. Drug discovery*, 13(7), pp. 513–32.

Kamalakaran, S., Varadan, V., Giercksky Russnes, H.E., Levy, D., Kendall, J., Janevski, A., Riggs, M., Banerjee, N., Synnestvedt, M., Schlichting, E., Kåresen, R., Shama Prasada, K., Rotti, H., Rao, R., Rao, L., Eric Tang, M.-H., Satyamoorthy, K., Lucito, R., Wigler, M., *et al.* (2011) 'DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables.', *Molecular oncology*, 5(1), pp. 77–92.

Kamps, R., Brandão, R.D., Bosch, B.J. van den, Paulussen, A.D.C., Xanthoulea, S., Blok, M.J. and Romano, A. (2017) 'Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification.', *International journal of molecular sciences*, 18(2).

Kawakatsu, T., Huang, S. shan C., Jupe, F., Sasaki, E., Schmitz, R.J.J., Urich, M.A.A., Castanon, R., Nery, J.R.R., Barragan, C., He, Y., Chen, H., Dubin, M., Lee, C.R., Wang, C., Bemm, F., Becker, C., O'Neil, R., O'Malley, R.C.C., Quarless, D.X.X., *et al.* (2016) 'Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions', *Cell*, 166(2), pp. 492–505.

Kawauchi, D., Robinson, G., Uziel, T., Gibson, P., Rehg, J., Gao, C., Finkelstein, D., Qu, C., Pounds, S., Ellison, D.W., Gilbertson, R.J. and Roussel, M.F. (2012) 'A mouse model of the most aggressive subgroup of human medulloblastoma.', *Cancer cell*, 21(2), pp. 168–80.

Kennedy, K.M. and Dewhirst, M.W. (2010) 'Tumor metabolism of lactate: the influence and therapeutic potential for MCT and CD147 regulation.', *Future oncology (London, England)*, 6(1), pp. 127–48.

KERSHMAN, J. (1938) 'THE MEDULLOBLAST AND THE MEDULLOBLASTOMA', *Archives of Neurology & Psychiatry*, 40(5), p. 937.

Khan, R.B., Patay, Z., Klimo, P., Huang, J., Kumar, R., Boop, F.A., Raches, D., Conklin, H.M., Sharma, R., Simmons, A., Sadighi, Z.S., Onar-Thomas, A., Gajjar, A. and Robinson, G.W. (2021) 'Clinical features, neurologic recovery, and risk factors of postoperative posterior fossa syndrome and delayed recovery: A prospective study', *Neuro-Oncology*, 23(9), pp. 1586–1596.

Khanna, V., Achey, R.L., Ostrom, Q.T., Block-Beach, H., Kruchko, C., Barnholtz-Sloan, J.S. and de Blank, P.M. (2017) 'Incidence and survival trends for medulloblastomas in the United States from 2001 to 2013', *Journal of Neuro-Oncology*, 135(3), pp. 433–441.

Khatua, S. (2016) 'Evolving molecular era of childhood medulloblastoma: time to revisit therapy', *Future Oncology*, 12(1), pp. 107–117.

Kijima, N. and Kanemura, Y. (2016) 'Molecular Classification of Medulloblastoma.', *Neurologia medico-chirurgica*, 56(11), pp. 687–697.

Kilaru, V., Knight, A.K., Katrinli, S., Cobb, D., Lori, A., Gillespie, C.F., Maihofer, A.X., Nievergelt, C.M., Dunlop, A.L., Conneely, K.N. and Smith, A.K. (2020) 'Critical Evaluation of Copy Number Variant Calling Methods Using DNA Methylation', *Genetic epidemiology*, 44(2), p. 148.

Kline, C.N., Packer, R.J., Hwang, E.I., Raleigh, D.R., Braunstein, S., Raffel, C., Bandopadhyay, P., Solomon, D.A., Aboian, M., Cha, S. and Mueller, S. (2017) 'Case-based review: pediatric medulloblastoma', *Neuro-Oncology Practice*, 4(3), pp. 138–150.

Knudson, A.G. and Jr. (1971) 'Mutation and cancer: statistical study of retinoblastoma.', *Proceedings of the National Academy of Sciences of the United States of America*, 68(4), pp. 820–3.

Koeller, K.K. and Rushing, E.J. (2003) 'From the Archives of the AFIP', *RadioGraphics*, 23(6), pp. 1613–1637.

Kongkham, P.N., Northcott, P.A., Croul, S.E., Smith, C.A., Taylor, M.D. and Rutka, J.T. (2010) 'The SFRP family of WNT inhibitors function as novel tumor suppressor genes epigenetically silenced in medulloblastoma', *Oncogene*, 29(20), pp. 3017–3024.

Konopka, T. (2020). umap: Uniform Manifold Approximation and Projection. R package version 0.2.7.0.

Kool, M., Korshunov, A., Remke, M., Jones, D.T.W., Schlanstein, M., Northcott, P.A., Cho, Y.-J., Koster, J., Schouten-van Meeteren, A., van Vuurden, D., Clifford, S.C., Pietsch, T., von Bueren, A.O., Rutkowski, S., McCabe, M., Collins, V.P., Bäcklund, M.L., Haberler, C., Bourdeaut, F., *et al.* (2012) 'Molecular subgroups of medulloblastoma: an international meta-analysis of transcriptome, genetic

aberrations, and clinical data of WNT, SHH, Group 3, and Group 4 medulloblastomas', *Acta Neuropathologica*, 123(4), pp. 473–484.

Koral, K., Gargan, L., Bowers, D.C., Gimi, B., Timmons, C.F., Weprin, B. and Rollins, N.K. (2008) 'Imaging Characteristics of Atypical Teratoid–Rhabdoid Tumor in Children Compared with Medulloblastoma', *American Journal of Roentgenology*, 190(3), pp. 809–814.

Korshunov, A., Chavez, L., Northcott, P.A., Sharma, T., Ryzhova, M., Jones, D.T.W., von Deimling, A., Pfister, S.M. and Kool, M. (2017) 'DNA-methylation profiling discloses significant advantages over NanoString method for molecular classification of medulloblastoma', *Acta Neuropathologica*, 134(6), pp. 965–967.

Korshunov, A., Sahm, F., Stichel, D., Schrimpf, D., Ryzhova, M., Zheludkova, O., Golanov, A., Lichter, P., Jones, D.T.W., von Deimling, A., Pfister, S.M. and Kool, M. (2018) 'Molecular characterization of medulloblastomas with extensive nodularity (MBEN)', *Acta Neuropathologica*, 136(2), pp. 303–313.

Koschmann, C., Bloom, K., Upadhyaya, S., Geyer, J.R. and Leary, S.E.S. (2016) 'Survival After Relapse of Medulloblastoma', *Journal of Pediatric Hematology/Oncology*, 38(4), pp. 269–273.

Kouzarides, T. (2007) 'Chromatin Modifications and Their Function', *Cell*, 128(4), pp. 693–705.

Krijthe, J. (2021) *GitHub - jkrijthe/Rtsne: R wrapper for Van der Maaten's Barnes-Hut implementation of t-Distributed Stochastic Neighbor Embedding* (no date). Available at: <https://github.com/jkrijthe/Rtsne> (Accessed: 10 January 2022).

Kristensen, B.W., Priesterbach-Ackley, L.P., Petersen, J.K. and Wesseling, P. (2019) 'Molecular pathology of tumors of the central nervous system', *Annals of Oncology*, 30(8), pp. 1265–1278.

Krueger, F. and Andrews, S.R. (2011) 'Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.', *Bioinformatics (Oxford, England)*, 27(11), pp. 1571–2.

Kuhn, M. (2021). caret: Classification and Regression Training. R package version 6.0-88.

Kulis, M. and Esteller, M. (2010) 'DNA Methylation and Cancer', in *Advances in genetics*, pp. 27–56.

Kuschel, L.P., Hench, J., Frank, S., Hench, I.B., Girard, E., Blanluet, M., Masliah-Planchon, J., Misch, M., Onken, J., Czabanka, M., Karau, P., Ishaque, N., Hain, E.G., Heppner, F., Idbaih, A., Behr, N., Harms, C., Capper, D. and Euskirchen, P. (2021) 'Robust methylation-based classification of brain tumors using nanopore sequencing', *medRxiv*, p. 2021.03.06.21252627.

---

Lafay-Cousin, L., Bouffet, E., Hawkins, C., Amid, A., Huang, A. and Mabbott, D.J. (2009) 'Impact of radiation avoidance on survival and neurocognitive outcome in infant medulloblastoma.', *Current oncology (Toronto, Ont.)*, 16(6), pp. 21–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20016743> (Accessed: 15 January 2019).

International Human Genome Sequencing Consortium: Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921

Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature methods*, 9(4), p. 357.

Längst, G. and Manelyte, L. (2015) 'Chromatin Remodelers: From Function to Dysfunction', *Genes*, 6(2), p. 299.

Lapidot, T., Sirard, C., Vormoor, J., Murdoch, B., Hoang, T., Caceres-Cortes, J., Minden, M., Paterson, B., Caligiuri, M.A. and Dick, J.E. (1994) 'A cell initiating human acute myeloid leukaemia after transplantation into SCID mice', *Nature*, 367(6464), pp. 645–648.

Lassaletta, A. and Ramaswamy, V. (2016) 'Medulloblastoma in adults: they're not just big kids', *Neuro-Oncology*, 18(7), pp. 895–897.

Laufer, B.I., Hwang, H., Jianu, J.M., Mordaunt, C.E., Korf, I.F., Hertz-Picciotto, I. and Lasalle, J.M. (2021) 'Low-pass whole genome bisulfite sequencing of neonatal dried blood spots identifies a role for RUNX1 in Down syndrome DNA methylation profiles', *Human Molecular Genetics*, 29(21), pp. 3465–3476.

Laufer, B.I., Hwang, H., Jianu, J.M., Mordaunt, C.E., Korf, I.F., Hertz-Picciotto, I. and Lasalle, J.M. (2021) 'Low-pass whole genome bisulfite sequencing of neonatal dried blood spots identifies a role for RUNX1 in Down syndrome DNA methylation profiles', *Human molecular genetics*, 29(21), pp. 3465–3476.

Lawrence, M., Daujat, S. and Schneider, R. (2016) 'Lateral Thinking: How Histone Modifications Regulate Gene Expression', *Trends in Genetics*, 32(1), pp. 42–56.

Legendre, C., Gooden, G.C., Johnson, K., Martinez, R.A., Liang, W.S. and Salhia, B. (2015) 'Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer.', *Clinical epigenetics*, 7(1), p. 100.

Lena, P. Di, Sala, C., Prodi, A. and Nardini, C. (2020) 'Methylation data imputation performances under different representations and missingness patterns', *BMC Bioinformatics*, 21(1).

Lever, J., Krzywinski, M. and Altman, N. (2017) 'Points of Significance: Principal component analysis', *Nature Methods*, 14(7), pp. 641–642.

Li, E. and Zhang, Y. (2014) 'DNA methylation in mammals.', *Cold Spring Harbor perspectives in biology*, 6(5), p. a019133.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079.

Li, K.K.-W., Lau, K.-M. and Ng, H.-K. (2013) 'Signaling pathway and molecular subgroups of medulloblastoma.', *International journal of clinical and experimental pathology*, 6(7), pp. 1211–22.

Li, X., Deng, W., Lobo-Ruppert, S.M. and Ruppert, J.M. (2007) 'Gli1 acts through Snail and E-cadherin to promote nuclear signaling by beta-catenin.', *Oncogene*, 26(31), pp. 4489–98.

Li, X., He, S. and Ma, B. (2020) 'Autophagy and autophagy-related proteins in cancer', *Molecular Cancer* 2020 19:1, 19(1), pp. 1–16.

Li, Y.-W., Kong, F.-M., Zhou, J.-P. and Dong, M. (2014) 'Aberrant promoter methylation of the vimentin gene may contribute to colorectal carcinogenesis: a meta-analysis', *Tumor Biology*, 35(7), pp. 6783–6790.

Lin, C.Y., Erkek, S., Tong, Y., Yin, L., Federation, A.J., Zapatka, M., Haldipur, P., Kawauchi, D., Risch, T., Warnatz, H.-J., Worst, B.C., Ju, B., Orr, B.A., Zeid, R., Polaski, D.R., Segura-Wang, M., Waszak, S.M., Jones, D.T.W., Kool, M., *et al.* (2016) 'Active medulloblastoma enhancers reveal subgroup-specific cellular origins.', *Nature*, 530(7588), pp. 57–62.

Lindsey, J.C., Michalski, A., Pizer, B., Bailey, S., Bown, N., Cuthbert, G., Wharton, S.B., Jacques, T.S., Joshi, A. and Clifford, S.C. (2021) 'Advanced molecular pathology for rare tumours: A national feasibility study and model for centralised medulloblastoma diagnostics', *Neuropathology and Applied Neurobiology*, 47(6), p. 736.

Liu, H., Zhou, J., Tian, W., Luo, C., Bartlett, A., Aldridge, A., Lucero, J., Osteen, J.K., Nery, J.R., Chen, H., Rivkin, A., Castanon, R.G., Clock, B., Li, Y.E., Hou, X., Poirion, O.B., Preissl, S., Pinto-Duarte, A., O'Connor, C., *et al.* (2021) 'DNA methylation atlas of the mouse brain at single-cell resolution', *Nature* 2021 598:7879, 598(7879), pp. 120–128.

Liu, Y., Siegmund, K.D., Laird, P.W. and Berman, B.P. (2012) 'Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data', *Genome Biology*, 13(7), pp. 1–14.

Lo Muzio, L. (2008) 'Nevoid basal cell carcinoma syndrome (Gorlin syndrome)', *Orphanet Journal of Rare Diseases*, 3(1), p. 32.

Locke, W. J., Guanzon, D., Ma, C., Liew, Y. J., Duesing, K. R., Fung, K. Y. C., & Ross, J. P. (2019). DNA Methylation Cancer Biomarkers: Translation to the Clinic. *Frontiers in Genetics*, 10, 1150. <https://doi.org/10.3389/FGENE.2019.01150/BIBTEX>

Louis, D.N., Ohgaki, H., Wiestler, O.D., Cavenee, W.K., Burger, P.C., Jouvet, A., Scheithauer, B.W. and Kleihues, P. (2007) 'The 2007 WHO classification of tumours of the central nervous system.', *Acta neuropathologica*, 114(2), pp. 97–109.

Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Webster, I.T., Cavenee, K., Ohgaki, H., Otmakhova, L., Wiestler, D., Kleihues, P., David, P. and Ellison, W. (no date) 'The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary'.

Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H.K., Pfister, S.M., Reifenberger, G., Soffietti, R., Von Deimling, A. and Ellison, D.W. (2021) 'The 2021 WHO Classification of Tumors of the Central Nervous System: a summary', *Neuro-oncology*, 23(8), pp. 1231–1251

Lövkvist, C., Dodd, I.B., Sneppen, K. and Haerter, J.O. (2016) 'DNA methylation in human epigenomes depends on local topology of CpG sites', *Nucleic acids research*, 44(11), pp. 5123–5132.

Lu, R., Duan, T., Wang, M., Liu, H., Feng, S., Gong, X., Wang, H., Wang, J., Cui, Z., Liu, Y., Li, C. and Ma, J. (2021) 'The application of multivariate adaptive regression splines in exploring the influencing factors and predicting the prevalence of HbA1c improvement', *Annals of Palliative Medicine*, 10(2), pp. 1296303–1291303.

Lu, Y., Labak, C.M., Jain, N., Purvis, I.J., Guda, M.R., Bach, S.E., Tsung, A.J., Asuthkar, S. and Velpula, K.K. (2017) 'OTX2 expression contributes to proliferation and progression in Myc-amplified medulloblastoma.', *American journal of cancer research*, 7(3), pp. 647–656.

Lu, Z., Zou, J., Li, S., Topper, M.J., Tao, Y., Zhang, H., Jiao, X., Xie, W., Kong, X., Vaz, M., Li, H., Cai, Y., Xia, L., Huang, P., Rodgers, K., Lee, B., Riemer, J.B., Day, C.P., Yen, R.W.C., *et al.* (2020) 'Epigenetic therapy inhibits metastases by disrupting premetastatic niches', *Nature*, 579(7798), pp. 284–290.

Luger, K., Dechassa, M.L. and Tremethick, D.J. (2012) 'New insights into nucleosome and chromatin structure: an ordered state or a disordered affair?', *Nature reviews. Molecular cell biology*, 13(7), pp. 436–47.

MacDonald, B.T., Tamai, K. and He, X. (2009) 'Wnt/beta-catenin signaling: components, mechanisms, and diseases.', *Developmental cell*, 17(1), pp. 9–26.

Maros, M.E., Capper, D., Jones, D.T.W., Hovestadt, V., von Deimling, A., Pfister, S.M., Benner, A., Zucknick, M. and Sill, M. (2020) 'Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data', *Nature Protocols 2020 15:2*, 15(2), pp. 479–512.

Martin, A.M., Raabe, E., Eberhart, C. and Cohen, K.J. (2014) 'Management of pediatric and adult patients with medulloblastoma.', *Current treatment options in oncology*, 15(4), pp. 581–94.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), pp. 10–12.

Mattocks, C.J., Morris, M.A., Matthijs, G., Swinnen, E., Corveleyn, A., Dequeker, E., Müller, C.R., Pratt, V. and Wallace, A. (2010) 'A standardized framework for the validation and verification of clinical molecular genetic tests', *European Journal of Human Genetics 2010 18:12*, 18(12), pp. 1276–1288.

McCabe, M.T., Brandes, J.C. and Vertino, P.M. (2009) 'Cancer DNA Methylation: Molecular Mechanisms and Clinical Implications', *Clinical cancer research: an official journal of the American Association for Cancer Research*, 15(12), p. 3927.

McInnes, L., Healy, J. and Melville, J. (2018) 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction'. Available at: <https://arxiv.org/abs/1802.03426v3> (Accessed: 31 January 2022).

McNeill, K.A. (2016) 'Epidemiology of Brain Tumors', *Neurologic Clinics*, 34(4), pp. 981–998.

Merkel, A., Fernández-Callejo, M., Casals, E., Marco-Sola, S., Schuyler, R., Gut, I.G. and Heath, S.C. (2019) 'gemBS: high throughput processing for DNA methylation data from bisulfite sequencing', *Bioinformatics (Oxford, England)*, 35(5), pp. 737–742.

Methyl-Seq DNA Library Kit (no date). Available at: <https://swiftbiosci.com/accel-ngs-methyl-seq-dna-library-kit-family/> (Accessed: 7 December 2021).

Meyer, B., Clifton, S., Locke, W., Luu, P.-L., Du, Q., Lam, D., Armstrong, N.J., Kumar, B., Deng, N., Harvey, K., Swarbrick, A., Ganju, V., Clark, S.J., Pidsley, R. and Stirzaker, C. (2021) 'Identification of DNA methylation biomarkers with potential to predict response to neoadjuvant chemotherapy in triple-negative breast cancer', *Clinical Epigenetics 2021 13:1*, 13(1), pp. 1–7.

Milde, T., Oehme, I., Korshunov, A., Kopp-Schneider, A., Remke, M., Northcott, P., Deubzer, H.E., Lodrini, M., Taylor, M.D., Von Deimling, A., Pfister, S. and Witt, O. (2010) 'HDAC5 and HDAC9 in Medulloblastoma: Novel Markers for Risk Stratification and Role in Tumor Cell Growth', *Diagnosis, Prognosis Clin Cancer Res*, 16(12).

