Aus dem KiTZ- Hopp Children's Cancer Center, Deutschen Krebsforschungszentrum

(Director/in: Prof. Dr. med. Andreas Kulozik, Prof. Dr. med. Olaf Witt, Prof. Dr.med. Stefan M.Pfister)

Abteilung. KiTZ - Hopp Children's Cancer center

(Abteilungsleiter: Prof. Dr. Med. Stefan M Pfister)

Analysis of copy number variation profiles in brain tumors in the context of a methylation based classifier

Inauguraldissertation

zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)

an der

Medizinischen Fakultät Heidelberg

der

Ruprecht-Karls-Universität

vorgelegt von

Bidii Stephen Ngalah

aus

Mombasa, Kenya

2021

Dekan: Prof. Dr. Hans-Georg Kräusslich

Doktorvater: PD. Dr. rer. nat. Obul Reddy Bandapalli

Dedication

In memory of my Dad and Mom, Engineer. Stephen Ngalah Ngenja and Nyamvula Ngome Jefwa. No matter the turbulence, you kept your eyes on me. You instilled honesty, hard-work, humbleness and generosity in my heart. Your ideals and wisdom will run through generations as your lineage serves humanity with dignity and integrity.

Table of Contents

List of figures

List of tables

Abbreviations

aCGH	Array comparative genomic hybridization
APC	Adenomatous polyposis coli
APD	Auditory processing deficit
APO	Apolipoprotein 1
ATRX	X-linked alpha-thalassemia chromatin remodelling gene
BMI1	B lymphoma Mo-MLV insertion region 1 homolog
BRAF	Proto-oncogene B-Raf
CCL2	Chemokine C-C- motif ligand 2
CCND1	Cyclin D1
CD133	Cluster of differentiation 133
CDKN2A	Cyclin Dependent Kinase Inhibitor 2A
CDKs	Cyclin-dependent kinases
cDNA	Complementary deoxyribonucleic acid
CHI3L2	Chitinase 3-like 2
Chr	Chromosome
CLEC-2	C-type lectin receptors-2
CMTM7	CKLF Like MARVEL Transmembrane Domain Containing 7
CN	Copy number
CNV	Copy number variations
COD	Coefficient of determination
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CSC	Cancer stem cells
CTNNB1	Catenin Beta 1
dH$_2$O	Distilled water
DNA	Desoxyribonucleic acid
dNTP	Deoxyribonucleoside Triphosphate
EF2	Elongation factor 2
EGFR	Epidermal Growth Factor Receptor
EWAS	Epigenome-wide association studies
FFPE	Formalin-Fixed Paraffin-Embedded
FISH	Fluorescence In Situ Hybridization

GABRA1	Gamma-aminobutyric acid Type A receptor subunit alpha-1
GBM	Glioblastoma multiforme
GBM-BDN	Glioblastoma blood normal control
GBM-PT	Glioblastoma primary tumour
GBM-RT	GBM recurrent tumour
GSCs	Glioma stem cell
GSK-3β	Glycogen synthase kinase 3 beta
Id1	Inhibitor of DNA Binding 1
NFI	Nuclear factor I
ICC	Intraclass correlation coefficient
IDH1/2	Isocitrate dehydrogenase
IHC	Immunohistochemistry
Kras	Kirsten RAt Sarcoma gene
LEV	Low elution volume
LGG	Low grade glioma
LGG-BDN	Low grade glioma -blood normal control
LGG-PT	Low grade glioma -primary tumour
LIN28A	Lin-28 Homolog A
LOH	Loss of heterozygosity
LRR	Log R Ratio
MAP2K1	Mitogen-activated protein kinase kinase 1
MGMT	O6- methylguanine-DNA methyltransferase
MGP	Matrix gla protein
MDM2	Mouse double minute 2 homolog
MMP	Matrix metalloproteinase
mRNA	Messenger ribonucleotide acid
NMB	Neuromedin B
NSCs	Neuronal stem cells
ORF	Open reading frame
P13K	Phosphoinositide 3-kinases
PCR	Polymerase chain reaction
PDGF	Platelet-derived growth factor
PDGFRA	Platelet derived growth factor receptor alpha

PNET	Primitive neuroectodermal tumors
pRb	Rb1 encoded protein
PTEN	Phosphatase and tensin homolog gene
qPCR	Quantitative real time polymerase chain reaction
Rb	Retinoblastoma
RNA	Ribonucleic acid
SOS	Son of sevenless
SOX2/SRY	Sex determining region Y-box 2
TERT	Telomerase reverse transcriptase
TFs	Transcription factors
TMZ	Temozolomide
TOP2A	Topoisomerase (DNA) II alpha
TP53	Tumor protein P53
TYMS	Thymidylate synthetase
WGS	Whole genome sequencing
WHO	World health organization
Wnt	Wingless and Int-1

1.0 Introduction

1.1 Cellular origin of brain tumors

The central nervous system is made up of the spine and brain. The largest part of the brain is the cerebrum which divides into four lobes namely, the frontal lobe which controls speech, reasoning, movement and even emotions, parietal lobe controls sense responses e.g touch, pain and environmental temperature[1], occipital and temporal lobe controls vision and memory respectively. At the back of the brain below the cerebrum, we find cerebellum. This part of the brain is in-charge of body coordination and balance.

Brain tumours are thought to arise from transformed neuronal stem cells (NSCs). This can lead to either benign or malignant stem cells. Regardless of the malignancy status, these cells are believed to reside in perivascular niches which are formed by the blood vessels of the brain[2]. The ability to identify alterations and differentiate cells of origins in the brain allow to compare both causality and tumorigenic potentials of these alterations and specific neural stem and progenitor cells. For example, *KRAS* activation and inactivation of *PTEN, P53* and *RB1* has been used as prospective makers to address the occurrence of cancer in different cells. Other common tumour suppressor genes used include *NF1*. It has been shown that astrocytes and oligodendrocytes progenitors are probably the source of gliomas as shown in figure 1.1[3]. Based on cell of origin, we therefore classify this tumor as astrocytic, oligodendroglial or mixed tumors.
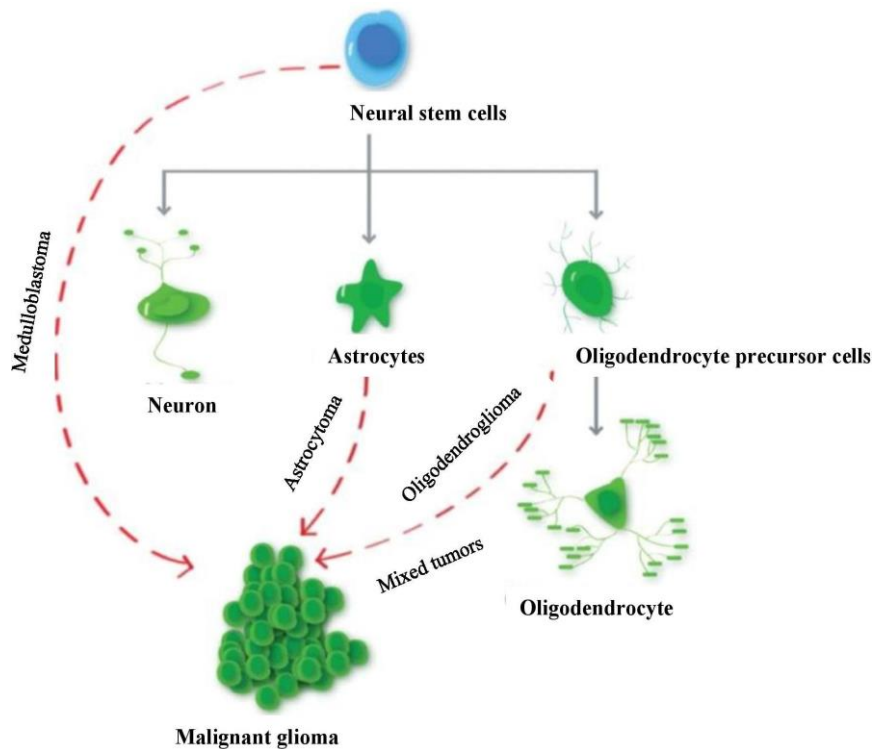
Figure 1.1 Neural stem cell lineages with dotted lines indicating the probable cells that could serve as source of glioma tumours.

It also indicates a glioma classification scheme based on cell of origin. This figure was adopted from H.zhong et al., 2015.

1.2 Brain tumor markers with focus to glioma markers distribution and molecular diagnosis

Central nervous tumors account for 1% of all human body tumors. They usually occur in the brain parenchymal and spinal cord[4]. Since histology proved not sufficient in predicting response and treatment outcomes, molecular expression has become common in characterization of tumors. This has attracted development of different technologies and algorithms in identifying this alteration e.g the 850K Illumina bead chip and conumee that aid in both methylation pattern brain tumor identification and copy number alterations as well.

Malignant glioma has an annual incidence of 6-7 cases per 100,000 and frequently occurs at age of 50-60 years in both females and males. Surgical resection and temozolomide therapy is the standard of care however most patients succumb to glioma within 15-20 months[5]. Histology

has proved to be not sufficient in stratifying patients based on clinical outcomes as patients with identical histological tumors often end up with different outcome[6]. Molecular characterization promises to significantly solve this problem[6].

Classification of glioma based on gene expression usually depends on key gene mutations and expression status such as *EGFR*, *NFI* and *PDGFRA* and *IDH1*[7]. This can classify the mesenchymal and proneural glioma subtypes. Most WHO grade I glioma are common in children and are characterized by *BRAF* mutations and fusion at *BRAF_KIAA1549* while grade II commonly known as low grade glioma (LGG) are seen in young adults and are characterized by R132 residue mutations at *IDH1*[8].

1.2.1 Loss of heterozygosity (LOH)

LOH at chromosome 10q has a prevalence of 60-80 % in both primary and secondary GBMs[9]. Since the frequency of this alteration is high it complicates classification of brain tumors although it's highly associated with primary events of GBM[9]. It is also a common genetic event in carcinogenesis of the majority of inherited cancer syndromes due to somatic loss of wild type alleles[10]. For example, a short arm at chromosome 11 that harbours tumour suppressor genes need to be inactivated in both alleles for LOH-wilms tumour to occur as shown in figure 1.2 below[11]. The presence of GBM is therefore well characterized by *IDH1/IDH2, ATRX* status, *EGFR, TP53* and *H3-K27M* mutations[12].
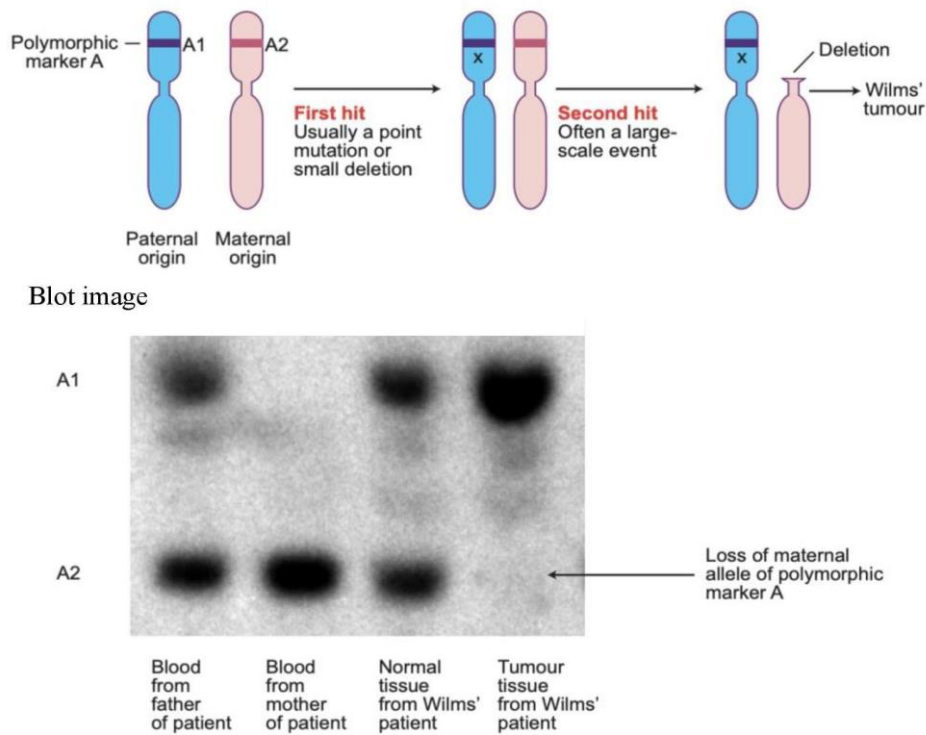
Figure 1.2 Process of development of wilm's tumour due to LOH in both alleles.

The second hit which deletes tumor suppressor genes at chromosome 11p13 is necessary for the pediatric kidney nephroblastoma to occur and not the initial small point mutations. Blot image illustrate the Southern blot that can be used to monitor both polymorphic paternal and maternal markers. It's clear that Wilms tumour has a deleted short arm of chromosome 11 due to lack of the band in tumour DNA blot. This figure was adopted from K. Brown et al., 2001.