Millard, N.E. and De Braganca, K.C. (2016) 'Medulloblastoma.', *Journal of child neurology*, 31(12), pp. 1341–53.

Miller, D.M., Thomas, S.D., Islam, A., Muench, D. and Sedoris, K. (2012) 'c-Myc and cancer metabolism.', *Clinical cancer research: an official journal of the American Association for Cancer Research*, 18(20), pp. 5546–53.

Miura, F., Enomoto, Y., Dairiki, R. and Ito, T. (2012) 'Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging', *Nucleic Acids Research*, 40(17).

Moran, S., Martínez-Cardús, A., Sayols, S., Musulén, E., Balañá, C., Estival-Gonzalez, A., Moutinho, C., Heyn, H., Diaz-Lagares, A., de Moura, M. C., Stella, G. M., Comoglio, P. M., Ruiz-Miró, M., Matias-Guiu, X., Pazo-Cid, R., Antón, A., Lopez-Lopez, R., Soler, G., Longo, F., ... Esteller, M. (2016). Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *The Lancet. Oncology*, 17(10), 1386–1395. [https://doi.org/10.1016/S1470-2045\(16\)30297-2](https://doi.org/10.1016/S1470-2045(16)30297-2)

Monjazeb, A.M., Zamora, A.E., Grossenbacher, S.K., Mirsoian, A., Sckisel, G.D. and Murphy, W.J. (2013) 'Immunoediting and antigen loss: overcoming the achilles heel of immunotherapy with antigen non-specific therapies.', *Frontiers in oncology*, 3, p. 197.



Moxon-Emre, I., Taylor, M.D., Bouffet, E., Hardy, K., Campen, C.J., Malkin, D., Hawkins, C., Laperriere, N., Ramaswamy, V., Bartels, U., Scantlebury, N., Janzen, L., Law, N., Walsh, K.S. and Mabbott, D.J. (2016) 'Intellectual Outcome in Molecular Subgroups of Medulloblastoma.', *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 34(34), pp. 4161–4170.

Müller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T. and Bock, C. (2019) 'RnBeads 2.0: Comprehensive analysis of DNA methylation data', *Genome Biology*, 20(1), pp. 1–12.

Nair, S.S., Luu, P.-L., Qu, W., Maddugoda, M., Huschtscha, L., Reddel, R., Chenevix-Trench, G., Toso, M., Kench, J.G., Horvath, L.G., Hayes, V.M., Stricker, P.D., Hughes, T.P. and Clark, S.J. (no date) 'Guidelines for whole genome bisulphite sequencing of intact and FFPE DNA on the Illumina HiSeq X Ten Epigenetics & Chromatin', *Epigenetics & Chromatin*, 10, p. 19.

National Cancer Institute. (2021). 'GDC Data Portal.' Available at: <https://portal.gdc.cancer.gov/repository> (Accessed 16th January 2022).

National Institute of Health. (2021). *Cell Biology and Cancer*. Available at: <https://science.education.nih.gov/supplements/webversions/CellBiology/guide/understanding1.html>. (Accessed 14th January 2022)

National Institute of Health. (2021). *Epigenomics*. Available: <http://commonfund.nih.gov/epigenomics/figure>. Last accessed 16th January 2022.

Negrini, S., Gorgoulis, V.G. and Halazonetis, T.D. (2010) 'Genomic instability — an evolving hallmark of cancer', *Nature Reviews Molecular Cell Biology*, 11(3), pp. 220–228.

Nelson, C.E., Hersh, B.M. and Carroll, S.B. (2004) 'The regulatory content of intergenic DNA shapes genome architecture', *Genome Biology*, 5(4), p. R25.

Newcastle University. *MAC: Methylation Array Classifier* (no date). Available at: <http://medullo.ncl.ac.uk:3838/mac/> (Accessed: 10 January 2022).

NHS (2022) *Cancer - NHS*. Available at: <https://www.nhs.uk/conditions/cancer/> (Accessed: 1 February 2022).

Northcott, P. A. *et al.* (2019) 'Medulloblastoma', *Nature Reviews Disease Primers* 2019 5:1. Nature Publishing Group, 5(1), pp. 1–20.

Northcott, P.A., Buchhalter, I., Morrissy, A.S., Hovestadt, V., Weischenfeldt, J., Ehrenberger, T., Gröbner, S., Segura-Wang, M., Zichner, T., Rudneva, V.A., Warnatz, H.J., Sidiropoulos, N., Phillips, A.H., Schumacher, S., Kleinheinz, K., Waszak, S.M., Erkek, S., Northcott, P.A., Hielscher, T., Dubuc, A., Mack, S., Shih, D., Remke, M., Al-Halabi, H., Albrecht, S., Jabado, N., Eberhart, C.G., Grajkowska, W., Weiss, W.A., Clifford, S.C., Bouffet, E., Rutka, J.T., Korshunov, A., Pfister, S. and Taylor, M.D. (2011) 'Pediatric and adult sonic hedgehog medulloblastomas are clinically and molecularly distinct', *Acta Neuropathologica*, 122(2), pp. 231–240.

Northcott, P.A., Dubuc, A.M., Pfister, S. and Taylor, M.D. (2012) 'Molecular subgroups of medulloblastoma.', *Expert review of neurotherapeutics*, 12(7), pp. 871–84.

Northcott, P.A., Lee, C., Zichner, T., Stütz, A.M., Erkek, S., Kawauchi, D., Shih, D.J.H., Hovestadt, V., Zapatka, M., Sturm, D., Jones, D.T.W., Kool, M., Remke, M., Cavalli, F.M.G., Zuyderduyn, S., Bader, G.D., VandenBerg, S., Esparza, L.A., Ryzhova, M., *et al.* (2014) 'Enhancer hijacking activates GF11 family oncogenes in medulloblastoma.', *Nature*, 511(7510), pp. 428–34.

Northcott, P.A., Shih, D.J.H., Peacock, J., Garzia, L., Morrissy, A.S., Zichner, T., Stütz, A.M., Korshunov, A., Reimand, J., Schumacher, S.E., Beroukhim, R., Ellison, D.W., Marshall, C.R., Lionel, A.C., Mack, S., Dubuc, A., Yao, Y., Ramaswamy, V., Luu, B., *et al.* (2012) 'Subgroup-specific structural variation across 1,000 medulloblastoma genomes.', *Nature*, 488(7409), pp. 49–56.

Northcott, P.A., Shih, D.J.H., Peacock, J., Garzia, L., Sorana Morrissy, A., Zichner, T., Stütz, A.M., Korshunov, A., Reimand, J., Schumacher, S.E., Beroukhim, R., Ellison, D.W., Marshall, C.R., Lionel, A.C., Mack, S., Dubuc, A., Yao, Y., Ramaswamy, V., Luu, B., *et al.* (2012) 'Subgroup specific structural variation across 1,000 medulloblastoma genomes', *Nature*, 488(7409), p. 49.

Northcott, P.A., Shih, D.J.H., Remke, M., Cho, Y.-J., Kool, M., Hawkins, C., Eberhart, C.G., Dubuc, A., Guettouche, T., Cardentey, Y., Bouffet, E., Pomeroy, S.L., Marra, M., Malkin, D., Rutka, J.T., Korshunov, A., Pfister, S. and Taylor, M.D. (2012) 'Rapid, reliable, and reproducible molecular sub-grouping of clinical medulloblastoma samples.', *Acta neuropathologica*, 123(4), pp. 615–26.

NovaSeq 6000 System (no date). Available at: <https://emea.illumina.com/systems/sequencing-platforms/novaseq.html> (Accessed: 7 December 2021).

Oh, S., Flynn, R.A., Floor, S.N., Purzner, J., Martin, L., Do, B.T., Schubert, S., Vaka, D., Morrissy, S., Li, Y., Kool, M., Hovestadt, V., Jones, D.T.W., Northcott, P.A., Risch, T., Warnatz, H.-J., Yaspo, M.-L., Adams, C.M., Leib, R.D., *et al.* (2016) 'Medulloblastoma-associated DDX3 variant selectively alters the translational response to stress.', *Oncotarget*, 7(19), pp. 28169–82.

Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R. V, Branco, M.R. and Reik, W. (no date) 'Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data'.

Olshen, A.B., Venkatraman, E.S., Lucito, R. and Wigler, M. (2004) 'Circular binary segmentation for the analysis of array-based DNA copy number data', *Biostatistics*, 5(4), pp. 557–572.

Ordóñez-Morán, P., Dafflon, C., Imajo, M., Nishida, E. and Huelsken, J. (2015) 'HOXA5 Counteracts Stem Cell Traits by Inhibiting Wnt Signaling in Colorectal Cancer', *Cancer cell*, 28(6), pp. 815–829.

Ordway, J.M., Bedell, J.A., Citek, R.W., Nunberg, A., Garrido, A., Kendall, R., Stevens, J.R., Cao, D., Doerge, R.W., Korshunova, Y., Holemon, H., McPherson, J.D., Lakey, N., Leon, J., Martienssen, R.A. and Jeddleloh, J.A. (2006) 'Comprehensive DNA methylation profiling in a human cancer genome identifies novel epigenetic targets', *Carcinogenesis*, 27(12), pp. 2409–2423.

Orr, B.A. (2020) 'Pathology, diagnostics, and classification of medulloblastoma', *Brain Pathology*, 30(3), pp. 664–678.

Ostrom, Q.T., Gittleman, H., Truitt, G., Boscia, A., Kruchko, C. and Barnholtz-Sloan, J.S. (2018) 'CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2011–2015', *Neuro-Oncology*, 20(suppl\_4), pp. iv1–iv86.

Ostrom, Q.T., Patil, N., Cioffi, G., Waite, K., Kruchko, C. and Barnholtz-Sloan, J.S. (2020) 'CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013–2017', *Neuro-Oncology*, 22(Suppl 1), p. iv1.

Oyharcabal-Bourden, V., Kalifa, C., Gentet, J.C., Frappaz, D., Edan, C., Chastagner, P., Sariban, E., Pagnier, A., Babin, A., Pichon, F., Neuenschwander, S., Vinchon, M., Bours, D., Mosseri, V., Le Gales, C., Ruchoux, M., Carrie, C. and Doz, F. (2005) 'Standard-risk medulloblastoma treated by adjuvant chemotherapy followed by reduced-dose craniospinal radiation therapy: a French Society of Pediatric Oncology Study', *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 23(21), pp. 4726–4734.

Packer, R.J. (2008) 'Childhood brain tumors: accomplishments and ongoing challenges.', *Journal of child neurology*, 23(10), pp. 1122–7.

Packer, R.J., Gurney, J.G., Punyko, J.A., Donaldson, S.S., Inskip, P.D., Stovall, M., Yasui, Y., Mertens, A.C., Sklar, C.A., Nicholson, H.S., Zeltzer, L.K., Neglia, J.P. and Robison, L.L. (2003) 'Long-Term Neurologic and Neurosensory Sequelae in Adult Survivors of a Childhood Brain Tumor: Childhood Cancer Survivor Study', *Journal of Clinical Oncology*, 21(17), pp. 3255–3261.

Packer, R.J., Sutton, L.N., Goldwein, J.W., Perilongo, G., Bunin, G., Ryan, J., Cohen, B.H., D'Angio, G., Kramer, E.D., Zimmerman, R.A., Rorke, L.B., Evans, A.E. and Schut, L. (1991) 'Improved survival with the use of adjuvant chemotherapy in the treatment of medulloblastoma', *Journal of Neurosurgery*, 74(3), pp. 433–440.

Pagès, F., Galon, J., Dieu-Nosjean, M.-C., Tartour, E., Sautès-Fridman, C. and Fridman, W.-H. (2010) 'Immune infiltration in human tumors: a prognostic factor that should not be ignored', *Oncogene*, 29(8), pp. 1093–1102.

Paska, A.V. and Hudler, P. (2015) 'Aberrant methylation patterns in cancer: a clinical view.', *Biochimica medica*, 25(2), pp. 161–76.

Pavlova, N.N. and Thompson, C.B. (2016) 'The Emerging Hallmarks of Cancer Metabolism.', *Cell metabolism*, 23(1), pp. 27–47.

Pei, Y., Moore, C.E., Wang, J., Tewari, A.K., Eroshkin, A., Cho, Y.-J., Witt, H., Korshunov, A., Read, T.-A., Sun, J.L., Schmitt, E.M., Miller, C.R., Buckley, A.F., McLendon, R.E., Westbrook, T.F., Northcott, P.A., Taylor, M.D., Pfister, S.M., Febbo, P.G., *et al.* (2012) 'An animal model of MYC-driven medulloblastoma.', *Cancer cell*, 21(2), pp. 155–67.

Pek, J.W. and Kai, T. (2011) 'DEAD-box RNA helicase Belle/DDX3 and the RNA interference pathway promote mitotic chromosome segregation.', *Proceedings of the National Academy of Sciences of the United States of America*, 108(29), pp. 12007–12.

Perez, E. and Capper, D. (2020) 'Invited Review: DNA methylation-based classification of paediatric brain tumours', *Neuropathology and Applied Neurobiology*, 46(1), pp. 28–47.

Perreault, S., Ramaswamy, V., Achrol, A.S., Chao, K., Liu, T.T., Shih, D., Remke, M., Schubert, S., Bouffet, E., Fisher, P.G., Partap, S., Vogel, H., Taylor, M.D., Cho, Y.J. and Yeom, K.W. (2014) 'MRI Surrogates for Molecular Subgroups of Medulloblastoma', *American Journal of Neuroradiology*, 35(7), pp. 1263–1269.

Peters, T.J., Buckley, M.J., Statham, A.L., Pidsley, R., Samaras, K., V Lord, R., Clark, S.J. and Molloy, P.L. (2015) 'De novo identification of differentially methylated regions in the human genome', *Epigenetics and Chromatin*, 8(1), pp. 1–16.

Pidsley, R., Lawrence, M.G., Zotenko, E., Niranjani, B., Statham, A., Song, J., Chabanon, R.M., Qu, W., Wang, H., Richards, M., Nair, S.S., Armstrong, N.J., Nim, H.T., Papargiris, M., Balanathan, P., French, H., Peters, T., Norden, S., Ryan, A., *et al.* (2018) 'Enduring epigenetic landmarks define the cancer microenvironment.', *Genome research*, 28(5), pp. 625–638.

Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., Van Dijk, S., Muhlhauser, B., Stirzaker, C. and Clark, S.J. (2016) 'Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling', *Genome Biology*, 17(1).

Pietsch, T., Taylor, M.D. and Rutka, J.T. (2004) 'Molecular pathogenesis of childhood brain tumors', *Journal of neuro-oncology*, 70(2), pp. 203–215.

Pleguezuelos-Manzano, C., Puschhof, J., Rosendahl Huber, A., van Hoeck, A., Wood, H.M., Nomburg, J., Gurjao, C., Manders, F., Dalmaso, G., Stege, P.B., Paganelli, F.L., Geurts, M.H., Beumer, J., Mizutani, T., Miao, Y., van der Linden, R., van der Elst, S., Ambrose, J.C., Arumugam, P., *et al.* (2020) 'Mutational signature in colorectal cancer caused by genotoxic pks+ *E. coli*', *Nature* 2020 580:7802, 580(7802), pp. 269–273.

QIAamp DNA FFPE Tissue Kit (no date). Available at: <https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/dna-purification/genomic-dna/qiaamp-dna-ffpe-tissue-kit/> (Accessed: 7 December 2021).

Qubit Assays | Thermo Fisher Scientific - UK (no date). Available at: <https://www.thermofisher.com/uk/en/home/industrial/spectroscopy-elemental-isotope-analysis/molecular-spectroscopy/fluorometers/qubit/> (Accessed: 7 December 2021).

Quinlan, A. and Rizzolo, D. (2017) 'Understanding medulloblastoma', *Journal of the American Academy of Physician Assistants*, 30(10), pp. 30–36.

R: The R Project for Statistical Computing (no date). Available at: <https://www.r-project.org/> (Accessed: 7 December 2021).

Raine, A., Liljedahl, U. and Nordlund, J. (2018) 'Data quality of whole genome bisulfite sequencing on Illumina platforms', *PLoS ONE*, 13(4).

Raman, L., Dheedene, A., De Smet, M., Van Dorpe, J. and Menten, B. (2019) 'WisecondorX: improved copy number detection for routine shallow whole-genome sequencing', *Nucleic Acids Research*, 47(4), pp. 1605–1614.

Ramaswamy, V., Remke, M., Adamski, J., Bartels, U., Tabori, U., Wang, X., Huang, A., Hawkins, C., Mabbott, D., Laperriere, N., Taylor, M.D. and Bouffet, E. (2016) 'Medulloblastoma subgroup-specific outcomes in irradiated children: who are the true high-risk patients?', *Neuro-oncology*, 18(2), pp. 291–7.

Rausch, T., Jones, D.T.W., Zapatka, M., Stütz, A.M., Zichner, T., Weischenfeldt, J., Jäger, N., Remke, M., Shih, D., Northcott, P.A., Pfaff, E., Tica, J., Wang, Q., Massimi, L., Witt, H., Bender, S., Pleier, S., Cin, H., Hawkins, C., *et al.* (2012) 'Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations.', *Cell*, 148(1–2), pp. 59–71.

*RefSeq: NCBI Reference Sequence Database* (no date). Available at: <https://www.ncbi.nlm.nih.gov/refseq/> (Accessed: 26 January 2022).

Reulecke, B.C., Erker, C.G., Fiedler, B.J., Niederstadt, T.-U. and Kurlemann, G. (2008) 'Brain Tumors in Children: Initial Symptoms and Their Influence on the Time Span Between Symptom Onset and Diagnosis', *Journal of Child Neurology*, 23(2), pp. 178–183.

Rini, B.I., Plimack, E.R., Stus, V., Gafanov, R., Hawkins, R., Nosov, D., Pouliot, F., Alekseev, B., Soulières, D., Melichar, B., Vynnychenko, I., Kryzhanivska, A., Bondarenko, I., Azevedo, S.J., Borchiellini, D., Szczylik, C., Markus, M., McDermott, R.S., Bedke, J., *et al.* (2019) 'Pembrolizumab plus Axitinib versus Sunitinib for Advanced Renal-Cell Carcinoma', *New England Journal of Medicine*, 380(12), pp. 1116–1127.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) 'limma powers differential expression analyses for RNA-sequencing and microarray studies', *Nucleic Acids Research*, 43(7), pp. e47–e47.

Robertson, P.L., Muraszko, K.M., Holmes, E.J., Sposto, R., Packer, R.J., Gajjar, A., Dias, M.S., Allen, J.C. and Children's Oncology Group (2006) 'Incidence and severity of postoperative cerebellar mutism syndrome in children with medulloblastoma: a prospective study by the Children's Oncology Group', *Journal of Neurosurgery: Pediatrics*, 105(6), pp. 444–451.

Robinson, G., Parker, M., Kranenburg, T.A., Lu, C., Chen, X., Ding, L., Phoenix, T.N., Hedlund, E., Wei, L., Zhu, X., Chalhoub, N., Baker, S.J., Huether, R., Kriwacki, R., Curley, N., Thiruvankatam, R., Wang, J., Wu, G., Rusch, M., *et al.* (2012) 'Novel mutations target distinct subgroups of medulloblastoma.', *Nature*, 488(7409), pp. 43–8.

Robinson, G.W., Orr, B.A., Wu, G., Gururangan, S., Lin, T., Qaddoumi, I., Packer, R.J., Goldman, S., Prados, M.D., Desjardins, A., Chintagumpala, M., Takebe, N., Kaste, S.C., Rusch, M., Allen, S.J., Onar-Thomas, A., Stewart, C.F., Fouladi, M., Boyett, J.M., *et al.* (2015) 'Vismodegib Exerts Targeted Efficacy Against Recurrent Sonic Hedgehog–Subgroup Medulloblastoma: Results From Phase II Pediatric Brain Tumor Consortium Studies PBTC-025B and PBTC-032', *Journal of Clinical Oncology*, 33(24), pp. 2646–2654.

Rocket HPC Service (high performance computing) | IT Service (NUIT) | Newcastle University (no date). Available at: <https://services.ncl.ac.uk/itservice/research/hpc/> (Accessed: 7 December 2021).

Roussel, M.F. and Robinson, G.W. (2013) 'Role of MYC in Medulloblastoma.', *Cold Spring Harbor perspectives in medicine*, 3(11), p. a014308.

Ruczinski, I., Kooperberg, C. and Leblanc, M. (2012) 'Logic Regression', <http://dx.doi.org/10.1198/1061860032238>, 12(3), pp. 475–511.

Ruthenburg, A.J., Allis, C.D. and Wysocka, J. (2007) 'Methylation of Lysine 4 on Histone H3: Intricacy of Writing and Reading a Single Epigenetic Mark', *Molecular Cell*, 25(1), pp. 15–30.

S220 Focused-ultrasonicator | Covaris (no date). Available at: <https://www.covaris.com/s220-focused-ultrasonicator-500217> (Accessed: 7 December 2021).

Sadi, A.M., Wang, D.-Y., Youngson, B.J., Miller, N., Boerner, S., Done, S.J. and Leong, W.L. (2011) 'Clinical relevance of DNA microarray analyses using archival formalin-fixed paraffin-embedded breast cancer specimens.', *BMC cancer*, 11, pp. 253:1–13.

Sahm, F., Schrimpf, D., Jones, D.T.W., Meyer, J., Kratz, A., Reuss, D., Capper, D., Koelsche, C., Korshunov, A., Wiestler, B., Buchhalter, I., Milde, T., Selt, F., Sturm, D., Kool, M., Hummel, M., Bewerunge-Hudler, M., Mawrin, C., Schüller, U., *et al.* (2016) 'Next-generation sequencing in routine brain tumor diagnostics enables an integrated diagnosis and identifies actionable targets', *Acta Neuropathologica*, 131(6), pp. 903–910.

---

Sample Quality Control, Electrophoresis, Bioanalyzer | Agilent (no date). Available at: <https://www.agilent.com/en/product/automated-electrophoresis/bioanalyzer-systems/bioanalyzer-instrument/2100-bioanalyzer-instrument-228250#productdetails> (Accessed: 7 December 2021).

Sarda, S., Das, A., Vinson, C. and Hannenhalli, S. (2017) 'Distal CpG islands can serve as alternative promoters to transcribe genes with silenced proximal promoters.', *Genome research*, 27(4), pp. 553–566.

Schröder, M. (2010) 'Human DEAD-box protein 3 has multiple functions in gene regulation and cell cycle control and is a prime target for viral manipulation', *Biochemical Pharmacology*, 79(3), pp. 297–306.

Schüller, U., Heine, V.M., Mao, J., Kho, A.T., Dillon, A.K., Han, Y.-G., Huillard, E., Sun, T., Ligon, A.H., Qian, Y., Ma, Q., Alvarez-Buylla, A., McMahon, A.P., Rowitch, D.H. and Ligon, K.L. (2008) 'Acquisition of granule neuron precursor identity is a critical determinant of progenitor cell competence to form Shh-induced medulloblastoma.', *Cancer cell*, 14(2), pp. 123–34.

Schumacher, A., Kapranov, P., Kaminsky, Z., Flanagan, J., Assadzadeh, A., Yau, P., Virtanen, C., Winegarden, N., Cheng, J., Gingeras, T. and Petronis, A. (2006) 'Microarray-based DNA methylation profiling: technology and applications.', *Nucleic acids research*, 34(2), pp. 528–42.

Schwalbe, E.C., Hicks, D., Rafiee, G., Bashton, M., Gohlke, H., Enshaei, A., Potluri, S., Matthiesen, J., Mather, M., Taleongpong, P., Chaston, R., Silmon, A., Curtis, A., Lindsey, J.C., Crosier, S., Smith, A.J., Goschzik, T., Doz, F., Rutkowski, S., *et al.* (2017) 'Minimal methylation classifier (MIMIC): A novel method for derivation and rapid diagnostic detection of disease-associated DNA methylation signatures', *Scientific Reports 2017 7:1*, 7(1), pp. 1–8.

Schwalbe, E.C., Lindsey, J.C., Nakjang, S., Crosier, S., Smith, A.J., Hicks, D., Rafiee, G., Hill, R.M., Iliasova, A., Stone, T., Pizer, B., Michalski, A., Joshi, A., Wharton, S.B., Jacques, T.S., Bailey, S., Williamson, D. and Clifford, S.C. (2017) 'Novel molecular subgroups for clinical classification and outcome prediction in childhood medulloblastoma: a cohort study', *The Lancet Oncology*, 18(7), pp. 958–971.

Schwalbe, E.C., Williamson, D., Lindsey, J.C., Hamilton, D., Ryan, S.L., Megahed, H., Garami, M., Hauser, P., Dembowska-Baginska, B., Perek, D., Northcott, P.A., Taylor, M.D., Taylor, R.E., Ellison, D.W., Bailey, S. and Clifford, S.C. (2013) 'DNA methylation profiling of medulloblastoma allows robust subclassification and improved outcome prediction using formalin-fixed biopsies.', *Acta neuropathologica*, 125(3), pp. 359–71.



Shao, C., Lacey, M., Dubeau, L. and Ehrlich, M. (2009) 'Hemimethylation footprints of DNA demethylation in cancer', *Epigenetics*, 4(3), pp. 165–175.

Sharma, S., Kelly, T.K. and Jones, P.A. (2010) 'Epigenetics in cancer.', *Carcinogenesis*, 31(1), pp. 27–36.

Sharma, T., Schwalbe, E.C., Williamson, D., Sill, M., Hovestadt, V., Mynarek, M., Rutkowski, S., Robinson, G.W., Gajjar, A., Cavalli, F., Ramaswamy, V., Taylor, M.D., Lindsey, J.C., Hill, R.M., Jäger, N., Korshunov, A., Hicks, D., Bailey, S., Kool, M., *et al.* (2019) 'Second-generation molecular subgrouping of medulloblastoma: an international meta-analysis of Group 3 and Group 4 subtypes', *Acta Neuropathologica*, 138(2), p. 309.

Shen, H. and Laird, P.W. (2013) 'Interplay between the cancer genome and epigenome.', *Cell*, 153(1), pp. 38–55.

Shen, W., Le, S., Li, Y. and Hu, F. (2016) 'SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation', *PLOS ONE*, 11(10), p. e0163962.

Simon, R. (2014) 'Class probability estimation for medical studies', *Biometrical journal. Biometrische Zeitschrift*, 56(4), pp. 597–600.

Simpson, F., Kerr, M.C. and Wicking, C. (2009) 'Trafficking, development and hedgehog', *Mechanisms of Development*, 126(5–6), pp. 279–288.

Singh, S.K., Clarke, I.D., Terasaki, M., Bonn, V.E., Hawkins, C., Squire, J. and Dirks, P.B. (2003) 'Identification of a cancer stem cell in human brain tumors.', *Cancer research*, 63(18), pp. 5821–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/14522905> (Accessed: 14 January 2019).

Skowron, P., Ramaswamy, V. and Taylor, M.D. (2015) 'Genetic and molecular alterations across medulloblastoma subgroups.', *Journal of molecular medicine (Berlin, Germany)*, 93(10), pp. 1075–84.

Smith, Z.D. and Meissner, A. (2013) 'DNA methylation: roles in mammalian development', *Nature Reviews Genetics*, 14(3), pp. 204–220.

Smoll, N.R. and Drummond, K.J. (2012) 'The incidence of medulloblastomas and primitive neuroectodermal tumours in adults and children', *Journal of Clinical Neuroscience*, 19(11), pp. 1541–1544.

Soto, J., Rodriguez-Antolin, C., Vallespín, E., de Castro Carpeño, J. and Ibanez de Caceres, I. (2016) 'The impact of next-generation sequencing on the DNA methylation-based translational cancer research', *Translational Research*, 169, pp. 1-18.e1.

Stark, Z., Dolman, L., Manolio, T. A., Ozenberger, B., Hill, S. L., Caulfield, M. J., Levy, Y., Glazer, D., Wilson, J., Lawler, M., Boughtwood, T., Braithwaite, J., Goodhand, P., Birney, E., & North, K. N. (2019). Integrating Genomics into Healthcare: A Global Responsibility. *American Journal of Human Genetics*, 104(1), 13. <https://doi.org/10.1016/J.AJHG.2018.11.014>

Starobova, H. and Vetter, I. (2017) 'Pathophysiology of Chemotherapy-Induced Peripheral Neuropathy.', *Frontiers in molecular neuroscience*, 10, p. 174.

Stekhoven, D.J. and Bühlmann, P. (2012) 'MissForest—non-parametric missing value imputation for mixed-type data', *Bioinformatics*, 28(1), pp. 112–118.

Stewart, B.W., Wild, C., International Agency for Research on Cancer and World Health Organization (no date) *World cancer report 2014*. Available at: <http://publications.iarc.fr/Non-Series-Publications/World-Cancer-Reports/World-Cancer-Report-2014> (Accessed: 14 January 2019).

Stirzaker, C., Taberlay, P.C., Statham, A.L. and Clark, S.J. (2014) 'Mining cancer methylomes: prospects and challenges.', *Trends in genetics: TIG*, 30(2), pp. 75–84.

Sun, C., Mezzadra, R. and Schumacher, T.N. (2018) 'Regulation and Function of the PD-L1 Checkpoint', *Immunity*, 48(3), pp. 434–452

Sun, D., Xi, Y., Rodriguez, B., Park, H.J., Tong, P., Meong, M., Goodell, M.A. and Li, W. (2014) 'MOABS: Model based analysis of bisulfite sequencing data', *Genome Biology*, 15(2), pp. 1–12.

Sun, S., Zane, A., Fulton, C. and Philipoom, J. (2021) 'Statistical and bioinformatic analysis of hemimethylation patterns in non-small cell lung cancer', *BMC Cancer*, 21(1), pp. 1–17.

Sun, Z., Cunningham, J., Slager, S. and Kocher, J.-P. (2015) 'Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis.', *Epigenomics*, 7(5), pp. 813–28.

Suzuki, M., Liao, W., Wos, F., Johnston, A.D., DeGrazia, J., Ishii, J., Bloom, T., Zody, M.C., Germer, S. and Grealley, J.M. (2018) 'Whole-genome bisulfite sequencing with improved accuracy and cost.', *Genome research*, 28(9), pp. 1364–1371.

Tabori, U., Sung, L., Hukin, J., Laperriere, N., Crooks, B., Carret, A.-S., Silva, M., Odame, I., Mpofu, C., Strother, D., Wilson, B., Samson, Y., Bouffet, E. and Canadian Pediatric Brain Tumor Consortium (2005) 'Medulloblastoma in the second decade of life: A specific group with respect to toxicity and management', *Cancer*, 103(9), pp. 1874–1880.

Tait, D.M., Thornton-Jones, H., Bloom, H.J., Lemerle, J. and Morris-Jones, P. (1990) 'Adjuvant chemotherapy for medulloblastoma: the first multi-centre control trial of the International Society of Paediatric Oncology (SIOP I).', *European journal of cancer (Oxford, England: 1990)*, 26(4), pp. 464–9.

Talevich, E., Shain, A.H., Botton, T. and Bastian, B.C. (2016) 'CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing', *PLOS Computational Biology*, 12(4), p. e1004873.

Tamayo-Orrego, L. and Charron, F. (2019) 'Recent advances in SHH medulloblastoma progression: tumor suppressor mechanisms and the tumor microenvironment', *F1000Research*, 8.

Taudt, A., Roquis, D., Vidalis, A., Wardenaar, R., Johannes, F. and Colome-Tatché-Tatché, M. (2018) 'METHimpute: Imputation-guided construction of complete methylomes from WGBS data', *BMC Genomics*, 19(1), pp. 1–14.

Taylor, M.D., Northcott, P.A., Korshunov, A., Remke, M., Cho, Y.-J., Clifford, S.C., Eberhart, C.G., Parsons, D.W., Rutkowski, S., Gajjar, A., Ellison, D.W., Lichten, P., Gilbertson, R.J., Pomeroy, S.L., Kool, M. and Pfister, S.M. (2012) 'Molecular subgroups of medulloblastoma: the current consensus.', *Acta neuropathologica*, 123(4), pp. 465–72.

Teodoro, J.G., Evans, S.K. and Green, M.R. (2007) 'Inhibition of tumor angiogenesis by p53: a new role for the guardian of the genome', *Journal of Molecular Medicine*, 85(11), pp. 1175–1186.

Teot, L.A., Sposto, R., Khayat, A., Qualman, S., Reaman, G. and Parham, D. (2007) 'The Problems and Promise of Central Pathology Review: Development of a Standardized Procedure for the Children's Oncology Group', *Pediatric and Developmental Pathology*, 10(3), pp. 199–207.

Thienpont, B., Van Dyck, L. and Lambrechts, D. (2016) 'Tumors smother their epigenome', <http://dx.doi.org/10.1080/23723556.2016.1240549>, 3(6).

Thomas, S., Izard, J., Walsh, E., Batich, K., Chongsathidkiet, P., Clarke, G., Sela, D.A., Muller, A.J., Mullin, J.M., Albert, K., Gilligan, J.P., DiGuilio, K., Dilbarova, R., Alexander, W. and Prendergast, G.C. (2017) 'The Host Microbiome Regulates and Maintains Human Health: A Primer and Perspective for Non-Microbiologists', *Cancer Research*, 77(8), pp. 1783–1812.

Thompson, E.M., Hielscher, T., Bouffet, E., Remke, M., Luu, B., Gururangan, S., McLendon, R.E., Bigner, D.D., Lipp, E.S., Perreault, S., Cho, Y.-J., Grant, G., Kim, S.-K., Lee, J.Y., Rao, A.A.N., Giannini, C., Li, K.K.W., Ng, H.-K., Yao, Y., *et al.* (2016) 'Prognostic value of medulloblastoma extent of resection after accounting for molecular subgroup: a retrospective integrated clinical and molecular analysis', *The Lancet Oncology*, 17(4), pp. 484–495.

Tickle, C. and Towers, M. (2017) 'Sonic Hedgehog Signaling in Limb Development', *Frontiers in Cell and Developmental Biology*, 5(FEB), p. 14.

Timp, W., Bravo, H.C., McDonald, O.G., Goggins, M., Umbricht, C., Zeiger, M., Feinberg, A.P. and Irizarry, R.A. (2014) 'Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors', *Genome Medicine*, 6(8), pp. 1–11.

Titus, A.J., Houseman, E.A., Johnson, K.C. and Christensen, B.C. (2016) 'methyLiftOver: cross-platform DNA methylation data integration.', *Bioinformatics (Oxford, England)*, 32(16), pp. 2517–9.

Todoric, J., Antonucci, L. and Karin, M. (2016) 'Targeting Inflammation in Cancer Prevention and Therapy Chronic Inflammation and Invasive Tumour Growth: Wounds that Do Not Heal'.

Triche, T.J., Weisenberger, D.J., Van Den Berg, D., Laird, P.W. and Siegmund, K.D. (2013) 'Low-level processing of Illumina Infinium DNA Methylation BeadArrays', *Nucleic Acids Research*, 41(7), pp. e90–e90.

*TruSeq Methyl Capture EPIC Library Prep Kit | NGS Methyl-Seq libraries* (no date). Available at: <https://emea.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-methyl-capture-epic.html> (Accessed: 28 January 2022).

Valastyan, S. and Weinberg, R.A. (2011) 'Tumor metastasis: molecular insights and evolving paradigms.', *Cell*, 147(2), pp. 275–92.

Valenta, T., Hausmann, G. and Basler, K. (2012) 'The many faces and functions of  $\beta$ -catenin.', *The EMBO journal*, 31(12), pp. 2714–36.

Valtz, N.L., Hayes, T.E., Norregaard, T., Liu, S.M. and McKay, R.D. (1991) 'An embryonic origin for medulloblastoma.', *The New biologist*, 3(4), pp. 364–71

Van Der Maaten, L. and Hinton, G. (2008) 'Visualizing Data using t-SNE', *Journal of Machine Learning Research*, 9, pp. 2579–2605.

Van Paemel, R., De Koker, A., Vandeputte, C., Van Zogchel, L., Lammens, T., Laureys, G., Vandesompele, J., Schleiermacher, G., Chicard, M., Van Roy, N., Vicha, A., Tytgat, G.A.M., Callewaert, N., De Preter, K. and De Wilde, B. (2020) 'Minimally invasive classification of pediatric solid tumors using reduced representation bisulfite sequencing of cell-free DNA: a proof-of-principle study', *bioRxiv*, (7), p. 795047.

Vardhanabhuti, S., Jeng, X.J., Wu, Y. and Li, H. (2014) 'Parametric modeling of whole-genome sequencing data for CNV identification', *Biostatistics*, 15(3), pp. 427–441.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., *et al.* (2001) 'The Sequence of the Human Genome', *Science*, 291(5507), pp. 1304–1351.

Vidal, E., Sayols, S., Moran, S., Guillaumet-Adkins, A., Schroeder, M.P., Royo, R., Orozco, M., Gut, M., Gut, I., Lopez-Bigas, N., Heyn, H. and Esteller, M. (2017) 'A DNA methylation map of human cancer at single base-pair resolution', *Nature Publishing Group*, 36, pp. 5648–5657.

Vinay, D.S., Ryan, E.P., Pawelec, G., Talib, W.H., Stagg, J., Elkord, E., Lichtor, T., Decker, W.K., Whelan, R.L., Kumara, H.M.C.S., Signori, E., Honoki, K., Georgakilas, A.G., Amin, A., Helferich, W.G., Boosani, C.S., Guha, G., Ciriolo, M.R., Chen, S., *et al.* (2015) 'Immune evasion in cancer: Mechanistic basis and therapeutic strategies', *Seminars in Cancer Biology*, 35, pp. S185–S198.

Virani, Shama, Virani, Shami, Colacino, J.A., Kim, J.H. and Rozek, L.S. (2012) 'Cancer epigenetics: a brief review.', *ILAR journal*, 53(3–4), pp. 359–69.

Walsh, C.P., Chaillet, J.R. and Bestor, T.H. (1998) 'Transcription of IAP endogenous retroviruses is constrained by cytosine methylation', *Nature Genetics*, 20(2), pp. 116–117.