1.2.2 *IDH1/IDH2* Mutations in glioblastoma

Isocitrate dehydrogenase-1 (*IDH1*) is an NDP-dependent enzyme found in the cytoplasm and located at chromosome 2q33.3 while *IDH2* is located at chromosome 15q26.1. These genes affect citrate metabolism thus leads to 2-hydroxyglutarate metabolism as shown in Figure 1.3 below [13,14]. The *IDH* enzyme is generally thought to catalyse the conversion of Isocitrate to 2-hydroxyglutarate which acts as oncometabolite thus causing tumorigenesis[15].
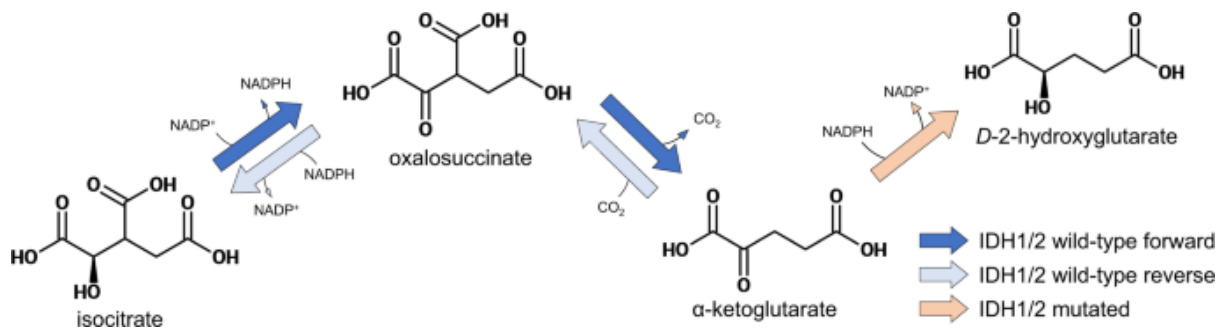
Figure 1.3 Forward oxidative decarboxylation reaction of wildtype IDH1/2 and reverse reductive carboxylation reaction of wildtype IDH1/2

The mutated *IDH1/2* form that leads to production of D-2-hydroxyglutarate. Figure was adapted from Molenaar, R.J et al., 2018.

Patients having *IDH1/IDH2* mutations tend to have better clinical outcomes than ones with *IDH1/IDH-2* wild type[14]. This is partly because presence of mutation at *IDH1/IDH2* enzyme reduce the ability of production of the coded proteins which are responsible for the conversion of isocitrate to alpha-ketogluterate and thereby producing *NADPH*, which in turn reduce reactive oxygen species[13,16]. The increased oxidative state due to presence of mutation is associated with carcinogenesis[17,18]. This also explains why the *IDH* mutation is rarely found at around primary glioma (6%) but pilocytic astrocytoma frequently found in progressed tumour such as secondary glioma (>80%), anaplastic astrocytoma and oligodendroglioma (>90%)[13]. The nature of mutation is a point mutation *IDH1- R132H* (arginine-histidine at position 132) and *IDH2 172* accounts for 3-5% of *IDH* mutations which are commonly found in oligodendrogliomas. *IDH1/2* mutations are common in young adolescents with glioma and rare in adults as shown in figure 1.3 [17,18]. Detection of *IDH1 R132H* mutation is by use of mouse anti human monoclonal antibody by Dianova with clone name H09 through immunohistochemistry techniques[17,18]. *IDH1/2* can be used to detect secondary glioblastoma using RT-PCR. The mutations are common in primary and secondary glioma as shown in figure 1.4 below[17,18].
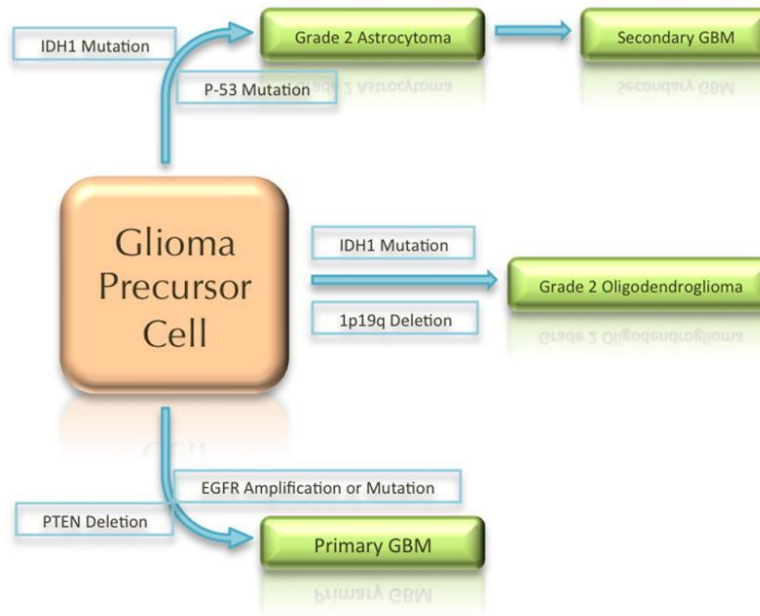
Figure 1.4 Main glioma pathways and distributions, where IDH1 mutation and chromosome 1p19q deletion defines grade 2 oligodendroglioma.

The *p53* mutation together with *IDH1* mutation is observed in grade 2 astrocytoma with the ability to progress to secondary glioma. *PTEN* deletion and *EGFR* amplification are common in primary glioblastoma and not *IDH* mutation. This figure was adapted from Cohen Adam et al., 2013

1.2.3 *P53* predict progression of astrocytoma to glioblastoma

Tumor protein *p53* gene is located at chromosome 17p13.1 and is known to regulate apoptosis, DNA repair and cell cycle induction[19]. Astrocytoma usually progresses, from premalignant state e.g low grade astrocytoma to malignant stages which include high grade astrocytoma, anaplastic astrocytoma to glioblastoma multiforme. Alteration of tumor suppressor gene *p53* has shown evidence in dissecting this tumor into different classes[20]. Polymerase chain reaction of loci coding for the *p53* is a common method to identify these mutations in chromosome 10 and 17 [20]. LOH loss on the other hand can be established using Southern DNA transfer analysis of somatic and tumor DNA from the same patients using polymorphic markers for various loci on chromosomes 10 and 17[21]. In most cases the somatic genetic mutations in the *p53* usually occur in the progression stage of astrocytoma and this explains why *p53* mutations occur at reasonable frequency > 25 % in anaplastic astrocytoma and glioblastoma (25-30%), low grade glioma (>70%) but very rare in low grade astrocytoma[9].

1.2.4 Methylguanine-DNA methyltransferase (*MGMT*)

MGMT gene is located on chromosome 10q26. Methylation of *MGMT* gene promoter is common in about 48% glioblastoma cases and in 85% IDH-mutants lower grade glioma[22]. *MGMT* enzyme is known to remove alkyl adducts from the O6 position of guanine thus aid in DNA repair and contribute in glioma alkylating drug susceptibility[13]. Monosomy of chromosome 10 and *MGMT* methylation do inhibit DNA repair by *MGMT* enzyme. It also contributes to temozolomide resistance[13]. The alterations are usually tested by using bisulfite sequencing and pyrosequencing.

1.2.5 Epidermal growth factor (*EGFR*)

EGFR is part of the *ErbB* receptor family and has tyrosine kinase activity. Mutations in *EGFR* are common in 60% primary tumors and about 10% in primary glioblastoma[23]. The common alteration of *EGFR* leading to its activations are amplification or over-expression of the gene in most of the affected cancers. The activation of *EGFR* can lead to cancer through dysregulation of the cell cycle and enhanced cell proliferation including invasion[24] as shown in figure 1.5. The frequency of these mutations differs from each site. Classical glioblastoma accounts for 95% of *EGFR* amplifications while 17% amplification accounts for neural and proneural glioblastoma. The dysregulation of *EGFR/HER1* occurs in most glioblastoma multiforme cases which is associated with growth and malignant transformation of the tumor[25]. Major probable mechanisms leading to dysregulation of *HER1/EGFR* include alteration of the receptor structure due to mutational changes. The common mutation being in-frame deletion of 801 bp in *DNA* sequence encoding extracellular domain thus leads to truncated and constitutive receptor *EGFRvIII* as shown in figure 1.5 and 1.6 below[25]. This mutated receptor is a target of most tyrosine kinase inhibitor therapies. Moreover, monoclonal antibodies can be helpful in blocking these targets.
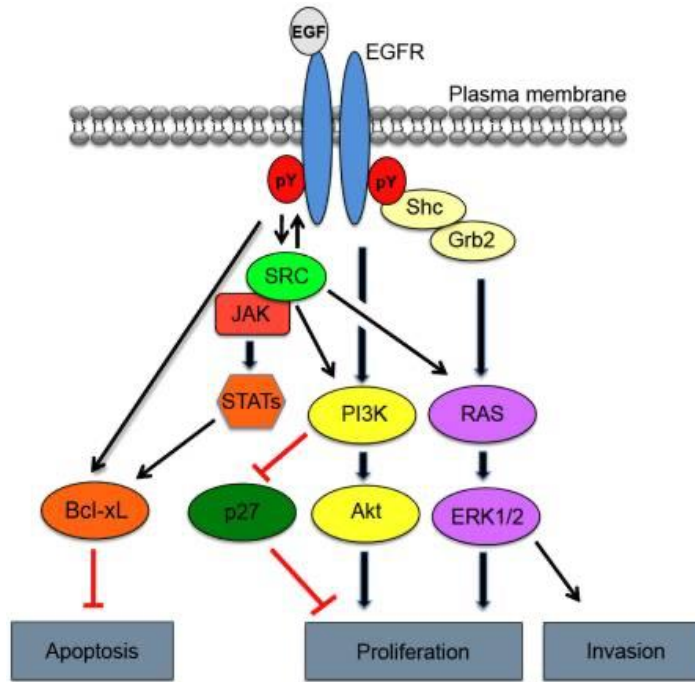
Figure 1.5 Sequence of events and gene products involved in transferring external signals to the nucleus as mediated by the EGFR pathway.

Figure was adapted from Hongsheng Xu et al., 2017.
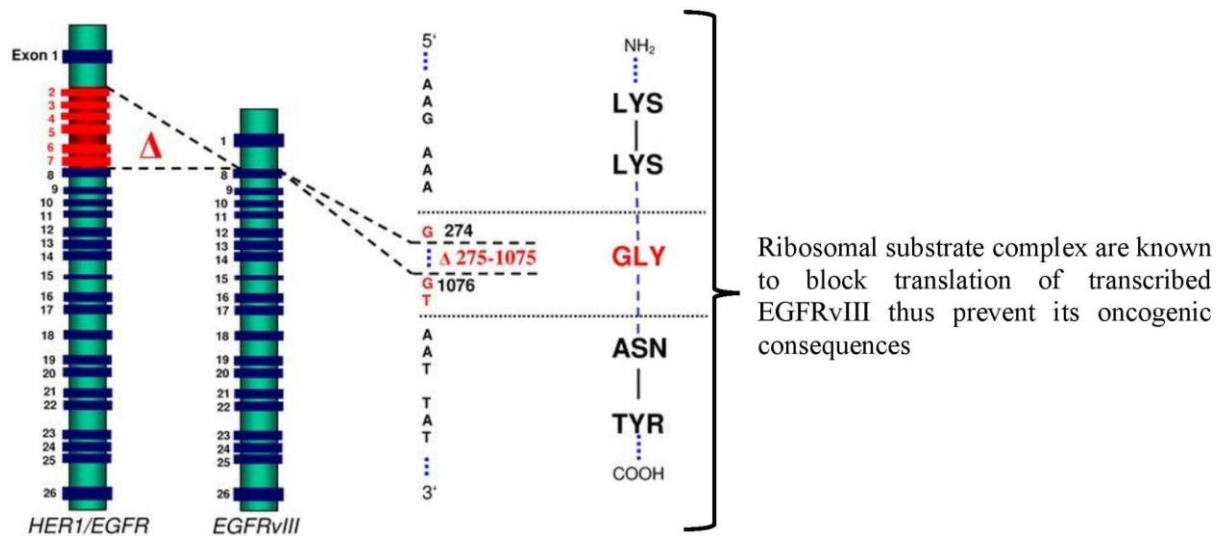
Figure 1.6. Structure of wild type HER1/EGFR and a mutated form of EGFR "EGFRvIII" genes.

This is brought about by deletion of 801 base pairs of exon 2-7. The right side shows the new glysine insertion as the result of the deletion at the fusion junction. This figure was adapted from Georg Karpel-Massler et al., 2010[25].

1.2.6 Stem cell makers of clinical relevance to glioma

The major molecular stem cell markers identified can be classified as membrane markers e.g *CD133*, podoplanin, *CD15*, and *A2B5*[26,27]. Second there are filament markers e.g nestin, RNA-binding proteins e.g Musashi-1 and transcription factors e.g *BMI1*, *SOX2*, and *Id1* [28–30]. Below are descriptions of some of these makers.

1.2.6.1 Nestin and *CD133* expression in glioblastoma multiforme

CD133 is a 5-TM glycoprotein antigen located in normal human hematopoietic cells, neural progenitor cells and also in malignant tissues[26,27]. *CD133+* (positive) cells have been recognized as a marker for the growth of cancer cells since they are required for tumour development. *CD133-* (negative) were shown to be unable to produce tumour[28–30]. It is also known to be a marker of poor survival in astrocytomas. In glioblastoma multiforme co-expression of *CD133/Ki67* and *CD133+* cells are also makers for poor disease outcome which has been reported to be <14 months overall survival[31].

Glioma stem cells (GSCs) are known to demonstrate high tumorigenic potential and this contributes to glioma invasiveness leading to incomplete surgical resection. *CD133* and nestin are known reliable markers for the GSCs[32–34]. Nestin is also a filament marker expressed in the primary central nervous system and expression increases with astrocytoma increases in malignancy[35–37]. It plays a crucial role in enhanced motility and differentiated status[36].