Wang, T., Guan, W., Lin, J., Boutaoui, N., Canino, G., Luo, J., Celedón, J.C. and Chen, W. (2015) 'A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data', *Epigenetics*, 10(7), pp. 662–669.

Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., Nolan, G.P., Bava, F.A. and Deisseroth, K. (2018) 'Three-dimensional intact-tissue sequencing of single-cell transcriptional states', *Science*, 361(6400).

Wang, X., Dubuc, A.M., Ramaswamy, V., Mack, S., Gendoo, D.M.A., Remke, M., Wu, X., Garzia, L., Luu, B., Cavalli, F., Peacock, J., López, B., Skowron, P., Zagzag, D., Lyden, D., Hoffman, C., Cho, Y.-

J., Eberhart, C., MacDonald, T., *et al.* (2015) 'Medulloblastoma subgroups remain stable across primary and metastatic compartments', *Acta Neuropathologica*, 129(3), pp. 449–457.

Wang, X., Laird, P.W., Hinoue, T., Groshen, S. and Siegmund, K.D. (2014) 'Non-specific filtering of beta-distributed data', *BMC Bioinformatics*, 15(1), p. 199.

Warburg, O., Wind, F. and Negelein, E. (1927) 'THE METABOLISM OF TUMORS IN THE BODY.', *The Journal of general physiology*, 8(6), pp. 519–30.

Ward, E., DeSantis, C., Robbins, A., Kohler, B. and Jemal, A. (2014) 'Childhood and adolescent cancer statistics, 2014', *CA: A Cancer Journal for Clinicians*, 64(2), pp. 83–103.

Waszak, S.M., Northcott, P.A., Buchhalter, I., Robinson, G.W., Sutter, C., Groebner, S., Grund, K.B., Brugières, L., Jones, D.T.W., Pajtler, K.W., Morrissy, A.S., Kool, M., Sturm, D., Chavez, L., Ernst, A., Brabetz, S., Hain, M., Zichner, T., Segura-Wang, M., *et al.* (2018) 'Spectrum and prevalence of genetic predisposition in medulloblastoma: a retrospective genetic study and prospective validation in a clinical trial cohort', *The Lancet Oncology*, 19(6), pp. 785–798.

Weishaupt, H., Johansson, P., Sundström, A., Lubovac-Pilav, Z., Olsson, B., Nelander, S. and Swartling, F.J. (2019) 'Batch-normalization of cerebellar and medulloblastoma gene expression datasets utilizing empirically defined negative control genes', *Bioinformatics*, 35(18), pp. 3357–3364.

Whole Genome Bisulfite Sequencing - Novogene (no date). Available at: <https://en.novogene.com/services/research-services/epigenome-sequencing/whole-genome-bisulfite-sequencing/#> (Accessed: 7 December 2021).

Wickham, H. (2016) 'ggplot2'. Cham: Springer International Publishing (Use R

Winkler, F. (2017) 'Hostile takeover: how tumours hijack pre-existing vascular environments to thrive', *The Journal of Pathology*, 242(3), pp. 267–272.

World Health Organization (2022) *Cancer*. Available at: <https://www.who.int/news-room/fact-sheets/detail/cancer> (Accessed: 30 January 2022).

Yang, Z.-J., Ellis, T., Markant, S.L., Read, T.-A., Kessler, J.D., Bourboulas, M., Schüller, U., Machold, R., Fishell, G., Rowitch, D.H., Wainwright, B.J. and Wechsler-Reya, R.J. (2008) 'Medulloblastoma Can Be Initiated by Deletion of Patched in Lineage-Restricted Progenitors or Stem Cells', *Cancer Cell*, 14(2), pp. 135–145.

Yang, Z., Jin, M., Zhang, Z., Lu, J. and Hao, K. (2017) 'Classification Based on Feature Extraction For Hepatocellular Carcinoma Diagnosis Using High-throughput Dna Methylation Sequencing Data', *Procedia Computer Science*, 107, pp. 412–417.

Yong, W.S., Hsu, F.M. and Chen, P.Y. (2016) 'Profiling genome-wide DNA methylation', *Epigenetics & Chromatin* 2016 9:1, 9(1), pp. 1–16.

Yuan, S., Norgard, R.J. and Stanger, B.Z. (2019) 'Cellular Plasticity in Cancer', *Cancer Discovery*, 9(7), pp. 837–851.

Zapotocky, M., Mata-Mbemba, D., Sumerauer, D., Liby, P., Lassaletta, A., Zamecnik, J., Krskova, L., Kyncl, M., Stary, J., Laughlin, S., Arnoldo, A., Hawkins, C., Tabori, U., Taylor, M.D., Bouffet, E., Raybaud, C. and Ramaswamy, V. (2018) 'Differential patterns of metastatic dissemination across medulloblastoma subgroups', *Journal of neurosurgery. Pediatrics*, 21(2), pp. 145–152.

Zeltzer, P.M., Boyett, J.M., Finlay, J.L., Albright, A.L., Rorke, L.B., Milstein, J.M., Allen, J.C., Stevens, K.R., Stanley, P., Li, H., Wisoff, J.H., Geyer, J.R., McGuire-Cullen, P., Stehens, J.A., Shurin, S.B. and Packer, R.J. (1999) 'Metastasis Stage, Adjuvant Treatment, and Residual Tumor Are Prognostic Factors for Medulloblastoma in Children: Conclusions From the Children's Cancer Group 921 Randomized Phase III Study', *Journal of Clinical Oncology*, 17(3), pp. 832–832.

Zeman, M.K. and Cimprich, K.A. (2014) 'Causes and consequences of replication stress.', *Nature cell biology*, 16(1), pp. 2–9.

Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., Stein, L.D. and Ferretti, V. (2019) 'The International Cancer Genome Consortium Data Portal', *Nature Biotechnology*, 37(4), pp. 367–369

Zhang, L., Xie, W.J., Liu, S., Meng, L., Gu, C. and Gao, Y.Q. (2017) 'DNA Methylation Landscape Reflects the Spatial Organization of Chromatin in Different Cells', *Biophysical journal*, 113(7), pp. 1395–1404.

Zhou, L., Ng, H.K., Drautz-Moses, D.I., Schuster, S.C., Beck, S., Kim, C., Chambers, J.C. and Loh, M. (2019) 'Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing', *Scientific Reports*, 9(1).

Zhuang, J., Widschwendter, M. and Teschendorff, A.E. (2012) 'A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform', *BMC Bioinformatics*, 13(1), pp. 1–14.

---

Zhukova, N., Ramaswamy, V., Remke, M., Pfaff, E., Shih, D.J.H., Martin, D.C., Castelo-Branco, P., Baskin, B., Ray, P.N., Bouffet, E., von Bueren, A.O., Jones, D.T.W., Northcott, P.A., Kool, M., Sturm, D., Pugh, T.J., Pomeroy, S.L., Cho, Y.-J., Pietsch, T., *et al.* (2013) 'Subgroup-Specific Prognostic Implications of *TP53* Mutation in Medulloblastoma', *Journal of Clinical Oncology*, 31(23), pp. 2927–2935.

Ziller, M.J., Hansen, K.D., Meissner, A. and Aryee, M.J. (2015) 'Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing.', *Nature methods*, 12(3), pp. 230–2, 1 p following 232.

Zou, L.S., Erdos, M.R., Taylor, D.L., Chines, P.S., Varshney, A., Parker, S.C.J., Collins, F.S. and Didion, J.P. (2018) 'BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues', *BMC Genomics*, 19(1), p. 390.



## **Chapter 8 – Appendix**

## 8.1 – Cohort data

### 8.1.1 – Sequencing quality control

#### 8.1.1.1 – WGBS Sequencing summary statistics – ICGC Cohort

Supplementary Table 2a: WGBS sequencing summary of clinical samples

Sample	Paired-end sequencing reads	Lambda phage non-conversion rate	Average CpG coverage	CpGs coverage $\geq 5$	CpGs coverage $\geq 10$	CpGs coverage $\geq 20$	Average CpG methylation rate	Average non-CpG methylation rate
ICGC_MB1	623,811,831	0.000353	27.88	0.951	0.924	0.759	0.742	0.001
ICGC_MB2	679,702,451	0.000304	34.23	0.956	0.939	0.841	0.742	0.001
ICGC_MB5	761,206,409	0.000266	29.56	0.954	0.932	0.773	0.765	0.004
ICGC_MB6	1,006,457,176	0.000245	47.64	0.960	0.953	0.923	0.786	0.001
ICGC_MB7	547,859,028	0.000201	25.75	0.950	0.910	0.689	0.706	0.002
ICGC_MB9	649,852,593	0.000221	30.49	0.948	0.905	0.750	0.655	0.001
ICGC_MB14	654,669,051	0.000190	34.38	0.952	0.922	0.783	0.643	0.001
ICGC_MB15	719,001,068	0.000224	29.71	0.956	0.936	0.790	0.776	0.003
ICGC_MB16	461,919,986	0.000218	24.66	0.951	0.903	0.644	0.640	0.001
ICGC_MB17	639,660,281	0.000154	27.23	0.954	0.934	0.771	0.748	0.002
ICGC_MB18	653,579,657	0.000157	31.14	0.955	0.937	0.819	0.617	0.001
ICGC_MB19	587,665,221	0.000248	29.10	0.954	0.927	0.771	0.748	0.001
ICGC_MB24	697,002,657	0.000225	27.40	0.954	0.926	0.737	0.729	0.005
ICGC_MB26	490,624,716	0.000305	25.90	0.932	0.877	0.679	0.714	0.002
ICGC_MB28	499,879,430	0.000212	23.87	0.951	0.917	0.683	0.717	0.001
ICGC_MB32	470,816,475	0.000167	25.52	0.947	0.891	0.633	0.677	0.002
ICGC_MB34	684,371,826	0.000243	32.55	0.955	0.938	0.819	0.736	0.001
ICGC_MB35	648,538,423	0.000229	33.69	0.956	0.941	0.850	0.731	0.001
ICGC_MB36	603,889,519	0.000226	24.91	0.953	0.910	0.600	0.615	0.001
ICGC_MB37	526,163,635	0.000267	27.98	0.936	0.888	0.720	0.709	0.001
ICGC_MB38	522,819,649	0.000228	26.47	0.953	0.921	0.715	0.762	0.003
ICGC_MB39	543,486,878	0.000149	26.94	0.955	0.930	0.731	0.726	0.002
ICGC_MB40	723,780,166	0.000261	33.48	0.956	0.943	0.871	0.738	0.002
ICGC_MB45	415,724,933	0.000243	23.56	0.939	0.874	0.613	0.685	0.001
ICGC_MB46	527,997,449	0.000181	28.28	0.943	0.888	0.681	0.737	0.002
ICGC_MB49	693,227,248	0.000216	29.59	0.956	0.939	0.820	0.709	0.001
ICGC_MB50	499,302,201	0.000223	27.76	0.948	0.905	0.712	0.648	0.000
ICGC_MB51	501,578,694	0.000218	27.79	0.927	0.867	0.681	0.727	0.001
ICGC_MB88	605,391,048	0.000204	26.50	0.952	0.920	0.708	0.679	0.001

<b>ICGC_MB90</b>	622,979,349	0.000221	28.03	0.952	0.917	0.713	0.673	0.002
<b>ICGC_MB95</b>	570,923,398	0.000229	24.36	0.950	0.908	0.640	0.705	0.003
<b>ICGC_MB107</b>	1,201,274,488	0.000218	38.72	0.956	0.941	0.853	0.686	0.001
<b>ICGC_MB112</b>	529,480,840	0.000217	21.46	0.943	0.877	0.551	0.664	0.001
<b>ICGC_MB113</b>	871,611,832	0.000220	30.15	0.954	0.918	0.718	0.665	0.003
<b>F1</b>	501,571,570	0.000255	24.94	0.949	0.917	0.721	0.759	0.001
<b>F2</b>	617,905,211	0.000206	28.15	0.954	0.938	0.814	0.773	0.001
<b>F3</b>	435,441,857	0.000228	18.33	0.949	0.886	0.420	0.761	0.001
<b>F4</b>	657,120,569	0.000268	29.33	0.955	0.941	0.841	0.776	0.001
<b>A1</b>	547,548,547	0.000259	22.93	0.917	0.807	0.550	0.723	0.014
<b>A2</b>	616,337,590	0.000277	27.66	0.949	0.917	0.752	0.750	0.012
<b>A3</b>	567,857,361	0.000286	26.07	0.945	0.900	0.700	0.744	0.012
<b>A4</b>	702,268,485	0.000245	31.71	0.952	0.928	0.808	0.749	0.011
<b>Minimum</b>	415,724,933	0.000149	18.33	0.917	0.807	0.420	0.615	0.000
<b>Average</b>	621,007,162	0.000231	28.47	0.950	0.914	0.730	0.715	0.003
<b>Maximum</b>	1,201,274,488	0.000353	47.64	0.960	0.953	0.923	0.786	0.014

## 8.1.2 – Methylation quality control

### 8.1.2.1 – NMB Methylation QC summary via RnBeads2.0

Sample_ID	Number of CpG Sites Covered	Mean Cytosine Coverage	No. CpGs with coverage $\geq 5$	No. CpGs with coverage $\geq 10$	No. CpGs with coverage $\geq 30$	No. CpGs with coverage $\geq 60$	Sites Masked with NA (5 Coverage Filter Applied)	Sites Masked with NA (3 Coverage Filter Applied)
<b>NMB_17</b>	26,586,855	6.27886559	16,438,339	4,527,018	14,354	5,308	12,553,586	5,146,677
<b>NMB_77</b>	26,733,652	6.807915133	18,137,956	5,552,429	16,741	6,015	10,664,301	4,023,884
<b>NMB_79</b>	26,785,108	6.759724807	18,154,790	5,457,494	15,181	5,493	10,954,705	4,036,484
<b>NMB_169</b>	26,408,687	6.274207461	15,975,444	4,119,770	57,447	37,375	12,534,493	5,249,066
<b>NMB_178</b>	26,806,819	7.430507924	19,048,803	6,960,309	21,299	6,748	9,982,634	3,819,716
<b>NMB_181</b>	26,409,438	6.376894162	16,486,049	4,295,771	49,288	38,986	12,233,771	4,975,057
<b>NMB_185</b>	26,882,443	6.966158582	18,285,961	5,900,001	19,925	6,859	11,178,852	4,227,338
<b>NMB_187</b>	26,915,152	7.086040569	18,953,356	6,078,388	19,346	6,983	10,343,566	3,765,899
<b>NMB_203</b>	26,886,886	6.901275626	18,428,986	5,803,352	17,290	6,248	10,443,129	3,918,522
<b>NMB_250</b>	27,016,146	7.816834977	19,992,410	8,135,706	24,690	6,183	9,748,030	3,536,307
<b>NMB_273</b>	26,696,034	6.546758893	17,140,816	4,838,300	21,240	11,413	11,833,758	4,682,422
<b>NMB_318</b>	26,440,301	6.463608754	16,050,019	4,857,876	37,871	5,700	12,539,272	5,391,869
<b>NMB_362</b>	26,743,882	6.622579474	17,509,363	5,143,802	17,509	6,461	11,558,581	4,509,850
<b>NMB_363</b>	26,771,580	6.385916446	17,260,329	4,562,290	15,123	5,613	11,937,556	4,585,606
<b>NMB_378</b>	26,790,271	6.579376371	17,373,067	5,111,123	21,846	6,922	11,640,699	4,541,187
<b>NMB_390</b>	26,786,479	6.860999312	18,174,770	5,785,345	18,314	6,356	11,032,396	4,262,627
<b>NMB_390_FFPE</b>	24,893,503	6.424897091	14,057,686	5,003,042	27,562	5,714	14,082,067	7,258,368
<b>NMB_416</b>	26,608,005	6.746878317	17,401,641	5,485,890	17,075	5,987	11,869,700	4,687,757

<b>NMB_433</b>	26,472,545	6.226843849	15,931,459	4,503,401	14,889	5,479	12,894,570	5,505,417
<b>NMB_436</b>	26,832,761	7.085801756	18,735,247	6,254,748	19,401	6,586	9,966,740	3,783,439
<b>NMB_529</b>	26,922,302	8.519824791	20,692,671	9,566,702	42,238	7,372	8,498,355	3,152,509
<b>NMB_549</b>	26,620,486	7.264023129	17,429,713	5,331,993	142,149	75,483	15,685,787	7,095,179
<b>NMB_610</b>	26,585,586	6.42453121	16,986,452	4,668,279	14,823	5,674	11,797,850	4,692,187
<b>NMB_725</b>	26,777,618	6.354703208	17,150,555	4,396,671	15,600	5,991	13,286,797	5,320,043
<b>NMB_729</b>	27,144,207	7.617011873	20,546,242	7,600,658	18,484	5,926	8,703,170	2,941,546
<b>NMB_733</b>	26,554,239	7.02696432	17,464,020	6,028,917	30,279	6,280	12,146,972	4,950,256
<b>NMB_758</b>	26,496,914	6.523205948	15,919,545	4,127,702	42,610	32,365	12,934,260	5,445,215
<b>NMB_769</b>	26,609,630	6.266817201	16,114,371	3,898,890	29,906	21,139	12,934,415	5,220,055
<b>NMB_772</b>	26,789,509	6.746752059	17,875,028	5,414,289	17,637	5,962	11,513,518	4,434,461
<b>NMB_774</b>	26,756,277	6.413380494	17,246,044	4,594,041	16,009	6,075	11,656,171	4,505,674
<b>NMB_787</b>	26,548,503	6.518779985	17,136,172	4,818,145	14,768	5,453	12,000,997	4,727,611
<b>NMB_803</b>	26,904,387	6.976522082	18,694,716	5,869,489	18,757	6,453	10,478,815	3,847,059
<b>NMB_858</b>	26,987,918	7.75348228	20,360,567	7,751,122	21,421	7,225	8,516,973	2,963,589
<b>NMB_867</b>	26,736,720	6.897012199	17,742,206	5,786,233	18,820	6,150	11,678,751	4,617,309
<b>NMB_870</b>	26,770,744	6.552362123	17,786,632	4,813,365	15,701	6,048	11,349,968	4,206,830
<b>NMB_890</b>	26,500,265	6.33615977	16,665,895	4,395,065	14,071	5,534	11,705,352	4,645,376

### 8.1.2.2 – ICGC methylation QC summary via RnBeads 2.0

Sample_ID	Number of CpG Sites Covered	Mean Cytosine Coverage	No. CpGs with coverage $\geq 5$	No. CpGs with coverage $\geq 10$	No. CpGs with coverage $\geq 30$	No. CpGs with coverage $\geq 60$	Sites Masked with NA
ICGC_A1	25,501,620	24.39562542	25,501,620	22,177,630	8,049,356	271,136	0
ICGC_A2	26,506,374	29.12012032	26,506,374	25,620,755	12,300,460	167,674	0
ICGC_A3	26,404,867	26.95651874	26,404,867	24,910,629	10,328,164	137,470	0
ICGC_A4	26,653,257	32.57667095	26,653,257	25,853,677	15,288,854	565,111	0
ICGC_F1	26,551,635	25.69125988	26,551,635	25,394,913	8,590,024	62,796	0
ICGC_F2	26,639,991	29.46192568	26,639,991	26,177,756	12,677,202	91,781	0
ICGC_F3	26,499,123	19.26177538	26,499,123	24,726,921	1,687,945	32,919	0
ICGC_F4	26,747,379	30.03018595	26,747,379	26,210,611	13,724,502	93,697	0
ICGC_MB1	26,569,816	29.2740427	26,569,816	25,790,459	12,417,461	169,142	0
ICGC_MB2	26,681,042	35.02786184	26,681,042	26,115,641	16,730,868	1,308,393	0
ICGC_MB5	26,625,869	30.37727595	26,625,869	25,903,991	12,613,604	612,284	0
ICGC_MB6	26,801,382	48.60023744	26,801,382	26,581,656	22,702,159	6,232,605	0
ICGC_MB7	26,568,081	26.49778368	26,568,081	25,184,757	9,460,681	252,673	0
ICGC_MB9	26,491,141	31.40323733	26,491,141	25,089,006	13,501,597	1,068,802	0
ICGC_MB14	26,564,945	35.75889146	26,564,945	25,704,765	15,132,288	2,572,119	0
ICGC_MB15	26,594,506	30.23855698	26,594,506	25,449,493	12,768,947	591,737	0
ICGC_MB16	26,596,573	25.34982353	26,596,573	24,971,879	8,058,461	379,198	0
ICGC_MB17	26,642,117	28.52756581	26,642,117	26,080,263	11,349,416	123,961	0
ICGC_MB18	26,653,107	31.909972	26,653,107	26,023,944	14,287,443	668,321	0
ICGC_MB19	26,641,708	30.37387318	26,641,708	25,890,801	12,944,289	589,988	0
ICGC_MB24	26,705,536	28.07686594	26,705,536	25,687,509	10,630,154	337,378	0
ICGC_MB26	25,983,078	27.10432806	25,983,078	24,177,774	10,216,849	244,272	0
ICGC_MB28	26,605,272	24.5552421	26,605,272	25,367,814	7,191,222	60,700	0

<b>ICGC_MB32</b>	26,414,794	26.50908018	26,414,794	24,717,770	8,891,994	611,287	0
<b>ICGC_MB34</b>	26,679,148	34.33558103	26,679,148	26,172,707	15,950,755	1,120,694	0
<b>ICGC_MB35</b>	26,705,695	35.15060469	26,705,695	26,279,804	17,579,554	807,527	0
<b>ICGC_MB36</b>	26,591,438	25.48934149	26,591,438	24,922,712	6,220,850	114,980	0
<b>ICGC_MB37</b>	26,131,156	29.79701862	26,131,156	24,792,395	12,966,514	260,363	0
<b>ICGC_MB38</b>	26,667,911	27.15444446	26,667,911	25,546,846	9,835,672	339,041	0
<b>ICGC_MB39</b>	26,717,768	27.57922982	26,717,768	25,766,837	9,998,409	340,449	0
<b>ICGC_MB40</b>	26,668,310	34.29185123	26,668,310	26,217,531	17,437,689	419,500	0
<b>ICGC_MB45</b>	26,101,958	24.52459724	26,101,958	24,007,367	7,795,472	130,561	0
<b>ICGC_MB46</b>	26,342,637	29.83820333	26,342,637	24,773,823	11,615,470	1,098,457	0
<b>ICGC_MB49</b>	26,663,104	30.30732315	26,663,104	26,056,023	13,217,825	306,600	0
<b>ICGC_MB50</b>	26,466,406	28.97450957	26,466,406	25,239,028	11,187,106	485,046	0
<b>ICGC_MB51</b>	25,873,535	29.26243128	25,873,535	23,981,920	11,708,765	744,065	0
<b>ICGC_MB88</b>	26,630,903	27.21376876	26,630,903	25,438,922	9,851,015	302,141	0
<b>ICGC_MB90</b>	26,581,518	28.71199512	26,581,518	25,259,467	11,006,045	643,516	0
<b>ICGC_MB95</b>	26,536,008	25.07211966	26,536,008	25,051,775	7,617,399	302,767	0
<b>ICGC_MB107</b>	26,784,259	39.60073407	26,784,259	26,244,589	18,050,218	3,329,892	0
<b>ICGC_MB112</b>	26,348,038	22.27310049	26,348,038	24,185,621	5,339,355	68,980	0
<b>ICGC_MB113</b>	26,645,016	31.54009331	26,645,016	25,645,261	12,143,616	1,602,015	0

## 8.1.3 – Imputation

### 8.1.3.1 - BoostMe model performance – ICGC Cohort – Effect of downsampling on RMSE

Root mean squared error for all per sample imputation models at each level of sequencing depth and coverage filter tested.

Key: 10x5 = 10x sequencing depth processed using a 5x filter

Sample_ID	30x	10x5	10x3	9x5	9x3	8x5	8x3	7x5	7x3	6x5	6x3	5x3	5x1	4x3	4x1	3x1	2x1	1x1
ICGC_A1	0.12	0.14	0.15	0.14	0.15	0.14	0.15	0.14	0.16	0.15	0.16	0.17	0.20	0.17	0.21	0.23	0.25	0.27
ICGC_A2	0.11	0.13	0.14	0.14	0.14	0.14	0.15	0.14	0.15	0.15	0.16	0.16	0.18	0.17	0.20	0.21	0.24	0.27
ICGC_A3	0.11	0.13	0.14	0.13	0.14	0.14	0.14	0.14	0.15	0.14	0.15	0.16	0.18	0.17	0.20	0.21	0.24	0.26
ICGC_A4	0.10	0.13	0.13	0.13	0.13	0.13	0.14	0.14	0.14	0.14	0.15	0.15	0.17	0.16	0.19	0.21	0.23	0.26
ICGC_F1	0.10	0.13	0.14	0.13	0.14	0.14	0.14	0.14	0.15	0.15	0.16	0.16	0.18	0.17	0.20	0.22	0.25	0.27
ICGC_F2	0.10	0.13	0.13	0.13	0.13	0.13	0.14	0.14	0.14	0.14	0.15	0.16	0.17	0.17	0.19	0.21	0.24	0.27
ICGC_F3	0.11	0.14	0.15	0.15	0.16	0.15	0.16	0.15	0.17	0.16	0.17	0.18	0.21	0.18	0.23	0.25	0.27	0.29
ICGC_F4	0.10	0.12	0.13	0.13	0.13	0.13	0.14	0.13	0.14	0.14	0.15	0.15	0.17	0.16	0.19	0.21	0.23	0.27
ICGC_MB1	0.12	0.14	0.15	0.15	0.15	0.15	0.16	0.15	0.16	0.16	0.17	0.17	0.19	0.18	0.21	0.22	0.25	0.28
ICGC_MB2	0.11	0.13	0.13	0.13	0.13	0.13	0.14	0.14	0.14	0.14	0.15	0.15	0.17	0.16	0.18	0.20	0.22	0.26
ICGC_MB5	0.10	0.13	0.13	0.13	0.13	0.13	0.14	0.14	0.14	0.14	0.15	0.16	0.18	0.16	0.19	0.21	0.24	0.27
ICGC_MB6	0.09	0.11	0.11	0.11	0.11	0.11	0.12	0.12	0.12	0.12	0.13	0.13	0.14	0.14	0.15	0.17	0.19	0.23
ICGC_MB7	0.11	0.14	0.14	0.14	0.15	0.14	0.15	0.15	0.16	0.15	0.16	0.17	0.19	0.17	0.21	0.22	0.25	0.28
ICGC_MB9	0.13	0.16	0.16	0.16	0.17	0.16	0.17	0.17	0.17	0.17	0.18	0.19	0.21	0.20	0.22	0.25	0.27	0.31
ICGC_MB14	0.16	0.18	0.19	0.19	0.19	0.19	0.20	0.19	0.20	0.20	0.21	0.21	0.24	0.22	0.25	0.27	0.30	0.33
ICGC_MB15	0.10	0.12	0.12	0.12	0.13	0.12	0.13	0.13	0.13	0.13	0.14	0.14	0.16	0.15	0.18	0.19	0.22	0.24
ICGC_MB16	0.15	0.17	0.18	0.18	0.18	0.18	0.19	0.19	0.19	0.19	0.20	0.21	0.23	0.21	0.25	0.27	0.29	0.32
ICGC_MB17	0.10	0.14	0.14	0.14	0.15	0.14	0.15	0.15	0.16	0.15	0.16	0.17	0.20	0.18	0.21	0.23	0.26	0.30
ICGC_MB18	0.14	0.16	0.17	0.17	0.17	0.17	0.18	0.18	0.18	0.18	0.19	0.19	0.21	0.20	0.23	0.25	0.28	0.31
ICGC_MB19	0.12	0.14	0.15	0.15	0.15	0.15	0.15	0.15	0.16	0.15	0.16	0.17	0.18	0.17	0.20	0.21	0.23	0.26



<b>ICGC_MB24</b>	0.11	0.13	0.14	0.13	0.14	0.14	0.15	0.14	0.15	0.14	0.15	0.16	0.18	0.17	0.20	0.22	0.24	0.26
<b>ICGC_MB26</b>	0.12	0.14	0.14	0.14	0.15	0.15	0.15	0.15	0.16	0.15	0.16	0.17	0.19	0.17	0.21	0.22	0.25	0.28
<b>ICGC_MB28</b>	0.13	0.15	0.16	0.16	0.16	0.16	0.17	0.16	0.17	0.17	0.18	0.18	0.21	0.19	0.22	0.24	0.27	0.29
<b>ICGC_MB32</b>	0.11	0.14	0.14	0.14	0.15	0.14	0.15	0.14	0.16	0.15	0.16	0.17	0.20	0.17	0.21	0.23	0.25	0.28
<b>ICGC_MB34</b>	0.12	0.15	0.15	0.15	0.15	0.15	0.16	0.16	0.16	0.16	0.17	0.17	0.19	0.18	0.20	0.22	0.25	0.28
<b>ICGC_MB35</b>	0.13	0.15	0.15	0.15	0.15	0.16	0.16	0.16	0.16	0.16	0.17	0.17	0.19	0.18	0.20	0.22	0.24	0.27
<b>ICGC_MB36</b>	0.18	0.21	0.21	0.21	0.22	0.21	0.22	0.22	0.22	0.22	0.23	0.24	0.26	0.24	0.28	0.30	0.32	0.35
<b>ICGC_MB37</b>	0.13	0.15	0.16	0.15	0.16	0.16	0.16	0.16	0.17	0.16	0.17	0.18	0.20	0.19	0.21	0.23	0.26	0.29
<b>ICGC_MB38</b>	0.10	0.12	0.13	0.12	0.13	0.13	0.13	0.13	0.14	0.13	0.14	0.15	0.17	0.15	0.19	0.20	0.23	0.25
<b>ICGC_MB39</b>	0.11	0.14	0.14	0.14	0.14	0.14	0.15	0.14	0.15	0.15	0.16	0.16	0.19	0.17	0.20	0.22	0.25	0.28
<b>ICGC_MB40</b>	0.11	0.13	0.13	0.13	0.14	0.14	0.14	0.14	0.15	0.15	0.15	0.16	0.17	0.17	0.18	0.20	0.23	0.27
<b>ICGC_MB45</b>	0.14	0.17	0.17	0.17	0.17	0.17	0.18	0.18	0.18	0.18	0.19	0.19	0.22	0.20	0.24	0.26	0.28	0.31
<b>ICGC_MB46</b>	0.10	0.12	0.13	0.12	0.13	0.13	0.14	0.13	0.14	0.13	0.14	0.15	0.18	0.16	0.19	0.21	0.23	0.25
<b>ICGC_MB49</b>	0.13	0.15	0.15	0.15	0.16	0.16	0.16	0.16	0.16	0.16	0.17	0.18	0.19	0.18	0.21	0.23	0.25	0.28
<b>ICGC_MB50</b>	0.14	0.16	0.17	0.16	0.17	0.17	0.17	0.17	0.18	0.17	0.18	0.19	0.21	0.20	0.23	0.25	0.27	0.30
<b>ICGC_MB51</b>	0.11	0.13	0.13	0.13	0.13	0.13	0.14	0.13	0.14	0.14	0.15	0.15	0.18	0.16	0.19	0.21	0.23	0.26
<b>ICGC_MB88</b>	0.13	0.16	0.16	0.16	0.17	0.17	0.17	0.17	0.18	0.17	0.18	0.19	0.21	0.20	0.23	0.25	0.27	0.30
<b>ICGC_MB90</b>	0.13	0.16	0.16	0.16	0.17	0.16	0.17	0.17	0.18	0.17	0.18	0.19	0.21	0.20	0.23	0.25	0.28	0.31
<b>ICGC_MB95</b>	0.12	0.15	0.15	0.15	0.16	0.15	0.16	0.16	0.17	0.16	0.17	0.18	0.21	0.19	0.23	0.25	0.27	0.30
<b>ICGC_MB107</b>	0.14	0.16	0.16	0.16	0.17	0.16	0.17	0.17	0.17	0.17	0.18	0.18	0.20	0.19	0.21	0.23	0.25	0.29
<b>ICGC_MB112</b>	0.15	0.17	0.18	0.17	0.18	0.18	0.19	0.18	0.19	0.18	0.20	0.20	0.24	0.21	0.25	0.27	0.29	0.32
<b>ICGC_MB113</b>	0.15	0.17	0.18	0.17	0.18	0.17	0.18	0.18	0.19	0.18	0.19	0.19	0.22	0.20	0.23	0.25	0.27	0.29

### 8.1.3.2 - BoostMe model performance – NMB Cohort – Effect of downsampling on RMSE

Root mean squared error for all per sample imputation models at each level of sequencing depth and coverage filter tested.

Key: 10x5 = 10x sequencing depth processed using a 5x filter

Sample_ID	30x	10x5	10x3	9x5	9x3	8x5	8x3	7x5	7x3	6x5	6x3	5x3	5x1	4x3	4x1	3x1	2x1	1x1
NMB_17	0.19	0.18	0.19	0.19	0.20	0.19	0.20	0.19	0.20	0.19	0.21	0.21	0.25	0.22	0.26	0.28	0.30	0.32
NMB_77	0.16	0.15	0.16	0.15	0.16	0.16	0.16	0.16	0.17	0.16	0.17	0.17	0.20	0.18	0.21	0.22	0.24	0.26
NMB_79	0.18	0.18	0.18	0.18	0.19	0.18	0.19	0.18	0.19	0.19	0.20	0.20	0.24	0.21	0.25	0.27	0.28	0.31
NMB_169	0.18	0.17	0.18	0.18	0.19	0.18	0.19	0.18	0.19	0.18	0.20	0.20	0.24	0.20	0.26	0.27	0.29	0.31
NMB_178	0.17	0.16	0.17	0.16	0.17	0.17	0.17	0.17	0.18	0.17	0.18	0.18	0.21	0.19	0.23	0.24	0.26	0.28
NMB_181	0.23	0.22	0.23	0.22	0.23	0.22	0.23	0.22	0.23	0.23	0.24	0.24	0.28	0.25	0.29	0.30	0.32	0.34
NMB_185	0.17	0.16	0.17	0.16	0.17	0.16	0.18	0.17	0.18	0.17	0.18	0.19	0.22	0.19	0.23	0.25	0.27	0.29
NMB_187	0.16	0.15	0.16	0.16	0.16	0.16	0.17	0.16	0.17	0.16	0.17	0.17	0.20	0.18	0.21	0.23	0.24	0.26
NMB_203	0.17	0.17	0.17	0.17	0.18	0.17	0.18	0.17	0.18	0.18	0.19	0.19	0.22	0.20	0.23	0.25	0.27	0.29
NMB_250	0.16	0.15	0.16	0.15	0.16	0.15	0.16	0.16	0.17	0.16	0.17	0.17	0.21	0.18	0.22	0.23	0.25	0.27
NMB_273	0.16	0.16	0.16	0.16	0.17	0.16	0.17	0.16	0.17	0.16	0.18	0.18	0.21	0.18	0.22	0.24	0.25	0.27
NMB_318	0.19	0.18	0.19	0.19	0.20	0.19	0.20	0.19	0.20	0.19	0.20	0.21	0.25	0.21	0.26	0.27	0.29	0.31
NMB_362	0.16	0.15	0.16	0.15	0.16	0.15	0.16	0.15	0.17	0.16	0.17	0.17	0.20	0.17	0.21	0.23	0.24	0.26
NMB_363	0.18	0.17	0.18	0.17	0.18	0.18	0.19	0.18	0.19	0.18	0.19	0.20	0.23	0.20	0.25	0.26	0.28	0.30
NMB_378	0.20	0.19	0.20	0.19	0.20	0.20	0.20	0.20	0.21	0.20	0.21	0.21	0.25	0.22	0.26	0.27	0.29	0.31
NMB_390	0.19	0.19	0.19	0.19	0.19	0.19	0.20	0.19	0.20	0.19	0.20	0.20	0.24	0.21	0.24	0.25	0.27	0.28
NMB_390_FFPE	0.18	0.18	0.18	0.18	0.18	0.18	0.19	0.19	0.19	0.19	0.19	0.20	0.22	0.20	0.23	0.24	0.26	0.28
NMB_416	0.15	0.14	0.15	0.14	0.16	0.15	0.16	0.15	0.16	0.15	0.16	0.17	0.20	0.17	0.21	0.23	0.24	0.26
NMB_433	0.19	0.18	0.19	0.18	0.19	0.18	0.19	0.19	0.20	0.19	0.20	0.20	0.24	0.21	0.25	0.26	0.27	0.29
NMB_436	0.18	0.18	0.18	0.18	0.19	0.18	0.19	0.19	0.19	0.19	0.20	0.20	0.23	0.21	0.24	0.25	0.27	0.29
NMB_529	0.19	0.18	0.19	0.19	0.19	0.19	0.20	0.19	0.20	0.19	0.20	0.21	0.24	0.21	0.25	0.27	0.29	0.32