1.2.6.2 Podoplanin (Aggrus)

It's a transmembrane glycoprotein that is associated with poor prognosis in glioblastoma. *In vitro* studies have shown that it influences tumor cell migration and proliferation however there is limited information about its role in the glioma formation *in vivo*[37]. Some of the methods used to study the glycoprotein *in vitro* included genome-wide expression analysis and clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR-associated nuclease 9 (Cas9)-mediated deletion of podoplanin in patient-derived human glioblastoma cells[37–39]. The manipulated cells can be combined with organotypic brain slice cultures and intracranial injections follow into mice. Using tissue microarrays, Podoplanin was shown to be highly expressed in high grade astrocytomas but not in low-grade astrocytomas and thus reflects its

role in prognostic in this brain entity. It has also been shown to play a role in spheroid formations and glioma invasion[39–41]. It is a reliable maker for tumour initiating cells and interaction between C-type lectin receptors-2 (CLEC-2) is important for cancer metastasis and can induce platelet aggregation[39]

1.2.6.3 Cluster of differentiation 15- (*CD15*) and A2B5 maker

Cancer stem cells (CSC) offer an alternative novel approach to the treatment of glioblastoma. Cluster of differentiation-15 is a potential marker that has limited information *in vivo* however implantation of these cells in the mouse brain was shown to initiate tumour formation thus suggesting the possibility of *CD15* as a potential CSC marker[42]. However, in the primary glioblastoma cell line, *CD15+* were shown to have similar proliferative ability as *CD15-* cells[43].

A2B5 on the other hand is a surface glycoside that acts as 0-2A neural progenitor cell markers[44]. A2B5+ cells present the potential to differentiate to oligodendrocytes and few of them to neurons[44,45]. It's also associated with poor prognosis[35]. Apart from differentiation, *A2B5+* cells can distinguish glioma subtypes since A2B5- does not present with neural stem-like cell properties[44,45].

1.2.7 Methods for identification of molecular markers in glioma
1.2.7.1 An overview to CNV analysis using aCGH and EPIC BeadChip(850k)

Copy number variations (CNV) are common in both health and disease conditions. CNV ranges from 50 base pairs (bp) to > 450 kilobase(kb) in the human genome[46]. They are known to influence gene expression patterns or indirectly gene dosage hence responsible for phenotypic changes and disease occurrence [46,47].

Microarray-based comparative genomic hybridization (aCGH) allows high resolution DNA copy number alteration association with chromosomal abnormalities at genome level [48,49]. It can be routinely applied in basic research and clinical practise to detect CNV and Methylation defects [50–53]. aCGH depends on total genomic DNA from a test and a reference genomic material. In principle the two are hybridized together to normal chromosome at metaphase[54]. Blocking DNA is used to suppress signals from repetitive DNA sequences. The resultant

fluorescence intensities ratio is therefore proportional to DNA copy number between the test and the reference genome[54]. Once the intensity ratio is obtained from the microarray hybridization process, several steps including logarithmic transformation, segmentation and detecting change point are crucial in order to determine regions of gain and losses. Different bioinformatic packages and statistical models are available for change point detection and detection of DNA CNV [55,56].

In 2016 the 450K method was replaced by EPIC BeadChip(850K). The 850K is advantageous since it almost doubles CpG sites to >850,000 from the initial 450,000 CpG[57]. We therefore used combined data from the two BeadChips for our analysis. Recently epigenome-wide association studies (EWAS) have been used to associate altered DNA methylation and health outcomes [58]. Also the reliability of Conumee, ChAMP and CNVnalysis450k to CNV evaluation have recently been done. Conumee showed high reliability for HumanMethylationEPIC array at 57% while CNVnalysis450k had the highest reliability in 450K data at 43%[58].

Different genotyping CNV calling algorithms rely on probe intensities, log R Ratio (LRR) and allele frequency (BAF), to call CNVs[59]. The common programs used for calling CNV from genotyping arrays include Birdsuite, ipattern and pennCNV[59]. The major setback of genotyping arrays is the inability to detect inversion and translocations.

The DNA methylation arrays data is usually analysed using ChAMP, Conumee and CNV analysis 450K packages. These methods have the ability to detect CNVs within genes but not intergenic CNVs. Both methods operate under the principle that total methylation signal, which is the sum of unmethylated signal and methylated signal, has a direct relationship with copy number state[110]. In our study we used the Conumee package (Hovestadt and Zapatka). Conumee identifies CNVs by normalisation of experimental samples to reference samples using multiple linear regression[53,58,60]. It then takes the log2 probe intensity ratio between the reference sample and test sample. The individual intensities are then combined to bins and the genome is segmented based on equal copy numbers (Hovestadt and Zapatka). Though the method is widely used, limited research has been conducted to compare the results with gold standard methods like aCGH, an area I focused on in phase I of this thesis.

1.2.7.2 Fluorescence in situ hybridization (FISH) in non-stem glioma makers detection

FISH is applied to detect or screen for genetic chromosomal anomalies such as amplifications, co-deletion or chromosomal loss using a set of chromosomal probes[61]. In principle the methods use a hybrid DNA probe which is labelled using fluorescence nucleotides directly or using a reporter molecule that can be detected by antibodies or other affinity molecules indirectly. In oligodendroglioma 1p/19q co- deletion of short chromosome arm p and long chromosome arm q is a reliable marker that is absent from any other non-glial tumours[62]. This can be detected using FISH. Proneural tumor which is characterised by *CDK4* amplification and *MET* amplification in mesenchymal tumor can also be detected by FISH[63]. Commercial probes for detection Polysomy 7,9 and 19 (Centromere (CEP) 7,9 and 10) together with *EGFR* amplification, 10q23/*PTEN* loss and 1p36/19q/13 co-deletions in pediatric astrocytic glioma are also available for FISH method too[64].

1.2.7.3 Markers and common commercial antibodies for immunohistochemistry

Protein analysis in gliomas is important to detect pathological conditions. Different biomarkers which are detected in tumor tissues can give information about pathological processes, response to chemotherapy and even disease recurrence[65].

Oncogenic gene alteration results from point mutations, duplication, insertion, amplifications, deletion, hypermethylation or hypomethylation and translocations[66]. All these events may result in amino acid changes that may cause abnormal protein production. The resultant protein provides the basis of designing appropriate antibodies to detect the product in pathological condition[66]. Therefore, immunohistochemical diagnostics involve use of specific antibodies to differentiate normal protein (wild type) from the mutants[67]. Some commercial antibodies available are listed below which can aid in the diagnosis of some of the glial tumours.

IDH1 R132H canonical point mutation antibody detects *IDH1* R132H which indicates presence of diffuse glioma with a sensitivity/specificity of 80-100%[68–70]. It mainly stains the cytoplasm and acts as a weak nuclei stain. *H3K27M* antibodies stain tumour nuclei and detect *H3.1K27* M mutations in diffuse midline gliomas with 100% specificity and sensitivity[71]. *BRAF* V600E is a cytoplasmic targeting antibody that detects *BRAF* V600E mutation which are common in PXA and glioblastoma multiforme (GBM) glial tumours[66]. P53 antibody used to detect *TP53* point mutation antibodies in glioma and astrocytic tumours which leads to

increased expression of protein and stabilisation[70,72]. It stains the nuclei. *ATRX* antibodies are used to detect loss of *ATRX* protein expression in tumour cells and only present in normal cells. The *ATRX* loss is common in diffuse astrocytic tumors and secondary GBM[73]. It can also be detected through PCR and sequencing. Sometimes western blot is employed. Other commercially available antibodies include *EGFR* for detection of overexpression, *CIC* and beta-catenin antibody to access the canonical Wnt pathway activation though its sensitivity and specificity is not well studied[66]. Moreover, homozygous deletion of *CDKN2A* and *RB* deletions or mutations in GBM can be detected by either immunohistochemistry technique or FISH [66,67].

1.2.7.4 Sequencing and PCR based method in glioma marker identifications

Mutations in cancer reside in both coding and non-coding regions. Most studies have focused on the coding region, however whole genome sequencing enables us to study even the non-coding regions and infer causal associations[74]. Whole genome sequencing (WGS) has been instrumental in deciphering both clonal and subclonal evolution in cancer, especially the GBM recurrence[75]. Different disease aetiology such as *TERT* promoter contribution in disease progression was established through whole genome sequencing approach[76]. Different markers have been identified and RNA functional annotation platform for the non-coding region is also available in about 129 species[77]. Oncogenic processes have been linked to both genomic and epigenetic adaptation that are linked to conserved functional outcomes [74]. Many glioma markers can be identified through sequencing and PCR[62]. These markers include; *MGMT* promoter methylation in GBM which predicts better response to Temozolomide (TMZ) and *P13K* activation mutation in GBM[62]. Other makers benefiting from sequencing approach is the *IDH* missense mutation identification at arginine 132 which is common in oligodendroglioma and secondary GBM and *NFI* mutation in mesenchymal and pilocytic astrocytoma[66,75,78].

1.3 Classification of brain tumor

A brain tumor affects the normal function of the brain. In-general, we have primary brain tumors which are classified as low-grade e.g the low-grade glioma (LGG) and high-grade tumor e.g glioblastoma. Low grade tumors grow slowly but can transform into high grade. Most people usually die of metastasis which is simply the spread of the tumor from its primary site. Metastasized tumors are therefore types of secondary tumors. It is important to precisely classify this tumor for patient's management.

The classification of brain tumors was built up from histogenesis where tumors were histologically differentiated using haematoxylin eosin-stained sections on a light microscope. Immunohistochemical expression, lineage associated protein and ultrastructural were later used in the classification of tumour[79]. In 2016 the world health organization (WHO) introduced the use of molecular parameters along histological features in the characterization of brain tumors. This classification stratified tumors further e.g based on histology, the diffused astrocytic and oligodendroglioma could be classified into WHO grade I and II[79].

Using molecular parameters such as isocitrate dehydrogenase 1 or 2 (*IDH*) mutation status, we could have three groups which include *IDH*-mutant with retained nuclear transcriptional regulator *ATRX,* IDH-mutant with lost *ATRX* and IDH-wild type with nuclear *ATRX* retained. The *ATRX* status is usually determined using immunohistochemistry techniques[80]. Using the *ATRX* retained *IDH*-mutant status this category of tumors could be further divided into two based on copy number variations at chromosome 1 and 19, where 1p/19q co-deletion is a characteristic feature of oligodendroglioma while absence of the co-deletion leads to IDH-Mutant astrocytoma as illustrated in Reifenberger, G. et al. (2016) as shown in the adapted figure 1.7 below[80].

On the other hand diffuse astrocytic glioma and glioblastoma which belongs to WHO grade IV could be further divided into glioblastoma *IDH*-wild type WHO grade IV, glioblastoma *IDH*_mutant WHO grade IV and diffuse midline glioma H3K27M mutant WHO grade IV. Middle line gliomas are usually located on the thymus, brainstem or spinal cord[79]. Apart from the Glioma family the WHO 2016 further introduced tumour groups based on brain invasion as a criterion for atypical meningioma[79].
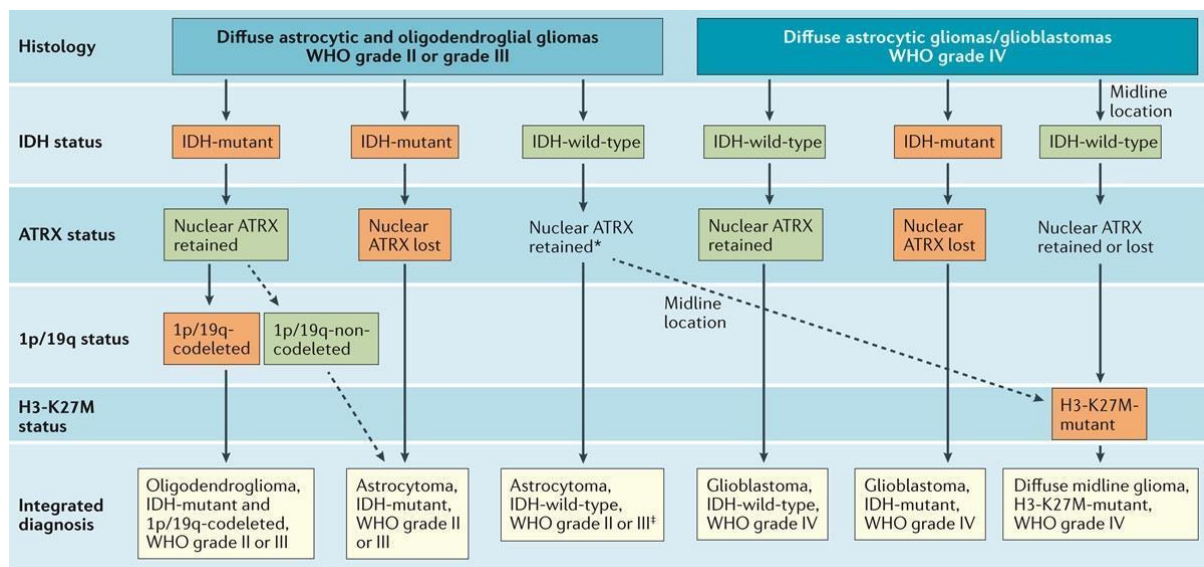
Figure 1.7 Histology and molecular classification integration of diffuse glioma based on WHO 2016 classification.

This figure was adapted from Reifenberger, G. et al. (2016).

1.3.1 Conflicts in brain tumour classification during diagnosis

Despite the development and improvement in brain tumour diagnosis there are still conflicts in brain tumour diagnosis in the clinics that were recently reported as shown in figure 8 below[50]. It is therefore important to integrate methylation patterns, molecular patterns including copy number alteration to strengthen the stratification of patients' risk beyond classes. This ambiguity in tumour classification usually costs the life of patients in some cases.
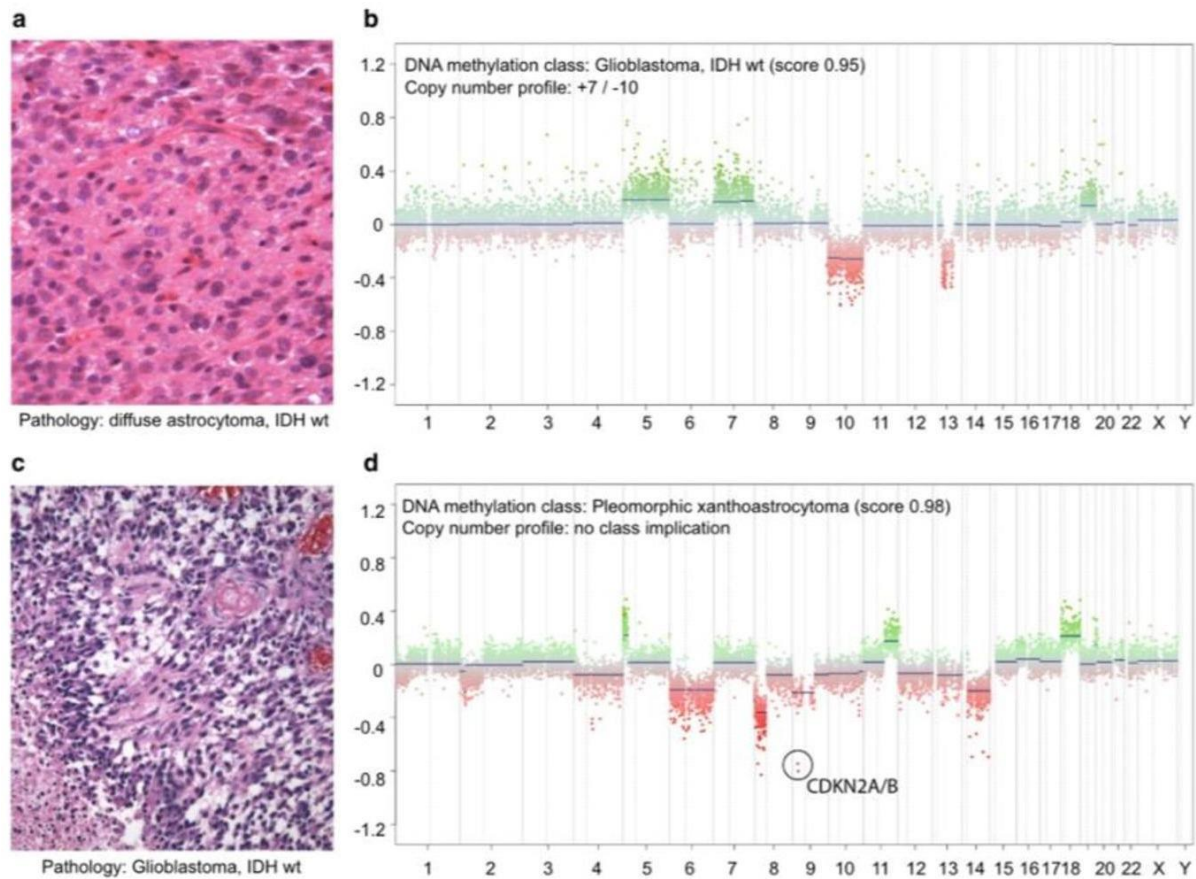
Figure 1.8. Conflicts in histological, methylation and copy number diagnosis of brain tumours.

Using microscopy diffuse astrocytoma, *IDH* wt (a) shows cell differentiation but may lack mitotic activity while glioblastoma, *IDH* wt (b) may show mitotic activity, necrosis and angiogenesis. The glioblastoma (c) and pleomorphic xanthoastrocytoma (d) are different entities from diffuse astrocytoma and glioblastoma respectively. This would affect patient management though such discrepancy may be resolved with inclusion of copy number profiles. This figure results were adapted from Capper et al.,2018

1.3.2 Methylation classifier and illumina epic beads in brain tumour diagnosis

Differential diagnosis of brain tumour depends on immunohistochemistry and molecular signatures such as *EGFR*, *MGMT* (promoter methylation status), 19q13.42,1p/19q codeletion, *MYC*, *MYCN* and *PDGFRA*. Immunohistochemistry involves the use of *CTNNB1* and *LIN28A* staining. These routine signatures have proved to be difficult to standardize. It was therefore hypothesised that methylome which is a combination of both somatically acquired DNA methylation changes could harbour unique properties that reflect the cell of origin hence can be a better classification of tumorous cells[81].

DNA methylation can result in dysregulation of cellular processes which are frequently taught to cause a disease. Addition of a methyl can lead to hypermethylation of DNA making it not accessible to key enzymes required in gene expression especially at promoter and enhancer regions. It has also been shown to contribute to silencing of specific tumour suppressor genes and this partly explains its role in cancer[82]. In diagnosis methylation patterns in the 850,000 CpG islands or > 450,000 CpG islands in illumina epic beads could be used to identify specific tumours and this is possible because the methylome remains stable in the cause of disease.

Recently the methylome classifier was developed based on clusters of this pattern and act as a novel tool that is used in research with promising application on the bed side even for cases which can't be classified by histopathology[83]. Following clinical experiences, it has been reported that the methylome is reliable in better classification of the tumour. In addition, copy number alteration would stratify patients better and provide better clinical management[84]. WHO brain tumour classes that have benefited much from the methylome based classification include the medulloblastoma and supratentorial primitive neuroectodermal tumors (PNET)[50]. The molecular classification was shown to be superior in-terms of risk stratification than traditional histopathology[85]. The details of this classifier can be found elsewhere https://www.molecularneuropathology.org/mnp. Details of the use of this classifier and results interpretation was also published [86].

The methylome classifier was developed by a random forest algorithm. It requires one to load data information of IDAT files which are obtained by processing patients' DNA from FFPE and other sources into 450k or 850k illumina epic beads. In principle the Epic beads contain probes which hybridize with patient DNA. This beadarray technology uses beadChips coated in microwells with multiple copies of oligonucleotide that targets specific genome locus. When a DNA fragment passes over the BeadChip, each probe binds to a complementary sequence in the sample DNA. During excitation by a laser, the nucleotide label emits a signal. This intensity of signal provides information about the allelic ratio of the position. In brain tumor research we have 450K and 850K epic arrays. The Infinium methylation arrays therefore detect cytosine methylation at single CpG site level.

The methylated bead type and unmethylated bead type hybridize to the test DNA as shown in figure 1.9 below. The methylated intensity is then divided by the sum of methylated and unmethylated intensity to get the beta value. As part of our findings in this thesis we provided

evidence that the copy number alteration could also be estimated from the infinium epic array data with > 95 percent agreement to comparative genome hybridization assay (data shown on results section). Unlike beta value calculation of the copy number alterations are calculated by addition of the methylated and unmethylated intensities and then compared with a reference flat genome sample[50].
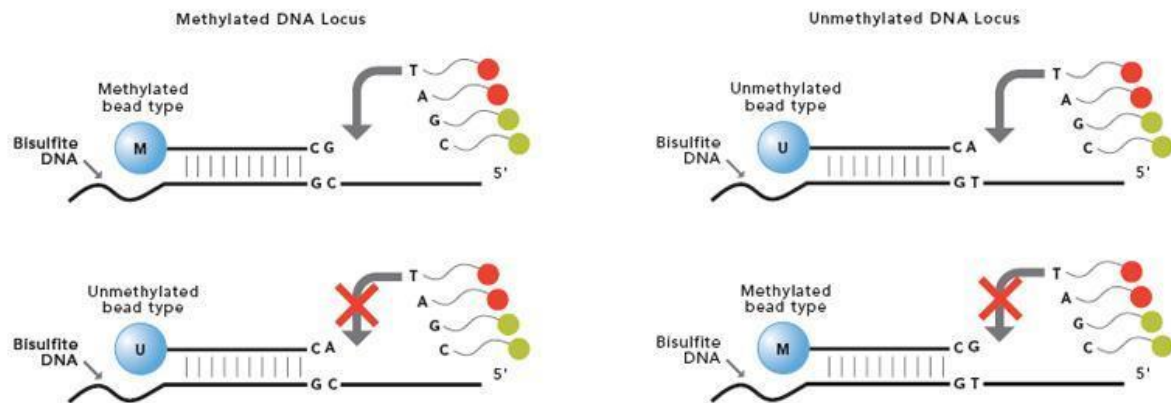
Figure 1.9. Cytosine is converted to uracil (U) which is read as thymine after bisulfite conversion in unmethylated CpG while for methylated CpG the cytosine is prevented from conversion as shown by methylated probe (M).

The U and M types of probes are used to detect methylation state of DNA samples through hybridization and termination by single base extension. Mismatch will terminate the reaction and green and red intensities produced based on CpG state. This figure was adapted from www.illumina.com

1.4 Molecular and histopathology diagnosis of diffuse glioma

The 2016 WHO classification divides diffuse glioma into three grades namely; WHO grade II, grade III and grade IV. In other words, these are low grade glioma, anaplastic tumorous cells i.e characterised by poorly defined cellular differentiation and glioblastoma respectively. In general, the diffuse gliomas are known to exhibit infiltrative growth in the neutrophil. Stratification of these families of tumour is important in precision medicine especially due to its possibility to enhance prolonged survival. During diagnosis, this type of tumour is difficult to characterise using the histology gold methods. It is therefore recommended to include molecular markers combined with the methylome classification[87].

Normal grey matter or white matter develop lesions during the cancer development stages. Malignancy is usually characterised by necrosis, angiogenesis and mitotic activity, while oligodendroglial tumors are characterised by extensive calcification which is a deviation from its normal round cells with dense chromatin pattern which stain negative with GFAP [87]. Moreover, Ki67 proliferation index is reliable in differentiation of grade II from grade III tumours.

Different biomarkers have been suggested in glial tumour for example the expression of platelet-derived growth factor receptor alpha (*PDGFRA*) has been suggested as mitotic marker in medulloblastoma during mitosis[88]. These markers were also reported to increases medulloblastoma migration and also enhance expression of *MAP2K1/MEK1, MAP2K2/MEK2, MAPK1 (p42 MAPK)* and *MAPK3 (p44 MAPK)* phosphorylation in a dose-dependent manner[88]. The chromosome 17p deletion has been shown to occur in up to 50% medulloblastoma cases with short survival however its reliability is not yet clear.

In glioblastoma patients, around 70-80% cases show over-expression of epidermal growth factor receptor *(EGFR)*. Common deletion at *CDKN2A* (p16) is seen and sometimes in a few cases *MDM2* amplification is evident[89]. For lower grade astrocytoma which occurs in younger patients (45years) progression to GBM is typically detected by mutations in *TP53*. TP53 usually acts as a stress inducible switch that allows cells to undergo G1 arrest or apoptosis. Most importantly, the low grade and secondary glioma also exhibit mutations in isocitrate dehydrogenase-1 *(IDH1)* and *IDH2* at high frequency[89].

1.4.1 An overview to molecular diagnosis of glioma

In the molecular era, the 2016 world health organisation classification of CNS tumours integrated molecular parameters along histology in the classification of tumours[79]. This was allowed accurate diagnosis and improve patient management. Among the categories of tumour that majorly benefited from this approach is oligoastrocytoma. This category had issues with inter observer histological discordances. Based on 1p/19q codeletion and IDH mutation status it was possible to differentiate astrocytoma from oligodendroglioma[90]. Patients with IDH mutation and 1p/19q codeletion are known to have favourable clinical outcomes in low grade glioma unlike those without IDH mutation which present like glioblastoma[90,91]. Those lower grade gliomas with no 1p/19q codeletion were shown to have IDH mutation with most of them

having additional TP53 and ATRX inactivation[112]. Recently, a machine learning approach for classification of central nervous system tumors based on the analysis of genome-wide DNA methylation patterns was developed and published [86]. Using this tool tumor classes (or subclasses) can now be dissected better and diagnostic accuracy is improved, though, for some cases the "classifier" is not able to give a conclusive prediction). A rough overview of the Copy Number Variations (CNV) of methylation classes used in the "classifier" was reported in [Capper et al., Acta Neuropathologica 2018][50], however, the CNV are not well studied and integrated into any standardized diagnostic procedures, yet. It seems promising to achieve improvements in methylation-based diagnostics by establishing an approach to systematically include CNV information.I therefore, proposed to find characteristic patterns of CNV in tumor cases and also evaluate if copy number alterations can aid in stratifying patients further.

1.5 Glioblastoma necrogenesis and FAS Expression

The presence of necrotic tumour in the astrocytic neoplasm is an important feature for glioblastoma diagnosis by a pathologist. There are two distinct forms of necrosis. One is characterised by presence of irregular small, band like shaped glial cells and the other form is characterised by extensive area of necrosis containing necrotic vessels and tumorous cells. The second type is more common in primary glioblastoma but absent in secondary glioblastoma[89]. *FAS/APO-1 (CD95, APT1)* is predominantly expressed in glial cells and known to play a role in regulating cell membrane apoptotic associated proteins. It is also more frequent in primary glioblastoma (100%) than in secondary glioblastoma (21%) suggesting a key clinical role during the initial stage of glioblastoma development[89].