<b>NMB_549</b>	0.23	0.22	0.23	0.21	0.22	0.22	0.22	0.22	0.23	0.22	0.23	0.23	0.27	0.24	0.28	0.29	0.31	0.33
<b>NMB_610</b>	0.21	0.20	0.21	0.21	0.21	0.21	0.21	0.21	0.22	0.22	0.22	0.23	0.26	0.23	0.27	0.28	0.30	0.32
<b>NMB_725</b>	0.18	0.17	0.18	0.17	0.18	0.17	0.18	0.18	0.19	0.18	0.19	0.19	0.23	0.20	0.24	0.26	0.27	0.29
<b>NMB_729</b>	0.18	0.17	0.18	0.17	0.18	0.17	0.18	0.18	0.19	0.18	0.19	0.20	0.23	0.20	0.24	0.26	0.28	0.31
<b>NMB_733</b>	0.17	0.16	0.17	0.16	0.17	0.16	0.17	0.16	0.17	0.16	0.17	0.18	0.21	0.18	0.22	0.23	0.25	0.27
<b>NMB_758</b>	0.22	0.22	0.22	0.22	0.23	0.22	0.23	0.22	0.23	0.23	0.24	0.24	0.28	0.25	0.29	0.30	0.32	0.34
<b>NMB_769</b>	0.21	0.20	0.21	0.21	0.22	0.21	0.22	0.21	0.22	0.22	0.23	0.23	0.27	0.24	0.29	0.30	0.32	0.34
<b>NMB_772</b>	0.17	0.16	0.17	0.16	0.17	0.16	0.17	0.16	0.17	0.17	0.18	0.18	0.21	0.19	0.22	0.24	0.25	0.27
<b>NMB_774</b>	0.19	0.18	0.19	0.18	0.19	0.19	0.19	0.19	0.20	0.19	0.20	0.21	0.24	0.21	0.25	0.27	0.29	0.31
<b>NMB_787</b>	0.16	0.15	0.16	0.15	0.16	0.15	0.16	0.15	0.16	0.15	0.16	0.17	0.20	0.17	0.21	0.22	0.23	0.25
<b>NMB_803</b>	0.18	0.17	0.18	0.17	0.18	0.17	0.18	0.18	0.19	0.18	0.19	0.19	0.22	0.20	0.23	0.25	0.26	0.28
<b>NMB_858</b>	0.18	0.17	0.18	0.18	0.18	0.18	0.18	0.18	0.19	0.18	0.19	0.20	0.22	0.20	0.23	0.25	0.26	0.29
<b>NMB_867</b>	0.15	0.15	0.15	0.15	0.16	0.15	0.16	0.15	0.16	0.15	0.16	0.17	0.20	0.17	0.21	0.22	0.24	0.26
<b>NMB_870</b>	0.19	0.18	0.19	0.18	0.19	0.19	0.19	0.19	0.20	0.19	0.20	0.20	0.23	0.21	0.25	0.26	0.27	0.30
<b>NMB_890</b>	0.17	0.16	0.17	0.16	0.17	0.16	0.17	0.16	0.17	0.17	0.18	0.18	0.21	0.18	0.22	0.23	0.24	0.26

### 8.1.3.3 – BoostMe model performance – ICGC cohort – Validation & test set performance for ICGC cohort (30x)

Sample ID	Validation Set				Test Set			
	RMSE	AUROC	AUPRC	Accuracy	RMSE	AUROC	AUPRC	Accuracy
ICGC_A1	0.12	0.98	1.00	0.96	0.12	0.98	1.00	0.96
ICGC_A2	0.11	0.99	1.00	0.96	0.11	0.99	1.00	0.96
ICGC_A3	0.11	0.99	1.00	0.96	0.11	0.99	1.00	0.96
ICGC_A4	0.11	0.99	1.00	0.96	0.11	0.99	1.00	0.96
ICGC_F1	0.10	0.99	1.00	0.96	0.10	0.99	1.00	0.96
ICGC_F2	0.10	0.99	1.00	0.97	0.10	0.99	1.00	0.97
ICGC_F3	0.11	0.99	1.00	0.96	0.11	0.99	1.00	0.96
ICGC_F4	0.10	0.99	1.00	0.97	0.10	0.99	1.00	0.97
ICGC_MB1	0.12	0.98	1.00	0.95	0.12	0.98	1.00	0.95
ICGC_MB2	0.11	0.99	1.00	0.96	0.11	0.99	1.00	0.96
ICGC_MB5	0.10	0.99	1.00	0.96	0.10	0.99	1.00	0.96
ICGC_MB6	0.09	0.99	1.00	0.97	0.09	0.99	1.00	0.97
ICGC_MB7	0.11	0.99	1.00	0.96	0.11	0.99	1.00	0.96
ICGC_MB9	0.13	0.97	0.99	0.92	0.13	0.97	0.99	0.92
ICGC_MB14	0.16	0.95	0.98	0.90	0.16	0.95	0.98	0.90
ICGC_MB15	0.10	0.99	1.00	0.97	0.10	0.99	1.00	0.97
ICGC_MB16	0.15	0.97	0.99	0.91	0.15	0.97	0.99	0.91
ICGC_MB17	0.10	0.98	1.00	0.96	0.10	0.98	1.00	0.96
ICGC_MB18	0.14	0.98	0.99	0.92	0.14	0.97	0.99	0.92
ICGC_MB19	0.12	0.98	1.00	0.96	0.12	0.98	1.00	0.95
ICGC_MB24	0.11	0.99	1.00	0.96	0.11	0.99	1.00	0.96
ICGC_MB26	0.12	0.98	1.00	0.95	0.12	0.99	1.00	0.95

<b>ICGC_MB28</b>	0.13	0.98	0.99	0.95	0.13	0.98	0.99	0.95
<b>ICGC_MB32</b>	0.11	0.99	1.00	0.96	0.11	0.99	1.00	0.96
<b>ICGC_MB34</b>	0.12	0.98	1.00	0.95	0.12	0.98	1.00	0.95
<b>ICGC_MB35</b>	0.13	0.98	0.99	0.95	0.13	0.98	0.99	0.95
<b>ICGC_MB36</b>	0.18	0.95	0.97	0.88	0.18	0.95	0.97	0.88
<b>ICGC_MB37</b>	0.13	0.98	0.99	0.94	0.13	0.98	0.99	0.94
<b>ICGC_MB38</b>	0.10	0.99	1.00	0.97	0.10	0.99	1.00	0.97
<b>ICGC_MB39</b>	0.11	0.99	1.00	0.96	0.11	0.99	1.00	0.96
<b>ICGC_MB40</b>	0.11	0.99	1.00	0.96	0.11	0.99	1.00	0.96
<b>ICGC_MB45</b>	0.14	0.97	0.99	0.93	0.14	0.97	0.99	0.93
<b>ICGC_MB46</b>	0.10	0.99	1.00	0.97	0.10	0.99	1.00	0.97
<b>ICGC_MB49</b>	0.13	0.98	0.99	0.95	0.13	0.98	0.99	0.94
<b>ICGC_MB50</b>	0.14	0.98	0.99	0.93	0.14	0.98	0.99	0.93
<b>ICGC_MB51</b>	0.10	0.99	1.00	0.96	0.11	0.99	1.00	0.96
<b>ICGC_MB88</b>	0.13	0.98	0.99	0.93	0.13	0.98	0.99	0.94
<b>ICGC_MB90</b>	0.13	0.97	0.99	0.93	0.13	0.97	0.99	0.93
<b>ICGC_MB95</b>	0.12	0.98	0.99	0.94	0.12	0.98	0.99	0.94
<b>ICGC_MB107</b>	0.14	0.97	0.99	0.93	0.14	0.97	0.99	0.93
<b>ICGC_MB112</b>	0.15	0.97	0.99	0.92	0.15	0.97	0.99	0.92
<b>ICGC_MB113</b>	0.15	0.97	0.99	0.93	0.15	0.97	0.99	0.93
<b>Average</b>	0.12	0.98	0.99	0.95	0.12	0.98	0.99	0.95

### 8.1.3.4 - BoostMe model performance – ICGC cohort – Validation & test set performance for ICGC cohort (10x – 3x filter)

Sample ID	Validation Set				Test Set			
	RMSE	AUROC	AUPRC	Accuracy	RMSE	AUROC	AUPRC	Accuracy
<b>NMB_17</b>	0.19	0.94	0.98	0.89	0.19	0.94	0.98	0.89
<b>NMB_77</b>	0.16	0.97	0.99	0.94	0.16	0.97	0.99	0.94
<b>NMB_79</b>	0.18	0.95	0.99	0.91	0.18	0.95	0.99	0.91
<b>NMB_169</b>	0.18	0.95	0.99	0.91	0.18	0.95	0.99	0.91
<b>NMB_178</b>	0.17	0.96	0.99	0.93	0.17	0.96	0.99	0.93
<b>NMB_181</b>	0.23	0.91	0.97	0.87	0.23	0.91	0.97	0.87
<b>NMB_185</b>	0.17	0.96	0.99	0.93	0.17	0.96	0.99	0.93
<b>NMB_187</b>	0.16	0.97	0.99	0.94	0.16	0.97	0.99	0.94
<b>NMB_203</b>	0.17	0.96	0.99	0.92	0.17	0.96	0.99	0.92
<b>NMB_250</b>	0.16	0.97	0.99	0.94	0.16	0.97	0.99	0.94
<b>NMB_273</b>	0.16	0.97	0.99	0.94	0.17	0.97	0.99	0.94
<b>NMB_318</b>	0.19	0.95	0.98	0.91	0.19	0.95	0.98	0.91
<b>NMB_362</b>	0.16	0.97	0.99	0.94	0.16	0.97	0.99	0.94
<b>NMB_363</b>	0.18	0.95	0.99	0.92	0.18	0.95	0.99	0.92
<b>NMB_378</b>	0.20	0.95	0.98	0.90	0.20	0.95	0.98	0.90
<b>NMB_390</b>	0.19	0.95	0.99	0.92	0.19	0.95	0.99	0.92
<b>NMB_390_FFPE</b>	0.18	0.96	0.99	0.93	0.18	0.96	0.99	0.93
<b>NMB_416</b>	0.15	0.97	0.99	0.95	0.15	0.97	0.99	0.95
<b>NMB_433</b>	0.19	0.95	0.99	0.92	0.19	0.95	0.99	0.91
<b>NMB_436</b>	0.18	0.95	0.99	0.92	0.18	0.95	0.99	0.92
<b>NMB_529</b>	0.19	0.95	0.98	0.90	0.19	0.95	0.98	0.90
<b>NMB_549</b>	0.23	0.93	0.97	0.87	0.23	0.93	0.97	0.87

---

<b>NMB_610</b>	0.21	0.95	0.98	0.89	0.21	0.95	0.98	0.89
<b>NMB_725</b>	0.18	0.96	0.99	0.92	0.18	0.96	0.99	0.92
<b>NMB_729</b>	0.18	0.96	0.98	0.91	0.17	0.96	0.98	0.91
<b>NMB_733</b>	0.17	0.97	0.99	0.94	0.17	0.97	0.99	0.94
<b>NMB_758</b>	0.22	0.94	0.96	0.87	0.22	0.94	0.96	0.87
<b>NMB_769</b>	0.21	0.94	0.96	0.87	0.21	0.94	0.96	0.87
<b>NMB_772</b>	0.17	0.97	0.99	0.93	0.17	0.97	0.99	0.93
<b>NMB_774</b>	0.19	0.96	0.98	0.91	0.19	0.96	0.98	0.90
<b>NMB_787</b>	0.16	0.97	0.99	0.95	0.16	0.97	0.99	0.95
<b>NMB_803</b>	0.18	0.96	0.99	0.93	0.18	0.96	0.99	0.93
<b>NMB_858</b>	0.18	0.96	0.99	0.92	0.18	0.96	0.99	0.92
<b>NMB_867</b>	0.15	0.97	0.99	0.95	0.15	0.97	0.99	0.95
<b>NMB_870</b>	0.19	0.96	0.99	0.92	0.19	0.96	0.99	0.92
<b>NMB_890</b>	0.17	0.97	0.99	0.94	0.17	0.97	0.99	0.94
<b>Average</b>	0.18	0.96	0.99	0.92	0.18	0.96	0.99	0.92

## 8.1.4 – Applying WGBS sample data to pre-existing array trained classifier models to predict the molecular subgroups of medulloblastoma

### 8.1.4.1 – 4-group classifier results

Coverage	Filter Applied	WGBS - ICGC		WGBS - NMB		Array - ICGC		Array - NMB	
		No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)
30x		10,000	<b>0.970 (0.057)</b>	NA	NA				
10x	5x	10,000	<b>0.968 (0.060)</b>	10,000	<b>0.967 (0.055)</b>				
	3x	10,000	<b>0.970 (0.055)</b>	10,000	<b>0.968 (0.057)</b>				
9x	5x	10,000	<b>0.967 (0.061)</b>	10,000	<b>0.964 (0.069)</b>				
	3x	10,000	<b>0.970 (0.055)</b>	10,000	<b>0.968 (0.055)</b>				
8x	5x	10,000	<b>0.964 (0.067)</b>	10,000	<b>0.958 (0.084)</b>				
	3x	10,000	<b>0.970 (0.055)</b>	10,000	<b>0.969 (0.049)</b>				
7x	5x	10,000	<b>0.960 (0.078)</b>	10,000	<b>0.956 (0.069)</b>	9,643	<b>0.964 (0.071)</b>	9,643	<b>0.949 (0.104)</b>
	3x	10,000	<b>0.969 (0.057)</b>	10,000	<b>0.969 (0.076)</b>				
6x	5x	10,000	<b>0.950 (0.100)</b>	10,000	<b>0.940 (0.101)</b>				
	3x	10,000	<b>0.968 (0.060)</b>	10,000	<b>0.968 (0.054)</b>				
5x	3x	10,000	<b>0.965 (0.065)</b>	10,000	<b>0.965 (0.068)</b>				
	1x	10,000	<b>0.971 (0.052)</b>	10,000	<b>0.963 (0.076)</b>				
4x	3x	10,000	<b>0.957 (0.083)</b>	10,000	<b>0.960 (0.072)</b>				
	1x	10,000	<b>0.971 (0.051)</b>	10,000	<b>0.965 (0.070)</b>				



<b>3x</b>	<b>1x</b>	10,000	<b>0.970 (0.053)</b>	10,000	<b>0.965 (0.065)</b>		
<b>2x</b>	<b>1x</b>	10,000	<b>0.967 (0.058)</b>	10,000	<b>0.967 (0.052)</b>		
<b>1x</b>	<b>1x</b>	10,000	<b>0.948 (0.093)</b>	10,000	<b>0.953 (0.086)</b>		

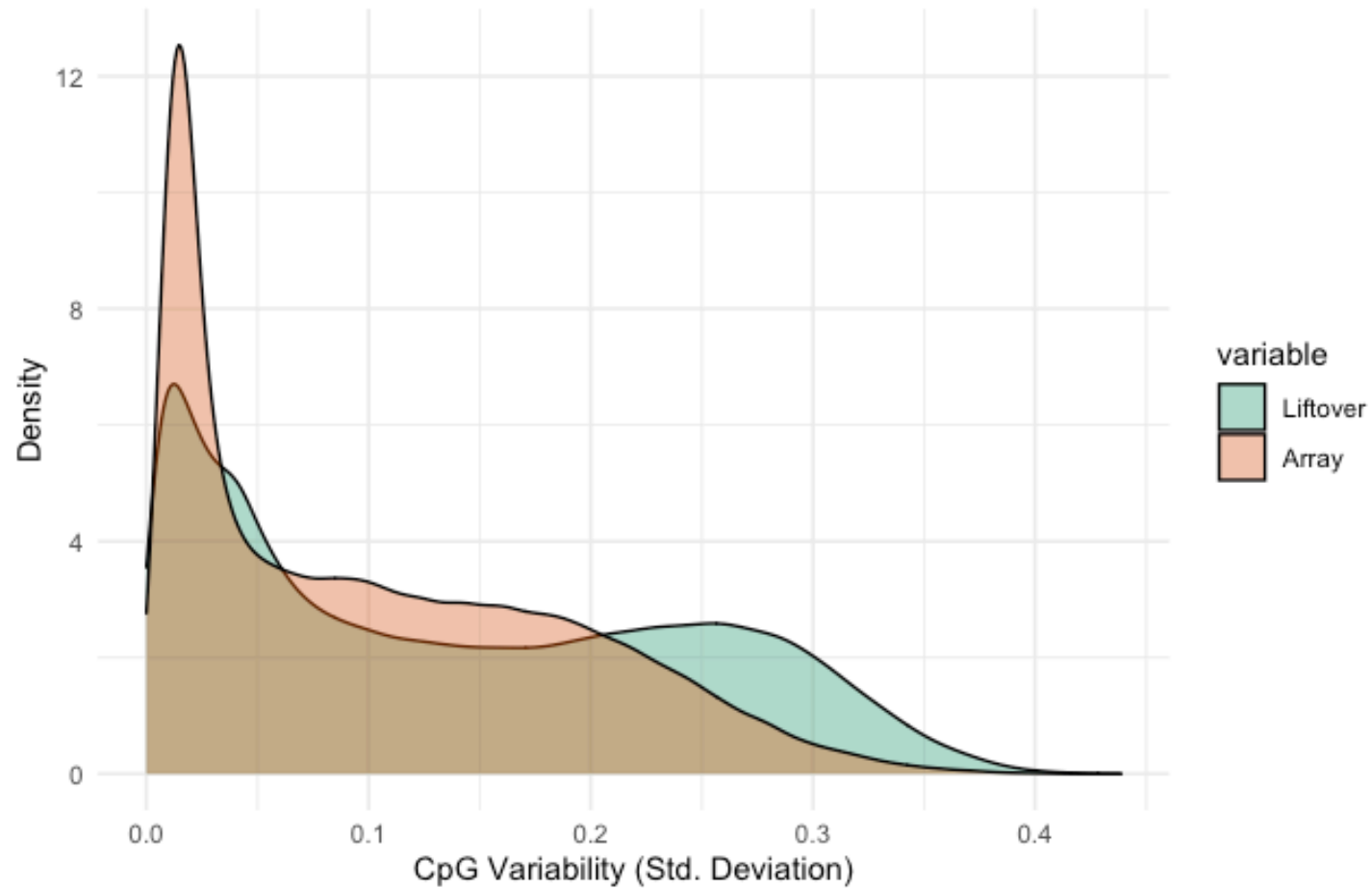
### 8.1.4.2 – 7-group classifier results

Coverage	WGBS - ICGC		WGBS - NMB		Array - ICGC		Array - NMB		
	Filter Applied	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)	No. probes matched to 10k feature set.	Mean Assignment probability (Std Dev)
<b>30x</b>		10,000	<b>0.873 (0.150)</b>	NA	NA				
<b>10x</b>	<b>5x</b>	10,000	<b>0.872 (0.150)</b>	10,000	<b>0.873 (0.144)</b>				
	<b>3x</b>	10,000	<b>0.874 (0.149)</b>	10,000	<b>0.905 (0.131)</b>				
<b>9x</b>	<b>5x</b>	10,000	<b>0.871 (0.149)</b>	10,000	<b>0.862 (0.147)</b>				
	<b>3x</b>	10,000	<b>0.874 (0.148)</b>	10,000	<b>0.903 (0.131)</b>				
<b>8x</b>	<b>5x</b>	10,000	<b>0.870 (0.148)</b>	10,000	<b>0.850 (0.144)</b>	9,651	<b>0.870 (0.150)</b>	9,651	<b>0.913 (0.115)</b>
	<b>3x</b>	10,000	<b>0.873 (0.147)</b>	10,000	<b>0.898 (0.136)</b>				
<b>7x</b>	<b>5x</b>	10,000	<b>0.873 (0.137)</b>	10,000	<b>0.818 (0.154)</b>				
	<b>3x</b>	10,000	<b>0.872 (0.148)</b>	10,000	<b>0.890 (0.141)</b>				
<b>6x</b>	<b>5x</b>	10,000	<b>0.871 (0.136)</b>	10,000	<b>0.763 (0.178)</b>				
	<b>3x</b>	10,000	<b>0.871 (0.148)</b>	10,000	<b>0.880 (0.148)</b>				
<b>5x</b>	<b>3x</b>	10,000	<b>0.870 (0.148)</b>	10,000	<b>0.862 (0.160)</b>				