1.6 Molecular stratification of glioblastoma along low grade and anaplastic astrocytoma

The World health organisation divided astrocytoma grades based on clinical presentation (non-infiltrative or infiltrative). WHO grade II which mainly include low grade astrocytoma e.g pilocytic astrocytoma. They are mainly found in children and have a low fatality rate due to their non-infiltrative ability. WHO grade III includes anaplastic astrocytoma. Grade IV astrocytoma is the most aggressive and mainly includes glioblastoma (GBM). Patients with grade III and IV have average survival of 5-2years respectively[92].

Astrocytoma is characterised by presence of *IDH*-mutations in >= 30% cases and *TERT* mutation in >=50% of patients, while glioblastoma is characterised by *MGMT* methylation[69,93]. Most molecular alterations in astrocytoma are enriched in the calcium signaling pathway which plays a critical role in its tumor progression and prognosis[93]. There are four molecular subtypes of glioblastoma namely; Proneural, mesenchymal, classical and neural. These subtypes are known to respond differently to treatments. Classical subtype is characterised by chromosome 7, and *EGFR* (>94%) amplification[94]. Classical subtype also shows chromosome 10 deletion and homozygous deletion of Ink4a/ARF locus[95]. Mesenchymal type is characterised by elevated expression of Chitinase 3-like 1 (*CH13L1*) and *MET* proto-oncogene. Moreover, *NFI* mutations are associated with mesenchymal type. In proneural types, *PDGFR* alpha amplifications, *IDH1* mutation and *TP53* loss of heterozygosity mutations are important in differentiating this group. Finally, the neural type is identified by differential expression of neural makers for example *GABRA1, SYT1, SLC12A5 and NEFL*[94].

The clinical presentation of glioblastoma can be differentiated further from other grades by differential gene expression of key genes, e.g Over expression of vascular cell adhesion molecule I (*VCAMI*) is unique in low grade glioma and anaplastic astrocytoma but not GBM[96]. Secreted modular calcium-binding I (*SMOCI*) and Adenosine A3 receptor (*ADORA3*) are over expressed mainly in anaplastic astrocytoma. Overexpression of Aquaporin (*AQP1*), Topoisomerase (DNA) II alpha (*TOP2A*), ATP- binding cassette, subfamily C, member 3 (*ABCC3*), Thymidylate synthetase (*TYMS*), stabling I (*STABI*), Chemokine (C-C- motif) ligand 2 (*CCL2*), Matrix metalloproteinase (*MMP*) and Matrix Gla protein (*MGP*) are mainly overexpressed in GBM[92]. The Chitinase 3-like 2 (*CHI3L2*) and neuromedin B (*NMB*) are overexpressed in all astrocytoma grades[92].

Aim of the study / research hypothesis

The aim of the study was to explore the importance of copy number variations in diagnostics of brain tumor entities. I also proposed to provide evidence of reliability of epic illumina data in copy number estimation by comparing with aCGH data. Different brain tumor entities were analyzed in detail for specific patterns of copy number alterations. I tested, if CNV profiles can be of help, to identify subclasses, establish a diagnosis and grading in cases that are or not classifiable well by the methylation-based tumor classifier. Based on this, I hoped to come up with initial profiles and a systematic approach for diagnostics that would include the CNVs in the methylation-based classifier. I also planned to correlate CNV profiles with clinical parameters and analyze if CN data is of use in grading within methylation classes.

2.0 Materials and Methods

2. 1 Comparative analysis of CNV status in paired data set between aCGH and Illumina epic (450K and 850K)

2.1.1 Study sample/ cohort

The cancer genome atlas database helped to retrieve the astrocytoma and glioma data. For the first phase, I used a total of 61 paired samples from 450/850K data and aCGH data for which I did all the analysis during my stay at DKFZ. Once tumors are collected for diagnostic purposes, they provide a source of samples for research investigation too. DNA is then extracted from formalin-fixed paraffin-embedded (FFPE) tissue with estimated 75% or above tumour content. An automated extraction is done with a Maxwell system (Promega, Fitchburg, WI, USA). Maxwell 16 FFPE Plus LEV DNA Purification Kit and the Maxwell 16 LEV RNA are used using the FFPE Kit according to the manufacturer's instructions[84].

2.1.2 CNV calling from aCGH data

Human methylation arrays imaging and array comparative genomic hybridization data are often accompanied by noise which challenges the CNV threshold determination. To reduce the noise and need to detect significant CNV I opted for log2 (test(T)/reference(R)) analysis of the intensity values in both data sets. I followed heuristic thresholds for calling copy number alterations (CNVs) in tumor samples[97]. In brief, humans have two copy numbers in all autosomes (diploid state) however in cancer duplications or losses occur. In principle the hybridized genomic segments in a microarray chip/slide are composed of Cy3 dyes for test and Cy5 dye for the diploid reference. Upon hybridization the test and reference intensities are obtained. The test intensity can be low or high compared to reference and this defines loss and gain of copy number as reference to diploid state[97]. For a single copy gain, assuming 50% tumor clonality then R was estimated to be >log2 (2:6/4). This corresponds to -1.0, -0.4, 0.0, 0.3 and 0.6. We allowed for random noise of + /- 0.1, and the default log2 cut off ratio for my alterations were; deletions < -1.0, Loss < -0.25, gain >= 0.2 and amplification>= 0.7. We further considered plateau (plot p) as described in the DNAcopy package[98] and more guidance from Antonie M. Snijders et al.,[51]. Following the sample plateaus (plot p) and description of cutoffs of microarrays data I settled on a cut-offs for loss (log2 ratio< -0.2), gain (log2 ratio >

0.2) and balance fall between (log2 ratio >= -2.0 to log2 ratio <0.2). To test the reliability of cutoff values, I analysed each individual sample and checked the plot p trends. I observed that most balanced segment alteration aligned to log2 ratio value close to 0.0 (+/-0.5) or between the specified balanced cutoff window. I therefore considered that an optimal established CGH baseline corrected cut off for our data was obtained at 0.2 and -0.2 log intensity values.

2.1.3 CNV calling from EPIC and HumanMethylation450 BeadChip data

I did genome-wide copy number analyses as previously described, using the in house conumee R package version 1.9.0 (Hovestadt V, Zapatka M). Conumee uses the combined unmethylated and methylated intensity and normalises it with a normal set of controls (to control bias). This is followed by combining close probes to form bins with a minimum number of probes and size[97,99–101]. Methylation data copy number alterations of genomic segments were inferred from the methylation array data after additional baseline correction with cutoff of 0.1 for gains and −0.1 for losses on log2-scale (https://github.com/dstichel/conumee). Copy number plots for the methylation set of samples were generated using the CNV.genomeplot for the complete genome and CNV. Detailplot for a specific chromosomal detail region. Segments and copy number profiles for the methylation data were obtained from these data in each individual set of samples used in the analysis. The method is described in details elsewhere http://bioconductor.org/packages/conumee/.

2.1.4 Creating G-range segments from both data set

Human genome is usually presented as a linear sequence divided into chromosomal range position[102]. To get a complete set of data between the two paired methods, I used the g-range package and disjoin to align segments with similar chromosomal range between the two data sets. I then conducted an annotation of the copy number state between the two data sets. To do this I opted to take care of any overlaps and adopt a simple match approach to compare each segment call at a given position between the two methods. Large segments were also weighed more than small segments. This was done by eliminating overlaps in the chromosome segment between the data points using the genomic range package and disjoin function with modifications[102].

2.1.5 Statistical analysis in paired segments and computing overall agreement analysis

All the statistical analysis was done in R version 3.6. The matched Copy Number Variations (CNV) profiles generated from micro-array based comparative genomic hybridization (aCGH) and CNV methylation data were compared in all the 61 paired samples. For Samples which baseline correction and proper pair match could not be achieved, I eliminated them from the analysis(n=6). A match or agreement was defined as a unique/similar loss, gain or balance call in a specific range segment between the two methods. Percent agreement was calculated to show the extent of agreement between the two methods[103].Overall the two methods were considered to agree if percent agreement of segments classified were $\geq$ 79.5 percent. Figure 2.1. below shows the analysis plan I used and further statistics interpretation.
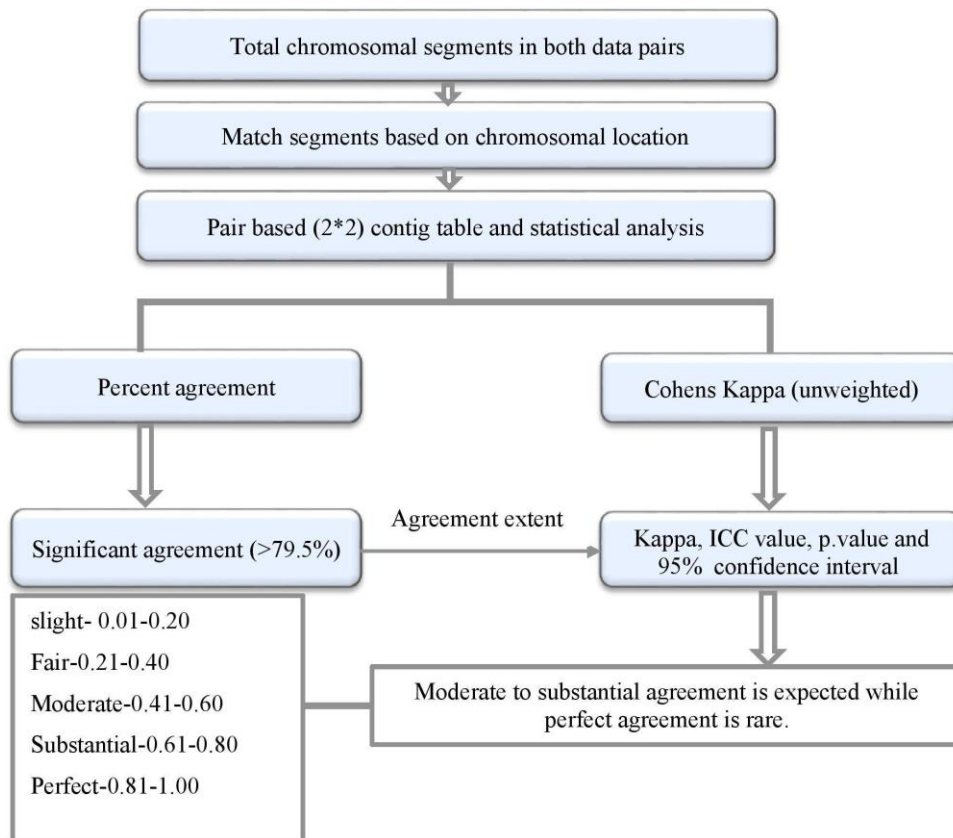
Figure 2.1. Steps used in the association and agreement calculations of paired chromosomal segments CNV calls between aCGH and Epic BeadChip.

Kappa statistics interpretations are indicated on the grey outline shapes with no fill.

To get more insight to the agreement, I further checked the interrater reliability of the two methods to ensure that the agreement is not by chance. This was assessed using unweighted Kappa statistics since I had more than two categories (gain, loss and deletion).

The Kappa value is usually similar to intraclass correlation coefficient (ICC) and this enabled me to confirm the agreement extent calculated beside checking reliability. I selected a two-way random effects ICC model which fitted well with our dataset. This model is reliable to assess clinical methods agreements or reliability especially for routinely planned clinical use methods[103]. The agreement estimates and their 95% confidence were calculated in Rstudio based on mean rating k=3 (balance, loss, gain).

2.2 Astrocytoma BeadChip EPIC data CNV analysis
2.2.1 Astrocytoma data retrieval

DNA methylation data (Illumina Epic Human Methylation 450k) was downloaded from the TCGA (n=116). The so generated raw methylation data was analyzed using the standard pipeline established in the department using Hovestadt V, Zapatka M. conumee: http://bioconductor.org/packages/conumee/, with minor modification in algorithms. All further statistical analyses were done using the programming language R. The copy number alteration was classified as gain, amplification, loss or deletions based on log2 intensity values in specific chromosomes. For example log2 of copy number > 0.1 was considered a gain while log2 of copy number < -0.1 was considered a loss. This analysis was done as described in the Bioconductor package with minor modification. https://bioconductor.org/packages/release/bioc/html/DNAcopy.html. Canonical pathway and network analyses were conducted using Qiagen's ingenuity pathway analysis (IPA) tool (www.qiagen.com/ingenuity). Hierarchical clustering was done on R version R 3.6 (http://www.R-project.org/).

2.2.2 Kaplan meier survival analysis

Survival was simply considered as the time point up to occurrence of an event (death) while survival analysis was the analysis of such occurrences in the data[104]. The data contained WHO grade classifications and time to event (death). With this type of data, I aimed to analyse the fraction of individuals that were living after being diagnosed with grade II or III astrocytoma based on WHO grading system. The ultimate goal was to see if the survival function differs

from each grade. This would be an indicator of severity of disease. In the event I found no difference in the survival functions I opted to screen for sub-groups using the copy number alteration and test their survival too. Kaplan meier estimates are reliable in computing the survival over time despite the challenges observed during follow up of individuals over time before the censored events[105,106]. This method simply computes probabilities of occurrence of events at a certain point of time and multiplying the successive probabilities with earlier probabilities to get the final estimates[105,106]. All the survival analysis was done on R.