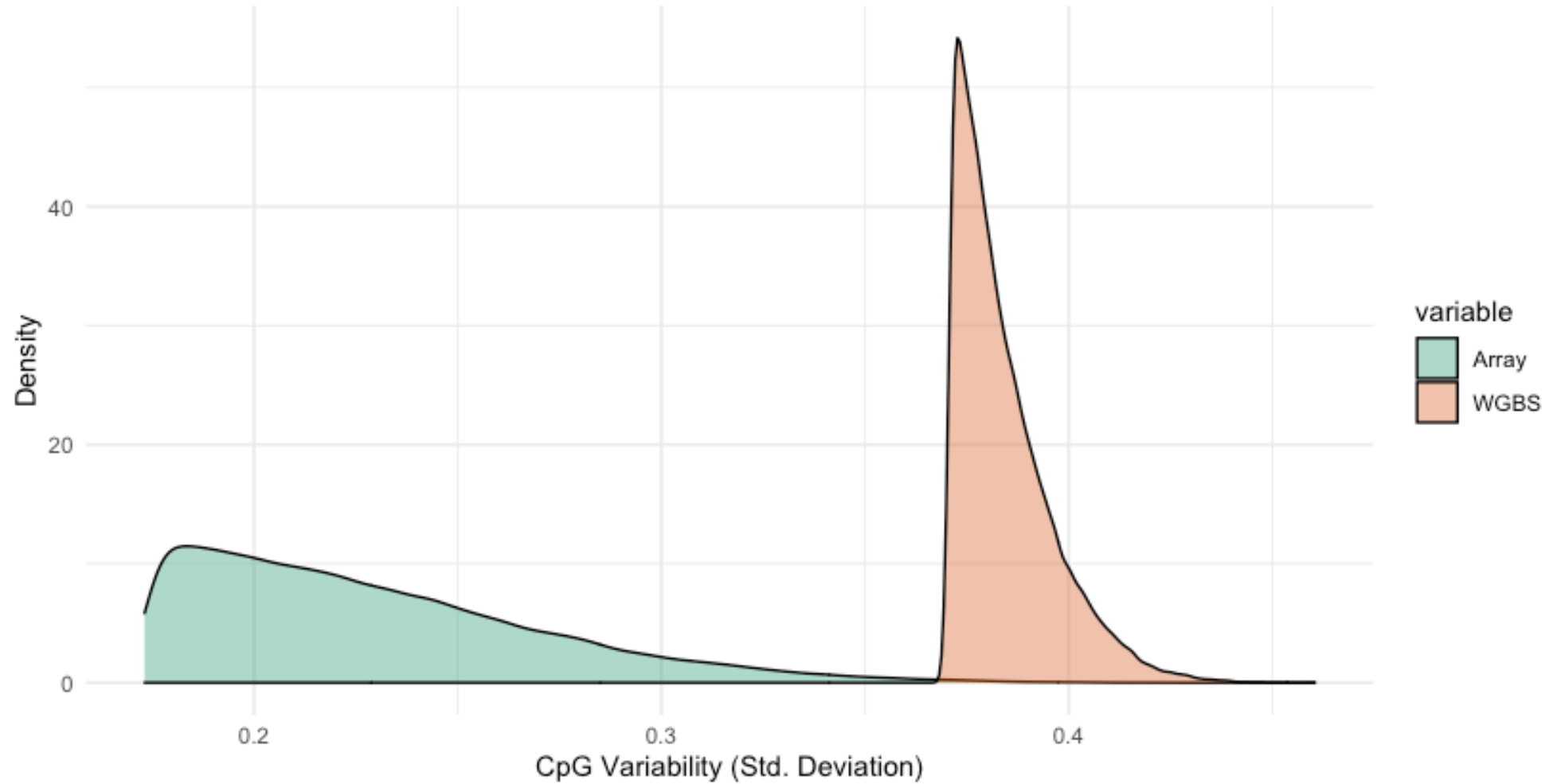
	<b>1x</b>	10,000	<b>0.872</b> (0.151)	10,000	<b>0.914</b> (0.127)	
<b>4x</b>	<b>3x</b>	10,000	<b>0.870</b> (0.142)	10,000	<b>0.841</b> (0.154)	
	<b>1x</b>	10,000	<b>0.870</b> (0.154)	10,000	<b>0.912</b> (0.129)	
<b>3x</b>	<b>1x</b>	10,000	<b>0.870</b> (0.152)	10,000	<b>0.908</b> (0.131)	
<b>2x</b>	<b>1x</b>	10,000	<b>0.868</b> (0.149)	10,000	<b>0.894</b> (0.137)	
<b>1x</b>	<b>1x</b>	10,000	<b>0.856</b> (0.159)	10,000	<b>0.847</b> (0.164)	

## 8.1.5 – CpG variability analysis

### 8.1.5.1 – CpG variability (Array vs Liftover cohort)



### 8.1.5.2 – CpG variability (Top 100,000 CpGs) – WGBS vs Array



### 8.1.5.3 – Chromosomal distribution of top 10,000 most variable features (WGBS / Liftover / Array cohorts)

\* denotes significantly different proportion compared to array – calculated using two-proportion z-test

Chromosome	Array	Liftover	WGBS
1	1,007	1081 *	938 *
2	864	902	788 *
3	372	389	467 *
4	432	498 *	410
5	612	648	970 *
6	908	797 *	582 *
7	650	700	705 *
8	389	395	513 *
9	143	147	318 *
10	600	565	450 *
11	623	679 *	392 *
12	615	637	461 *
13	359	374	434 *
14	301	350 *	318
15	262	293	496
16	437	333 *	238 *
17	571	512 *	342 *
18	117	132	315 *
19	356	233 *	228 *
20	156	138 *	313 *
21	94	105	148 *
22	132	92 *	174 *

## 8.1.6 – Copy number analysis

### 8.1.6.1 – Example conumee segmentation results – NMB\_362

chromosome	start	end	size	log2	call
chr1	560,684	148,927,230	148,366,546	-0.009	Balanced
chr1	149,077,230	149,379,823	302,593	0.213	Gain
chr1	149,629,823	249,195,311	99,565,488	-0.017	Balanced
chr10	105,000	36,950,000	36,845,000	0.222	Gain
chr10	37,675,000	46,175,000	8,500,000	-0.009	Balanced
chr10	46,313,482	47,660,823	1,347,341	0.162	Gain
chr10	48,252,854	135,462,374	87,209,520	-0.001	Balanced
chr11	130,000	1,575,000	1,445,000	-0.32	Loss
chr11	1,625,000	48,550,000	46,925,000	-0.495	Loss
chr11	48,950,000	134,873,258	85,923,258	0.014	Balanced
chr12	172,870	133,770,948	133,598,078	-0.007	Balanced
chr13	19,110,000	115,079,939	95,969,939	-0.005	Balanced
chr14	19,325,000	107,119,770	87,794,770	0.007	Balanced
chr15	20,275,000	102,410,696	82,135,696	-0.002	Balanced
chr16	105,000	35,017,901	34,912,901	0.015	Balanced
chr16	46,517,901	87,525,000	41,007,099	-0.388	Loss
chr16	87,575,000	90,222,377	2,647,377	-0.297	Loss
chr17	25,000	9,825,000	9,800,000	-0.391	Loss
chr17	9,900,000	16,575,000	6,675,000	-0.511	Loss
chr17	16,725,000	19,025,000	2,300,000	-0.357	Loss
chr17	19,150,000	81,122,605	61,972,605	0.25	Gain

<b>chr18</b>	80,000	77,983,624	77,903,624	0.002	<b>Balanced</b>
<b>chr19</b>	180,000	59,084,492	58,904,492	0.021	<b>Balanced</b>
<b>chr2</b>	130,000	243,026,238	242,896,238	-0.015	<b>Balanced</b>
<b>chr20</b>	155,000	62,907,760	62,752,760	0.011	<b>Balanced</b>
<b>chr21</b>	10,848,948	48,084,948	37,236,000	-0.009	<b>Balanced</b>
<b>chr22</b>	16,373,925	51,197,283	34,823,358	0.003	<b>Balanced</b>
<b>chr3</b>	155,000	197,831,215	197,676,215	-0.018	<b>Balanced</b>
<b>chr4</b>	80,000	190,972,138	190,892,138	-0.033	<b>Balanced</b>
<b>chr5</b>	55,000	140,225,000	140,170,000	-0.032	<b>Balanced</b>
<b>chr5</b>	140,275,000	140,825,000	550,000	-0.157	<b>Balanced</b>
<b>chr5</b>	140,875,000	180,777,630	39,902,630	-0.017	<b>Balanced</b>
<b>chr6</b>	155,000	170,952,534	170,797,534	0.003	<b>Balanced</b>
<b>chr7</b>	55,000	159,039,332	158,984,332	0.159	<b>Gain</b>
<b>chr8</b>	105,000	146,252,011	146,147,011	-0.004	<b>Balanced</b>
<b>chr9</b>	80,000	141,076,716	140,996,716	-0.005	<b>Balanced</b>
<b>chrX</b>	2,484,119	154,780,280	152,296,161	-0.02	<b>Balanced</b>

## 8.1.6.2 – Example CNVkit segmentation results – NMB\_362

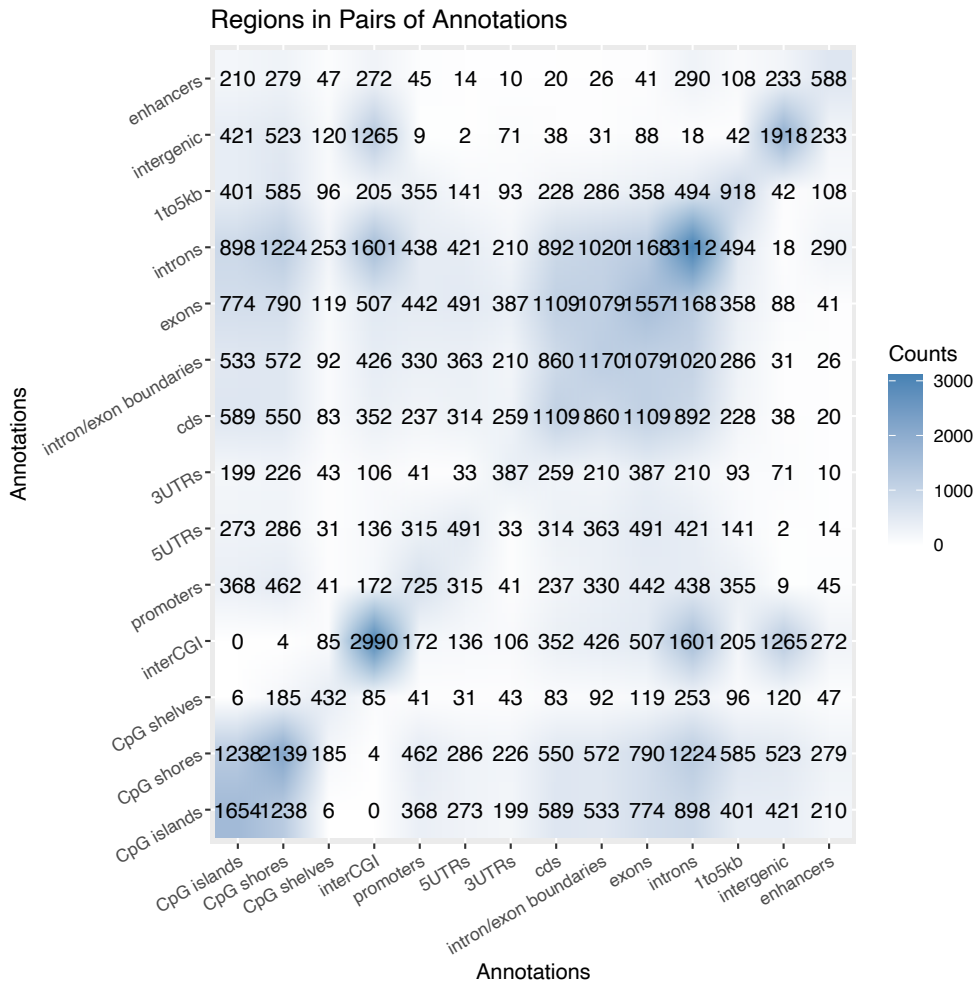
chromosome	start	end	size	log2	cn	call	probe/bins	Focal gene of interest?	stdev	iqr	mad	bivar	mse	sem	ci_lo	ci_hi
chr1	750,100	249,225,300	248,475,200	-0.01	2	Balanced	4342		0.12	0.09	0.07	0.07	0.01	0.00	-0.02	-0.01
chr2	10,000	243,052,100	243,042,100	0.01	2	Balanced	4626		0.08	0.08	0.06	0.06	0.01	0.00	0.00	0.01
chr3	60,000	15,364,321	15,304,321	-0.59	1	Loss	303		0.08	0.10	0.08	0.08	0.01	0.00	-0.60	-0.58
chr3	15,364,321	197,798,000	182,433,679	0.00	2	Balanced	3561		0.07	0.08	0.06	0.06	0.00	0.00	0.00	0.00
chr4	69,600	3,761,390	3,691,790	0.31	3	Gain	71		0.14	0.11	0.08	0.09	0.02	0.02	0.29	0.34
chr4	3,761,390	190,795,300	187,033,910	0.01	2	Balanced	3642		0.14	0.08	0.06	0.06	0.02	0.00	0.00	0.01
chr5	85,500	121,635,512	121,550,012	0.02	2	Balanced	2314		0.07	0.08	0.06	0.06	0.00	0.00	0.01	0.02
chr5	121,635,512	121,885,507	249,995	0.96	4	Gain	5	SNCAIP	0.21	0.26	0.05	0.02	0.05	0.11	0.74	1.09
chr5	121,885,507	180,599,700	58,714,193	-0.01	2	Balanced	1150		0.08	0.09	0.06	0.07	0.01	0.00	-0.02	-0.01
chr6	461,610	98,079,775	97,618,165	-0.01	2	Balanced	1765		0.07	0.09	0.06	0.07	0.01	0.00	-0.01	-0.01
chr6	98,079,775	170,915,300	72,835,525	0.37	3	Gain	1452		0.07	0.08	0.06	0.06	0.00	0.00	0.36	0.37
chr7	49,700	159,128,663	159,078,963	0.56	3	Gain	3015		0.10	0.08	0.06	0.07	0.01	0.00	0.56	0.57
chr8	185,300	146,304,022	146,118,722	0.01	2	Balanced	2785		0.08	0.09	0.06	0.07	0.01	0.00	0.01	0.02
chr9	10,000	210,024	200,024	-0.45	1	Loss	4		0.12	0.19	0.15	0.12	0.01	0.07	-0.57	-0.35
chr9	210,024	141,053,300	140,843,276	-0.01	2	Balanced	2157		0.08	0.09	0.07	0.07	0.01	0.00	-0.02	-0.01
chr10	60,000	13,048,645	12,988,645	-0.42	1	Loss	260		0.07	0.09	0.07	0.07	0.01	0.00	-0.43	-0.41
chr10	13,048,645	135,437,600	122,388,955	0.00	2	Balanced	2270		0.10	0.09	0.07	0.07	0.01	0.00	0.00	0.01
chr11	196,300	49,065,071	48,868,771	-0.98	1	Loss	970		0.13	0.11	0.08	0.09	0.02	0.00	-0.99	-0.97
chr11	49,065,071	134,946,516	85,881,445	-0.01	2	Balanced	1615		0.09	0.09	0.07	0.07	0.01	0.00	-0.01	0.00
chr12	187,000	133,825,400	133,638,400	-0.01	2	Balanced	2578		0.08	0.09	0.07	0.07	0.01	0.00	-0.02	-0.01
chr13	19,194,200	115,109,878	95,915,678	0.01	2	Balanced	1870		0.07	0.08	0.06	0.06	0.00	0.00	0.01	0.02
chr14	20,303,800	107,289,540	86,985,740	0.00	2	Balanced	1737		0.08	0.09	0.06	0.07	0.01	0.00	-0.01	0.00
chr15	20,166,200	102,411,600	82,245,400	-0.03	2	Balanced	1527		0.09	0.10	0.07	0.07	0.01	0.00	-0.03	-0.02



chr16	60,000	83,643,390	83,583,390	-0.04	2	Balanced	1301	0.11	0.09	0.07	0.07	0.01	0.00	-0.05	-0.04
chr16	83,643,390	90,155,800	6,512,410	-0.57	1	Loss	129	0.14	0.13	0.10	0.12	0.02	0.01	-0.60	-0.55
chr17	0	18,279,231	18,279,231	-0.30	2	Balanced	362	0.09	0.10	0.08	0.08	0.01	0.00	-0.30	-0.29
chr17	18,279,231	18,928,600	649,369	-0.55	1	Loss	13	0.25	0.24	0.22	0.25	0.06	0.07	-0.72	-0.39
chr17	19,140,800	70,852,805	51,712,005	0.53	3	Gain	925	0.12	0.10	0.07	0.08	0.01	0.00	0.52	0.54
chr17	70,852,805	81,195,210	10,342,405	0.90	4	Gain	205	0.18	0.09	0.07	0.08	0.03	0.01	0.87	0.92
chr18	127,000	78,017,248	77,890,248	0.03	2	Balanced	1477	0.07	0.08	0.06	0.06	0.00	0.00	0.03	0.04
chr19	210,155	59,118,983	58,908,828	-0.10	2	Balanced	1090	0.14	0.12	0.09	0.10	0.02	0.00	-0.11	-0.09
chr20	60,000	26,184,300	26,124,300	-0.58	1	Loss	515	0.09	0.10	0.08	0.08	0.01	0.00	-0.59	-0.57
chr20	29,521,147	62,887,700	33,366,553	-0.04	2	Balanced	658	0.08	0.10	0.07	0.08	0.01	0.00	-0.05	-0.03
chr21	14,670,197	48,119,895	33,449,698	0.00	2	Balanced	668	0.07	0.09	0.06	0.07	0.01	0.00	-0.01	0.00
chr22	16,962,100	51,220,000	34,257,900	-0.05	2	Balanced	662	0.11	0.10	0.07	0.08	0.01	0.00	-0.06	-0.05
chrX	2,699,520	154,931,043	152,231,523	-1.03	1	Loss	2916	0.28	0.11	0.08	0.08	0.08	0.01	-1.04	-1.02

## 8.1.7 – Differential methylation analysis

### 8.1.7.1 – Degree to which DMR annotations overlap – SHH example



### 8.1.7.2 – Gene ontology analysis – Subtype VII - VII\_1 vs VII\_2 DMRs (Array)

Term	Implicated Genes	p-value	Adjusted p-value	Odds Ratio	Combined Score
acyl-CoA biosynthetic process (GO:0071616)	1 / 14	0.003	0.006	512.385	3012.375
fatty-acyl-CoA metabolic process (GO:0035337)	1 / 19	0.004	0.006	369.963	2062.218
fatty acid derivative biosynthetic process (GO:1901570)	1 / 24	0.005	0.006	289.464	1545.991
fatty-acyl-CoA biosynthetic process (GO:0046949)	1 / 30	0.006	0.006	229.506	1174.653

### 8.1.7.3 – Gene ontology analysis – Subtype VII – VII\_1 vs VII\_2 DMRs (WGBS)

Term	Implicated Genes	p-value	Adjusted p-value	Odds Ratio	Combined Score
nervous system development (GO:0007399)	116/455	0.000	<b>0.000</b>	2.489	80.269
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	178/848	0.000	<b>0.000</b>	1.947	55.575
regulation of transcription from RNA polymerase II promoter (GO:0006357)	271/1478	0.000	<b>0.000</b>	1.660	43.146
positive regulation of transcription, DNA-templated (GO:0045893)	213/1120	0.000	<b>0.000</b>	1.723	41.130
neuron differentiation (GO:0030182)	43/139	0.000	<b>0.000</b>	3.206	60.640
heart development (GO:0007507)	43/149	0.000	<b>0.000</b>	2.901	48.297
chordate embryonic development (GO:0043009)	23/56	0.000	<b>0.000</b>	4.965	82.166
renal system development (GO:0072001)	23/59	0.000	<b>0.000</b>	4.551	70.105