2.2.3 Hierarchical clustering

I used unsupervised clustering approach for the 116 Astrocytoma samples based on CNV.I further used hclust in stats package and agnes in cluster package for the agglomerative clustering approach. I adopted the agglomerative clustering which combines two neighbour cluster into a larger cluster in what is commonly known as bottom to top approach. At first, I applied the four major distance measures (Euclidean, Manhattan, Minkowski and pearson) with three dissimilarity combination (Complete, average and ward) since I was not able to accurately choose the best methods for my copy number data clustering approach at the first time. All could be used however Pearson Ward-D2 provided better clusters with distinct copy number variations. I therefore concentrated with the Pearson correlation metric and ward as the agglomeration method. This combination allowed me to measure linear dependencies within the CNVs and also to minimise the total within cluster variance by merging clusters with minimum variance between clusters [107].

2.2.4 CNV frequency and heatmap analysis

This was conducted using the copynumber package (Gro Nilsen, 2020) in R. Briefly using this package, I first organised each data set into a matrix. I ensured that each row represented a probe and each column had chromosome and corresponding genomic position. The other columns had Log2 transformed copy number intensity values and specific sample identifier. After initial data cleaning considerations, I continued with segmentation and common break points were obtained. This was done in each sample with a multipcf function. I applied a gamma parameter to determine a penalty for each discontinuity obtained during segmentation. I then visualized the data by using plotSample and plotChrom in order to plot the data along the genome for each given chromosome. For heat maps I used plotHeatmap function.

2.2.5 Identification of the canonical pathways associated with the altered genes

To get more insight to the identified altered region I further retrieved genomic ranges of human genes using the biomart library genes. I then used the ranges to annotate our altered segments. This approach allowed me to compile the gene list which I uploaded on IPA in order to identify the pathways associated with the altered region. I adapted ingenuity pathway analysis in order to explain the pattern expression especially in the altered genes as shown in figure 2.2 below. I also aimed to identify biological key regulators and relevant pathways that could explain the role played by the altered genes in astrocytoma grade II progression to glioblastoma multiforme. Details on the application of ingenuity pathway can be found in user data sheet[108]
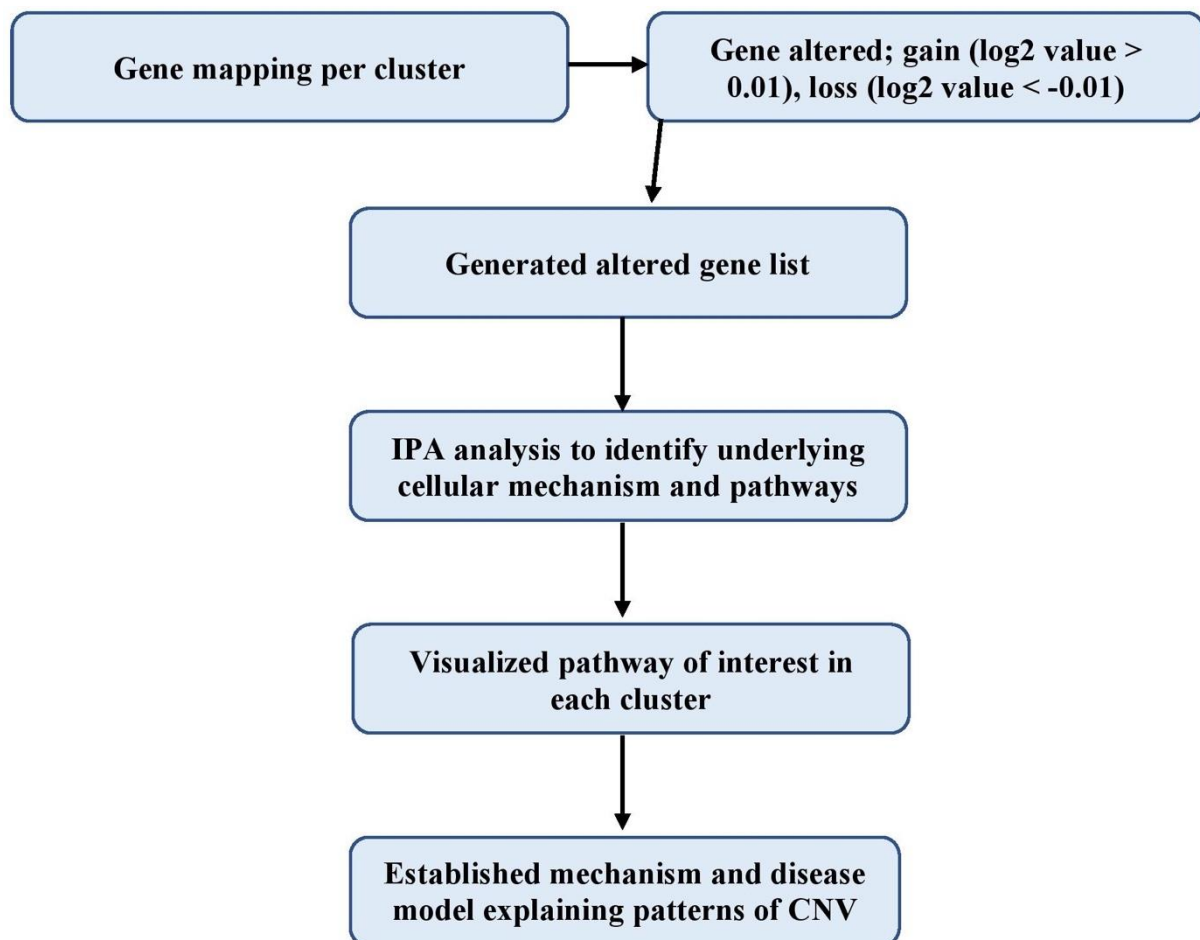
Figure 2. 2 Groupwise canonical pathway analysis workflow adapted in our analysis for each sub-group identified in the astrocytoma samples.

2.2.6 Methylation profiling classifier to confirm the astrocytoma classes

The methylation classifier was built based on a random forest algorithm. Based on 450K and 850K data unsupervised clustering was done on a reference cohort and each loaded sample is classified by comparing with this cohort [50,86]. For correct prediction of a tumour class a threshold of ≥ 0.9 is required between the sample and the reference set. Therefore, to confirm the methylation classes of each sample per group, I uploaded the 450K or Epic output data of green and red-channels (.idat files). The upload was done in the classifier as directed in the classifier website. www.molecularneuropathology.org. This allowed us to get the methylation classes along each CNV clustered group sample.

2.3 Identification of subclasses in low-grade and high-grade glioma based on CNV between cases and controls.

2.3.1 Data retrieval from the GDC national cancer institute (NCI) data portal

The retrieval and analysis of circo plot was adopted from the TCGA workflow with minor correction. The glioblastoma (GBM), low grade glioma (LGG) and sarcoma (SARC) data set were retrieved from the GDC NCI data portal (https://gdc.nci.nih.gov/) using Bioconductor package TCGAbiolinks[8]. I used the main functions GDCquery, GDCdownload and GDCprepare respectively to conduct the search. I further downloaded and loaded the data as an R object as described in the workflow. All analyses were done in R version 3.6.

2.3.2 Controls choice basis for CNV profiles and frequency in our data set

I selected case control samples from the GDOC databases. The samples were mainly from the TCGA project and Rembrandt databases. I considered blood normal samples as control since I assumed with high confidence that the samples came from diploid cells. Unlike tumour–adjacent tissue samples blood normal control is usually not contaminated by tumours during early tumorigenesis events. Moreover, it also fulfilled the batch effects considerations. Batch effects ensure that the samples come from the same laboratory and were processed with the same reagents.

2.3.3 Determination of the localization of the mutation frequency

The circo plot provides the chromosome position and specific gene altered per chromosome. However, to identify the frequency of each mutation based on position located in more than 400 cases and controls per group. I used the DNA copy package to process the log-intensity values[97,107]. Prior to determining the frequency and gene localisation, I first established the possible cluster per each glioblastoma multiforme group i.e I identified subclusters per each primary tumour, recurrent tumour or metastasized tumors.

2.3.4 Hierarchical clustering

I used a similar approach as applied in our astrocytoma data set as reported on section 2.2.3 above. In brief I adapted unsupervised clustering. I then clustered samples based on their log2 intensities values which gave basis on the copy number variations. I used hclust in the stats package and Agnes in cluster package for the agglomerative clustering approach. I adapted the agglomerative clustering which combines two neighbour clusters into a larger cluster in what is commonly known as bottom to top approach. Since I previously showed that Pearson correlation matrix and ward agglomeration method provided better clusters thus I depended much on results from this method. This combination allowed me to measure linear dependencies within the CNVs and also to minimise the total within cluster variance by merging clusters with minimum variance between clusters [97].

2.3.5 CNV frequency and heatmap analysis

This was conducted using the copy number package in R[97,98]. Briefly using this package, I first organised each data set into a matrix. I ensured that each row represented a probe and each column had chromosome and corresponding genomic position. The other columns had Log2 transformed copy number intensity values and specific sample identifier. After initial data cleaning considerations, I continued with segmentation and common break points were obtained. This was done in each sample with a multipcf function. I applied a gamma parameter to determine a penalty for each discontinuity obtained during segmentation. I then visualized the data by using plotSample and plotChrom in order to plot the data along the genome for each given chromosome. For heat maps I used the plotHeatmap function.

3.0 Results

3.1 Reliability of EPIC BeadChip (450/850K) and aCGH data in copy number variation calling in brain tumor samples-phase 1

3.1.1 Rationale for phase I results

Following technological developments, different methods have been developed that use EPIC Beadchip (850K) and HumanMethylation450 Beadchip (450K) data to call CNVs. The three major analysis packages include ChAMP, Conumee, and CNV analysis450k. However no major research has been done to validate or compare the CNV calls of these packages with the gold standard aCGH data on DNAcopy package. In these results set, I therefore aimed to provide evidence on the comparison between microarray-based genomic hybridization (aCGH) CNV calls and 850K combined with 450K data Conumee CNV calls analysed using paired samples (n=61). In specific I opted to determine percent agreement and reliability of the CNV calls from 450K/850 K methylation epic and aCGH paired data as shown in results below.

3.1.2 Genome wide aCGH data coverage in the CNV analysis in 61 samples

The sum of the length of all the autosome and sex chromosomes based on GRCh37 is 3095677412bp (3095.68Mb) per individual. I used the chromosomal length as reference to compute genome wide coverage of the two methods. The average segment length generated by the aCGH was 2771397022bp (2771.40Mb) per sample. This accounted for 90% genome wide segment length coverage and 96% in the autosomes only as compared with the GRCh37 total chromosomal length.

3.1.3 Genome wide methylation data coverage in the CNV analysis in 61 samples

The average genome wide chromosomal length covered by the methylation 450K or 850K data was 2746234758bp (2746.24Mb). This accounted for 89% of the total chromosomal genome length and 93% coverage for the autosomes.

3.1.4 Pairwise comparative analysis of paired data between aCGH and methylation results

A total of 4314 short segments had missing information on the aCGH data but had data for the methylation. This segment with missing data from aCGH was 1366702bp (1.37Mb) long on average. For the methylation dataset I observed short segments (3529) of around 2105653 (2.11Mb) on average that had no data for methylation but had data for the aCGH method. A summary of the chromosomal length covered by both methods is shown in table 1 below. The total percent matched and unmatched lengths of mapped chromosome positions were 74% and 19% respectively of the total length of chromosomes observed. The remaining 7% length accounted for the missing data in either of the cases. The methylation data contributed 3% of all lengths missing data while aCGH data contributed 4% lengths of the missing data. The average length covered by each method is shown in table 1 below. To calculate percent agreement, I considered the paired complete matched and unmatched data. I observed 80% agreement between the two methods.

Table 1 Summary of chromosomal length covered by the two methods

	Match	aCGH no data	Methylation no data	Unmatched
Segments	4310	4314	3529	1934
Average length (Mb)	2117.87	96.65	121.82	531.71
Percent coverage	74%	3%	4%	19%

3.1.5 Individual case analysis, agreement and chance agreement determination

To test the extent of agreement, I computed kappa statistics for each paired sample. I observed 92% of the sample had true agreement ($p < 0.05$) while in the remaining 8% (5/61) the samples had high kappa p-value ($p > 0.05$) which may be attributed to experimental errors or chance. I further calculated the percent agreement of the sample pairs which had true agreement. Interestingly I observed 88.3% agreement between the two methods based on chromosomal length weighted agreement and 80% unweighted agreement. To get more insight into the data I computed the interrater coefficient correlation (ICC) which was similar to the kappa k=0.57 and 95% C.I of 0.52-0.61. In summary for all the sample paired segments unweighted Kappa

and ICC value was 0.57 with a 95% confidence interval of 0.54-0.61 as shown in figure 3.1 below. In specific 7%,11%, 40%,35% and 7% of the paired samples were rated as slight, fair, moderate, substantial and perfect agreement respectively by kappa (k-value).
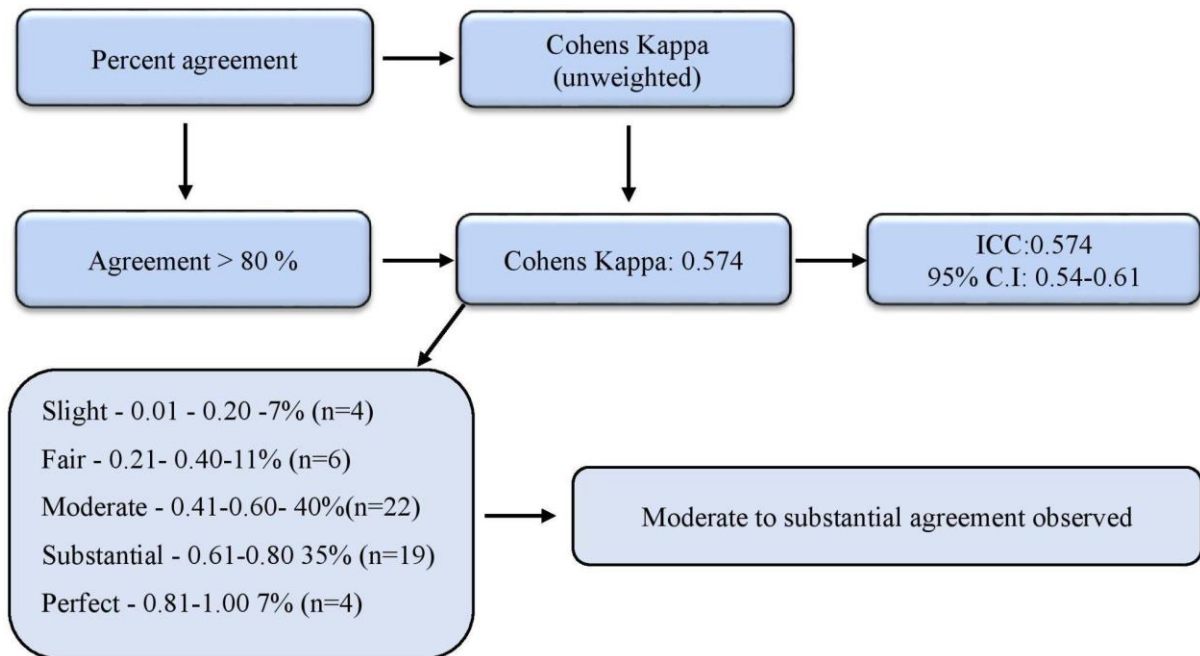
Figure 3.1 Percent agreement and Cohen's kappa value of 0.574 (p<0.05) that rule out any possibility of chance agreement between the two methods.

Often the ICC value is used along the Cohens value. Substantial to moderate agreement observed is sufficient for clinical diagnostic methods applications.

Moreover, a balance to gain or balance to loss disagreement is usually considered less severe than a loss to gain disagreements. I therefore computed the percent disagreement between the two methods. In an unweighted calculation I observed 13.5% were loss to balance/balance to loss, 5.7% were gain to balance/balance to gain and 0.8% accounted for gain to loss or loss to gain mismatches. Based on chromosomal length I observed 8.4%, 3.3% and 0.1%; loss-balance, gain to balance and loss to gain mismatch respectively. Table 2 below shows a summary of the segment's agreement.

Table 2 Overall segments agreement and disagreement between the paired data sets.

aCGH _mb	Epic-450/850K data segment-mb			Total	%Agreement	Proportion of mismatch
	Balanced	Gain	Loss			
Balanced	1783.65	48.24	188.6	**2020.49**	88.23	Loss-balance (8.4 %)
Gain	30.8	177.9	1.94	**210.64**		Gain-balance (3.3%)
Loss	11.47	0.68	149.25	**161.4**		Gain-loss (0.1%)
Total	**1825.92**	**226.82**	**339.79**	**2392.53**		5/61- kappa (p >0.05)

3.1.6 Selected summarized copy number profiles generated by the two methods

Copy number variation profiles in a selected set of samples with perfect and slight agreement are shown in figure 3.2 to figure 3.5 below. I observed that 90% of the samples (55/61) showed similar copy number profiles patterns across the genome. For the remaining 10% of the samples (5/61); we could not establish either a correct pair or baseline correction for our intensity value.
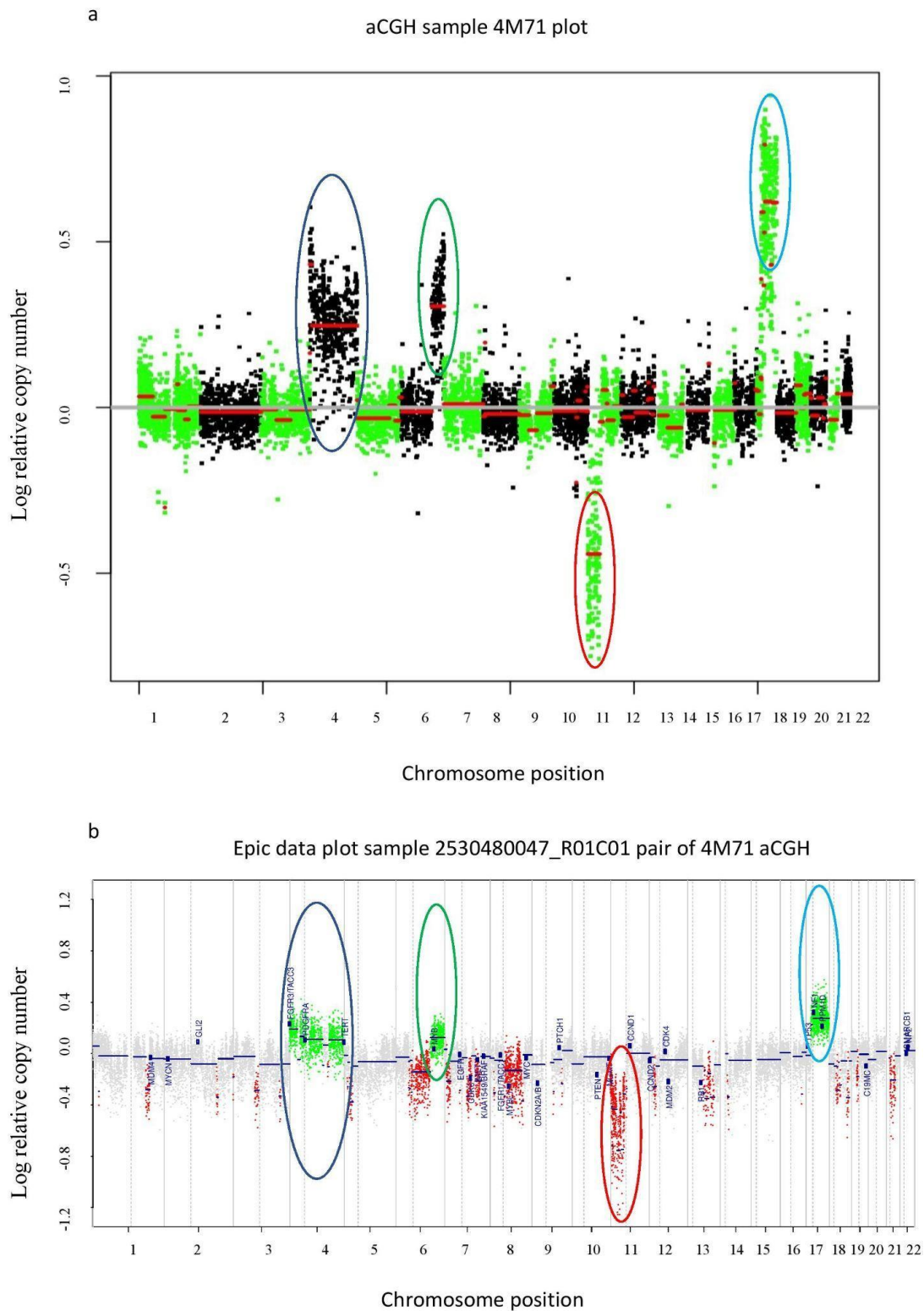
Figure 3.2 Sample chip ID:201530480047_R01C01.ID:X4M71; Summary of aCGH data CNV profile.

The green and black bins in (a) show alteration of chromosome number starting from chromosome 1-22 with corresponding chromosome position on x-axis b. Epic array methylation data CNV profile plot pair of the aCGH plot in (a). The two profiles indicate the correct sample pair match with moderate agreement.
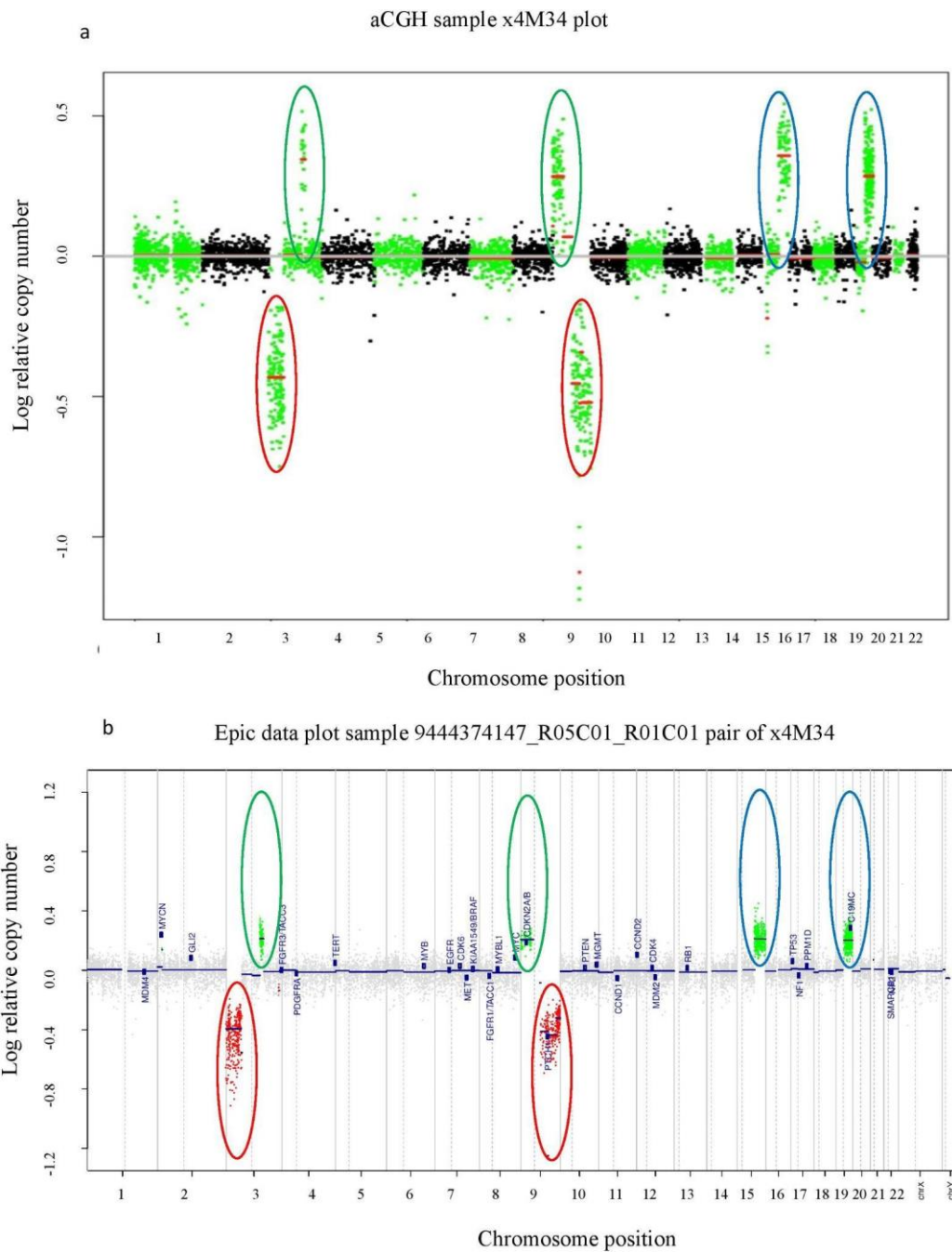
Figure 3.3 Sample chip ID:9444374147_R05C01.ID:X4M34; Summary of a. CGH data CNV profile.