<b>negative regulation of cell differentiation (GO:0045596)</b>	42/150	0.000	<b>0.000</b>	2.780	42.734
<b>embryonic organ development (GO:0048568)</b>	21/53	0.000	<b>0.000</b>	4.671	67.802
<b>gland development (GO:0048732)</b>	26/75	0.000	<b>0.000</b>	3.781	54.774
<b>skeletal system development (GO:0001501)</b>	40/146	0.000	<b>0.000</b>	2.696	37.999
<b>digestive tract development (GO:0048565)</b>	18/43	0.000	<b>0.000</b>	5.121	69.624
<b>epithelium development (GO:0060429)</b>	33/115	0.000	<b>0.001</b>	2.871	37.323
<b>urogenital system development (GO:0001655)</b>	12/22	0.000	<b>0.001</b>	8.522	109.997
<b>embryonic organ morphogenesis (GO:0048562)</b>	15/33	0.000	<b>0.001</b>	5.922	75.815
<b>kidney development (GO:0001822)</b>	22/63	0.000	<b>0.001</b>	3.819	48.209
<b>cell morphogenesis (GO:0000902)</b>	21/59	0.000	<b>0.001</b>	3.933	49.011

<b>axon guidance (GO:0007411)</b>	40/158	0.000	<b>0.001</b>	2.420	28.923
<b>skeletal system morphogenesis (GO:0048705)</b>	14/31	0.000	<b>0.001</b>	5.851	69.858
<b>ear morphogenesis (GO:0042471)</b>	12/24	0.000	<b>0.002</b>	7.101	83.197
<b>central nervous system development (GO:0007417)</b>	50/217	0.000	<b>0.002</b>	2.140	24.905
<b>negative regulation of transcription, DNA-templated (GO:0045892)</b>	142/813	0.000	<b>0.002</b>	1.527	17.357
<b>negative regulation of transcription from RNA polymerase II promoter (GO:0000122)</b>	105/565	0.000	<b>0.002</b>	1.641	18.634
<b>positive regulation of nucleic acid-templated transcription (GO:1903508)</b>	95/502	0.000	<b>0.002</b>	1.676	18.694
<b>respiratory system development (GO:0060541)</b>	13/29	0.000	<b>0.002</b>	5.770	63.923
<b>regulation of Wnt signaling pathway (GO:0030111)</b>	30/109	0.000	<b>0.002</b>	2.706	29.906
<b>circulatory system development (GO:0072359)</b>	30/109	0.000	<b>0.002</b>	2.706	29.906

<b>digestive system development (GO:0055123)</b>	11/22	0.000	<b>0.003</b>	7.098	76.932
<b>embryonic limb morphogenesis (GO:0030326)</b>	14/34	0.000	<b>0.003</b>	4.972	52.927
<b>muscle tissue morphogenesis (GO:0060415)</b>	10/19	0.000	<b>0.004</b>	7.884	83.038
<b>positive regulation of synaptic transmission (GO:0050806)</b>	21/67	0.000	<b>0.005</b>	3.247	33.218
<b>generation of neurons (GO:0048699)</b>	33/130	0.000	<b>0.005</b>	2.425	24.692
<b>axonogenesis (GO:0007409)</b>	49/223	0.000	<b>0.005</b>	2.011	20.333
<b>branching involved in ureteric bud morphogenesis (GO:0001658)</b>	10/20	0.000	<b>0.005</b>	7.096	70.648
<b>dorsal/ventral pattern formation (GO:0009953)</b>	10/20	0.000	<b>0.005</b>	7.096	70.648
<b>ureteric bud morphogenesis (GO:0060675)</b>	10/20	0.000	<b>0.005</b>	7.096	70.648
<b>embryonic digestive tract development (GO:0048566)</b>	10/20	0.000	<b>0.005</b>	7.096	70.648

<b>carbohydrate biosynthetic process (GO:0016051)</b>	14/36	0.000	<b>0.006</b>	4.520	44.665
<b>heart morphogenesis (GO:0003007)</b>	17/50	0.000	<b>0.007</b>	3.661	35.445
<b>limb morphogenesis (GO:0035108)</b>	10/21	0.000	<b>0.008</b>	6.450	60.809
<b>mesonephros development (GO:0001823)</b>	7/11	0.000	<b>0.009</b>	12.407	115.167
<b>negative regulation of multicellular organismal process (GO:0051241)</b>	31/125	0.000	<b>0.010</b>	2.349	21.579
<b>branching morphogenesis of an epithelial tube (GO:0048754)</b>	15/43	0.000	<b>0.011</b>	3.805	34.355
<b>inner ear morphogenesis (GO:0042472)</b>	10/22	0.000	<b>0.012</b>	5.912	52.850
<b>regulation of epithelial to mesenchymal transition (GO:0010717)</b>	20/68	0.000	<b>0.013</b>	2.962	26.127
<b>synapse assembly (GO:0007416)</b>	19/63	0.000	<b>0.013</b>	3.069	27.049
<b>regulation of mesenchymal stem cell differentiation (GO:2000739)</b>	5/6	0.000	<b>0.013</b>	35.425	310.390



<b>autonomic nervous system development (GO:0048483)</b>	8/15	0.000	<b>0.013</b>	8.104	70.968
<b>digestive tract morphogenesis (GO:0048546)</b>	8/15	0.000	<b>0.013</b>	8.104	70.968
<b>positive regulation of stem cell differentiation (GO:2000738)</b>	7/12	0.000	<b>0.016</b>	9.925	84.566
<b>cardiac right ventricle morphogenesis (GO:0003215)</b>	7/12	0.000	<b>0.016</b>	9.925	84.566
<b>embryonic digestive tract morphogenesis (GO:0048557)</b>	6/9	0.000	<b>0.017</b>	14.174	119.620
<b>columnar/cuboidal epithelial cell differentiation (GO:0002065)</b>	6/9	0.000	<b>0.017</b>	14.174	119.620
<b>regulation of potassium ion transport (GO:0043266)</b>	13/36	0.000	<b>0.017</b>	4.012	33.661
<b>cardiac muscle tissue development (GO:0048738)</b>	13/36	0.000	<b>0.017</b>	4.012	33.661
<b>sympathetic nervous system development (GO:0048485)</b>	8/16	0.000	<b>0.021</b>	7.091	57.993
<b>calcium ion import (GO:0070509)</b>	11/28	0.000	<b>0.021</b>	4.591	37.342

<b>cellular response to retinoic acid (GO:0071300)</b>	14/42	0.000	<b>0.025</b>	3.550	28.292
<b>oligosaccharide biosynthetic process (GO:0009312)</b>	12/33	0.000	<b>0.026</b>	4.055	32.091
<b>neuron fate commitment (GO:0048663)</b>	7/13	0.000	<b>0.026</b>	8.270	65.010
<b>lung development (GO:0030324)</b>	11/29	0.000	<b>0.028</b>	4.336	33.709
<b>development of primary male sexual characteristics (GO:0046546)</b>	14/43	0.000	<b>0.028</b>	3.427	26.376
<b>oligosaccharide metabolic process (GO:0009311)</b>	18/63	0.000	<b>0.028</b>	2.842	21.832
<b>respiratory tube development (GO:0030323)</b>	10/25	0.000	<b>0.028</b>	4.729	36.272
<b>negative regulation of epithelial to mesenchymal transition (GO:0010719)</b>	10/25	0.000	<b>0.028</b>	4.729	36.272
<b>ureteric bud development (GO:0001657)</b>	8/17	0.000	<b>0.028</b>	6.302	48.262
<b>thyroid gland development (GO:0030878)</b>	6/10	0.000	<b>0.028</b>	10.630	81.151

<b>gas transport (GO:0015669)</b>	6/10	0.000	<b>0.028</b>	10.630	81.151
<b>forelimb morphogenesis (GO:0035136)</b>	6/10	0.000	<b>0.028</b>	10.630	81.151
<b>mesenchymal to epithelial transition (GO:0060231)</b>	6/10	0.000	<b>0.028</b>	10.630	81.151
<b>long-term synaptic potentiation (GO:0060291)</b>	9/21	0.000	<b>0.028</b>	5.319	40.528
<b>embryonic skeletal system development (GO:0048706)</b>	9/21	0.000	<b>0.028</b>	5.319	40.528
<b>kidney epithelium development (GO:0072073)</b>	5/7	0.000	<b>0.028</b>	17.711	134.918
<b>neuron fate specification (GO:0048665)</b>	5/7	0.000	<b>0.028</b>	17.711	134.918
<b>regulation of neuron projection development (GO:0010975)</b>	28/119	0.001	<b>0.030</b>	2.189	16.451
<b>homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156)</b>	18/64	0.001	<b>0.031</b>	2.780	20.775
<b>activation of protein kinase B activity (GO:0032148)</b>	11/30	0.001	<b>0.031</b>	4.108	30.541

<b>gonad development (GO:0008406)</b>	15/49	0.001	<b>0.032</b>	3.132	23.211
<b>endocrine system development (GO:0035270)</b>	10/26	0.001	<b>0.035</b>	4.433	32.369
<b>anterior/posterior axis specification (GO:0009948)</b>	8/18	0.001	<b>0.038</b>	5.672	40.749
<b>positive regulation of synaptic transmission, glutamatergic (GO:0051968)</b>	8/18	0.001	<b>0.038</b>	5.672	40.749
<b>cardiac muscle tissue morphogenesis (GO:0055008)</b>	12/36	0.001	<b>0.045</b>	3.548	24.863
<b>metanephric mesenchyme development (GO:0072075)</b>	6/11	0.001	<b>0.047</b>	8.503	59.155

## 8.1.8 – Variant calling

### 8.1.8.1 – Diagnostic panel gene list

MDM4  
KIF26B  
NBAS  
MYCN  
MDM4  
GLI2  
LRP1B  
NEB  
CTNNB1  
MDM4  
CACNA1D  
PIK3CA  
FBXW7  
APC  
SNCAIP  
CSNK2B  
TNXB  
EYA4  
ABCA13  
CDK6  
SMO  
KMT2C  
SHH  
EPPK1  
CDKN2A  
CDKN2B  
PTCH1  
TSC1  
PTEN  
LDB1  
SUFU  
ATM  
KMT2D  
MDM2  
BRCA2  
OTX2  
SPTB  
AKT1  
RYR3

MAN2C1  
TSC2  
CREBBP  
PALB2  
GTF3C1  
CTDNEP1  
GPS2  
TP53  
PRKAR1A  
TCF4  
ALPK2  
CDH7  
SMARCA4  
FCGBP  
LTBP4  
NLRP5  
CBFA2T2  
SMARCB1  
EP300  
BCOR  
DDX3X  
KDM6A  
ZMYM3  
FLNA

## 8.1.9 – Development of WGBS-trained classifier models

### 8.1.9.1 – Mean validation performance – XGBoost (Tree booster)

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>1x</b>	0.815	0.861	0.477	0.987
<b>2x</b>	0.879	0.909	0.383	0.994
<b>3x</b>	0.895	0.922	0.337	0.996
<b>4x</b>	0.906	0.929	0.326	0.997
<b>5x</b>	0.897	0.923	0.316	0.997
<b>6x</b>	0.857	0.893	0.367	0.991
<b>7x</b>	0.913	0.934	0.326	0.995
<b>8x</b>	0.866	0.899	0.342	0.995
<b>9x</b>	0.877	0.907	0.344	0.996
<b>10x</b>	0.883	0.912	0.352	0.994
<b>30x</b>	0.873	0.906	0.323	0.991
<b>WGBS</b>	0.873	0.906	0.323	0.991
<b>Liftover</b>	0.940	0.955	0.278	0.999
<b>Array</b>	0.948	0.961	0.263	0.996

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>10x (High Filter)</b>	0.845	0.884	0.392	0.991
<b>10x (Low Filter)</b>	0.873	0.906	0.323	0.991
<b>Array</b>	0.948	0.961	0.263	0.996

### 8.1.9.2 – Mean validation performance – XGBoost (Linear booster)

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>1x</b>	0.747	0.808	0.593	0.966
<b>2x</b>	0.799	0.849	0.498	0.978
<b>3x</b>	0.799	0.848	0.448	0.983
<b>4x</b>	0.808	0.855	0.447	0.983
<b>5x</b>	0.812	0.858	0.431	0.985
<b>6x</b>	0.802	0.850	0.458	0.979
<b>7x</b>	0.870	0.902	0.360	0.990
<b>8x</b>	0.828	0.871	0.412	0.984
<b>9x</b>	0.809	0.856	0.472	0.975
<b>10x</b>	0.846	0.885	0.424	0.983
<b>30x</b>	0.834	0.875	0.369	0.980
<b>WGBS</b>	0.834	0.875	0.369	0.980
<b>Liftover</b>	0.843	0.882	0.395	0.980
<b>Array</b>	0.843	0.882	0.420	0.983

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>10x (High Filter)</b>	0.759	0.818	0.508	0.965
<b>10x (Low Filter)</b>	0.834	0.875	0.369	0.980
<b>Array</b>	0.843	0.882	0.420	0.983



### 8.1.9.3 – Mean validation performance – Support vector machine (Linear kernel)

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>1x</b>	0.933	0.950	0.360	0.998
<b>2x</b>	0.961	0.970	0.351	0.999
<b>3x</b>	0.961	0.970	0.357	0.998
<b>4x</b>	0.950	0.963	0.363	0.998
<b>5x</b>	0.941	0.956	0.370	0.997
<b>6x</b>	0.959	0.969	0.369	0.997
<b>7x</b>	0.961	0.971	0.368	0.998
<b>8x</b>	0.957	0.968	0.373	0.998
<b>9x</b>	0.956	0.967	0.380	0.997
<b>10x</b>	0.955	0.967	0.377	0.997
<b>30x</b>	0.950	0.963	0.382	0.997
<b>WGBS</b>	0.950	0.963	0.382	0.997
<b>Liftover</b>	0.979	0.984	0.343	0.999
<b>Array</b>	0.980	0.985	0.344	1.000

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>10x (High Filter)</b>	0.937	0.953	0.392	0.996
<b>10x (Low Filter)</b>	0.950	0.963	0.382	0.997
<b>Array</b>	0.980	0.985	0.344	1.000

### 8.1.9.4 – Mean validation performance – Support vector machine (Radial kernel)

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>1x</b>	0.539	0.649	0.974	0.895
<b>2x</b>	0.587	0.683	0.959	0.894
<b>3x</b>	0.623	0.711	0.944	0.895
<b>4x</b>	0.549	0.644	0.938	0.920
<b>5x</b>	0.557	0.651	0.927	0.919
<b>6x</b>	0.680	0.759	0.907	0.894
<b>7x</b>	0.686	0.764	0.905	0.895
<b>8x</b>	0.681	0.760	0.909	0.895
<b>9x</b>	0.682	0.761	0.916	0.894
<b>10x</b>	0.684	0.762	0.914	0.894
<b>30x</b>	0.695	0.771	0.899	0.899
<b>WGBS</b>	0.695	0.771	0.899	0.899
<b>Liftover</b>	0.664	0.756	0.881	0.922
<b>Array</b>	0.686	0.768	0.895	0.928

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>10x (High Filter)</b>	0.697	0.774	0.878	0.901
<b>10x (Low Filter)</b>	0.695	0.771	0.899	0.899
<b>Array</b>	0.686	0.768	0.895	0.928

### 8.1.9.5 – Mean validation performance – Logic regression

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>1x</b>	0.975	0.981	0.054	1.000
<b>2x</b>	0.979	0.984	0.046	1.000
<b>3x</b>	0.973	0.980	0.054	1.000
<b>4x</b>	0.982	0.987	0.046	1.000
<b>5x</b>	0.979	0.984	0.050	1.000
<b>6x</b>	0.975	0.981	0.056	1.000
<b>7x</b>	0.975	0.981	0.057	1.000
<b>8x</b>	0.980	0.985	0.048	1.000
<b>9x</b>	0.975	0.981	0.045	1.000
<b>10x</b>	0.975	0.981	0.052	1.000
<b>30x</b>	0.978	0.984	0.043	1.000
<b>WGBS</b>	0.978	0.984	0.043	1.000
<b>Liftover</b>	0.976	0.982	0.061	1.000
<b>Array</b>	0.959	0.969	0.070	1.000

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>10x (High Filter)</b>	0.978	0.984	0.043	1.000
<b>10x (Low Filter)</b>	0.969	0.977	0.068	1.000
<b>Array</b>	0.959	0.969	0.070	1.000

### 8.1.9.6 – Mean validation performance – Random Forest

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>1x</b>	0.870	0.904	0.650	0.994
<b>2x</b>	0.945	0.959	0.519	0.997
<b>3x</b>	0.948	0.961	0.458	0.997
<b>4x</b>	0.959	0.969	0.423	0.997
<b>5x</b>	0.966	0.974	0.395	0.998
<b>6x</b>	0.937	0.953	0.397	0.997
<b>7x</b>	0.955	0.966	0.384	0.998
<b>8x</b>	0.954	0.965	0.377	0.998
<b>9x</b>	0.959	0.969	0.370	0.998
<b>10x</b>	0.963	0.972	0.359	0.998
<b>30x</b>	0.964	0.973	0.360	0.997
<b>WGBS</b>	0.964	0.973	0.360	0.997
<b>Liftover</b>	0.949	0.962	0.349	0.999
<b>Array</b>	0.973	0.980	0.297	0.999

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>10x (High Filter)</b>	0.940	0.955	0.386	0.996
<b>10x (Low Filter)</b>	0.964	0.973	0.360	0.997
<b>Array</b>	0.973	0.980	0.297	0.999

### 8.1.9.7 – Mean validation performance – Elastic net

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>1x</b>	0.940	0.955	0.219	0.999
<b>2x</b>	0.973	0.980	0.158	0.999
<b>3x</b>	0.959	0.969	0.143	1.000
<b>4x</b>	0.963	0.973	0.134	0.999
<b>5x</b>	0.978	0.984	0.120	1.000
<b>6x</b>	0.961	0.971	0.135	0.999
<b>7x</b>	0.968	0.976	0.113	1.000
<b>8x</b>	0.967	0.975	0.111	1.000
<b>9x</b>	0.961	0.971	0.114	1.000
<b>10x</b>	0.964	0.973	0.122	0.999
<b>30x</b>	0.956	0.967	0.117	0.999
<b>WGBS</b>	0.956	0.967	0.117	0.999
<b>Liftover</b>	0.980	0.985	0.106	1.000
<b>Array</b>	0.980	0.985	0.084	1.000

Coverage	Kappa (Mean)	Accuracy (Mean)	logLoss (Mean)	AUC (Mean)
<b>10x (High Filter)</b>	0.951	0.963	0.117	0.999
<b>10x (Low Filter)</b>	0.956	0.967	0.117	0.999
<b>Array</b>	0.980	0.985	0.084	1.000

**END OF DOCUMENT**