The green and black colours on the plot shows alteration of chromosome number starting from chromosome 1-22 with corresponding genomic position on x-axis b. methylation data CNV profile that shows perfect pair match and sample perfect agreement. A plateau with baseline reference was achieved with most of the balanced segment aligned to the log2 (R/T) zero.

a
aCGH sample X4M36

b
Epic data plot sample 9610361029_R04C02 pair of X4M36 aCGH
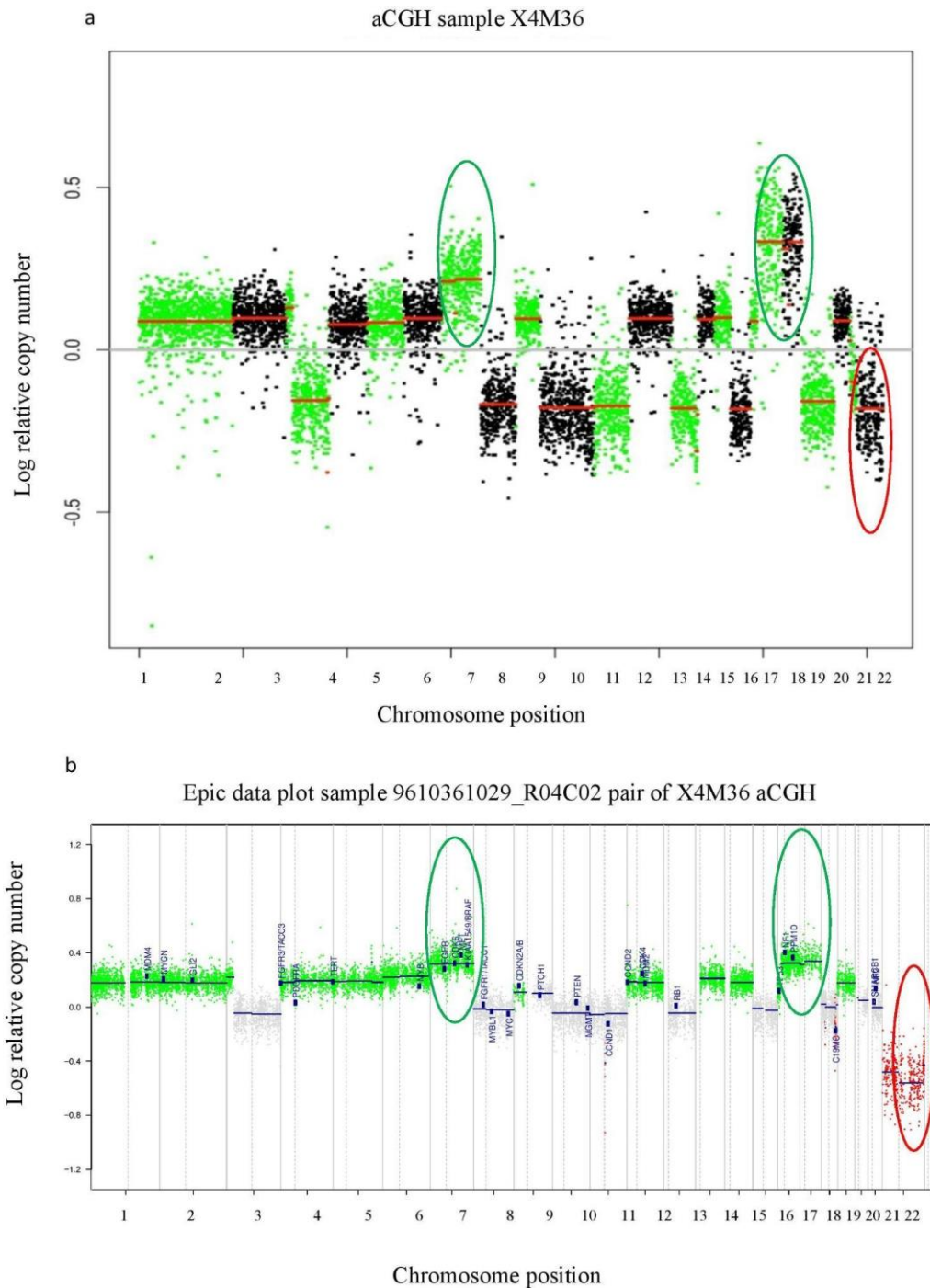
Figure 3.4 Sample chip ID:9610361029_R04C02.ID:X4M36, shows summary of a. CGH data CNV profile.

The green and black colours on the plot shows alteration of chromosome number starting from chromosome 1-22 with corresponding genomic position on x-axis b. Methylation data CNV profile that shows perfect pair match and agreement.In this figure segment plateau with baseline correction was not achieved. Balanced segment aligned to the log2 (R/T) below or above zero. In normal baseline correction, balanced segments align to 0.0.

a   aCGH sample X4M6 plot

b   Epic data plot sample 201533500059_R07C01 pair of x4M6 aCGH
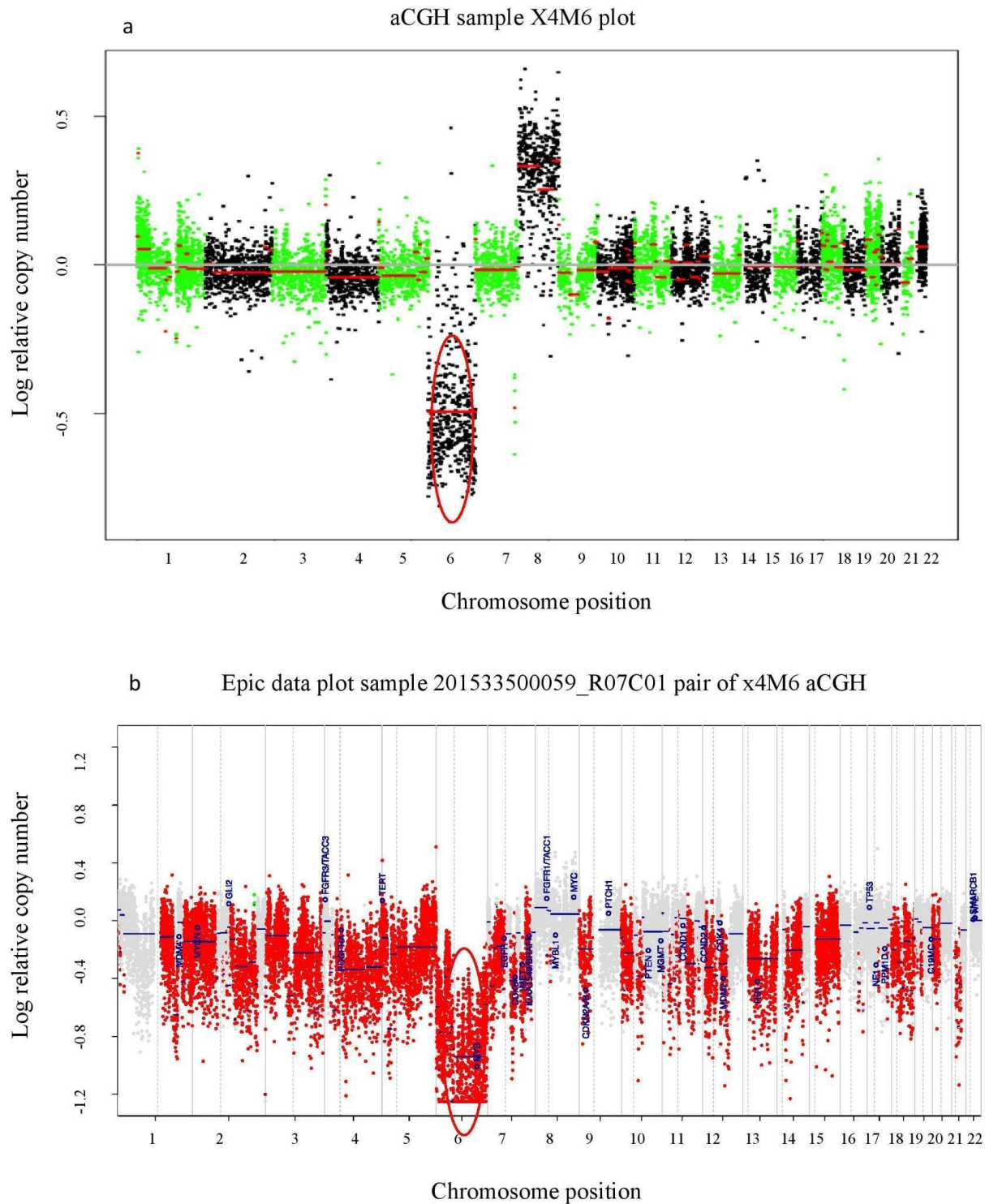
Figure 3.5. Sample chip ID:201533500059_R07C01.ID:X4M6; Summary of a. CGH data CNV profile.

The green and black colours on the plot shows alteration of chromosome number starting from chromosome 1-22 with corresponding genomic position on x-axis b. Methylation data CNV profile with failed baseline correction that shows perfect pair match and sample poor agreement with the aCGH data.

3.2 Copy number alterations profile and implications in astrocytoma patient stratification and diagnosis

3.2.1 Specific aims in phase II

In this phase I proposed to evaluate survival of WHO grades before and after clustering using the log2 copy number intensity value. I also aimed to identify relevant glioma pathways explained by the CNV in our astrocytoma data set and determine composition of methylation classes within our clusters. I further proposed to identify unique alteration patterns in each cluster.

3.2.2 Results on grading and survival analysis

To test whether the current world health organization grading system had an impact on survival, I conducted Kaplan Meier survival analysis in 116 samples between WHO astrocytoma grade II (n=46) and III (n=70). I found out that there was no significant difference in survival (p>0.05) between WHO astrocytoma grade II and grade III (Figure 3.6 a below). I observed an average survival of 1.7 years. The survival curves are shown in figure 3.6 a-c below. Although group III had considerable short survival time, it was not significantly different from grade II tumor's (p=0.84) as shown in figure 3.6 (3a). To identify subgroups from the WHO astrocytoma grade cases, I conducted hierarchical clustering (Pearson coefficient correlation ward D2 linkage) of the mixed WHO grade II and III using the log 2 intensity CNV. I was able to identify 7 clusters in the WHO grade II and grade III data sets as shown in the cluster dendrograms in figure 3.6 (3b). After identification of the clusters, I conducted survival analysis once again to find out whether the cluster had differences in survival time as shown in figure 3.6 (3c). I observed a significant difference in survival within the clusters. However only group 3 had a large sample size (n=56). In brief, cluster 1 (n=5) had median overall survival of 5 months while other clusters had median survival of 11, 10, 42, 8, 8 and 19 months for cluster 2-7 respectively. In the data set, I considered overall survival > 17 months as better survival which was only observed in cluster 4 and 7. Minimum survival of groups was 3, 1, 1, 4, 2, 1 and 1 months respectively while Maximum survival was 9, 87, 95, 108, 84, 68 and 80 months from cluster 1-7 respectively.
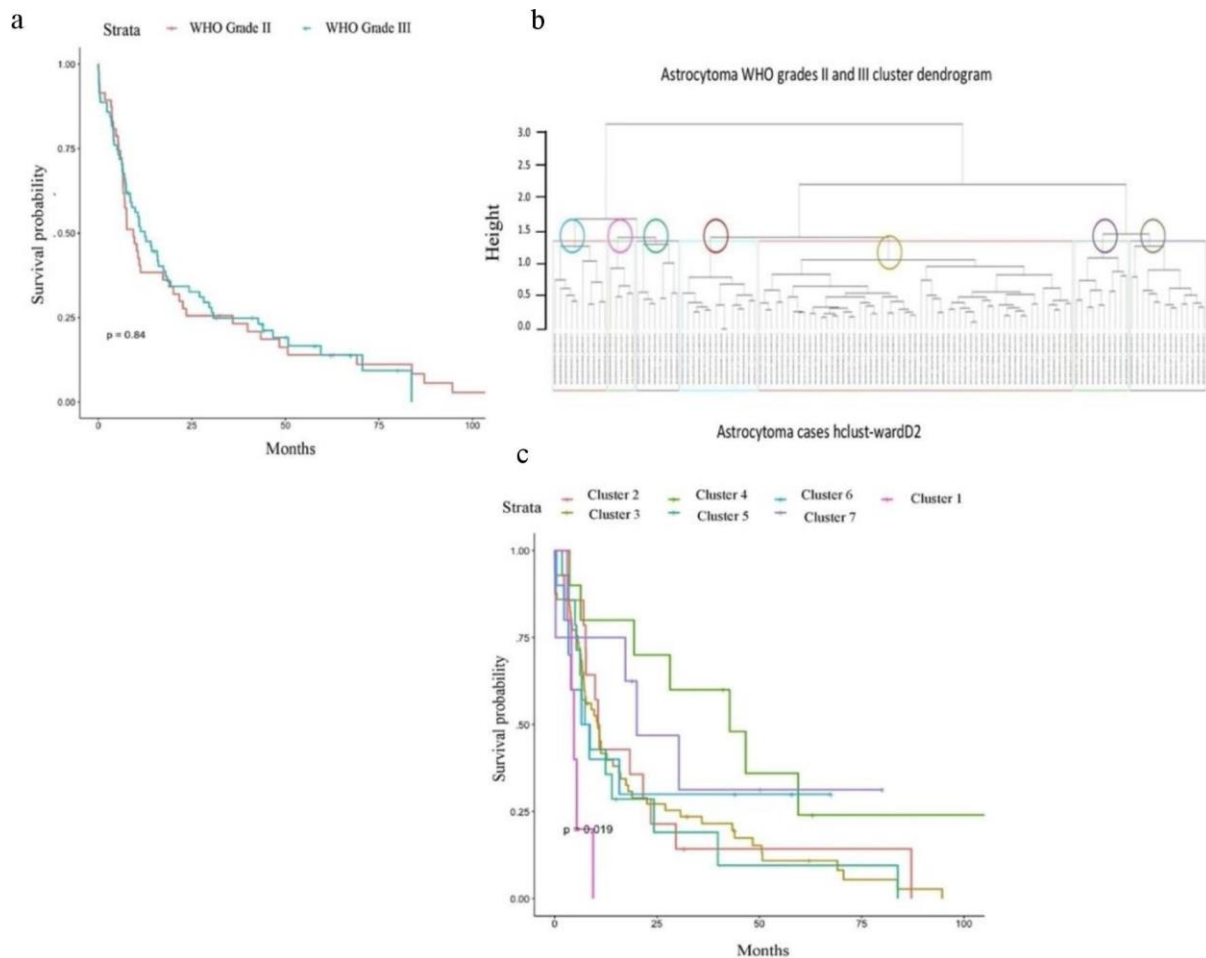
Figure 3.6. WHO survival curves before clustering, subgroups and groupwise survival curves after clustering.

3a. WHO grade II survival curve is shown by the brown line while WHO grade III survival curve is shown by the blue-line. There was no difference in survival between the two groups (p = 0.84). 3b; Shows the seven hierarchical clusters in 116 astrocytoma samples after clustering based on log2 intensity values. The circular node colour is in line with the survival analysis curve colours after hierarchical clustering in figure 3c; Shows the identified 7 clusters survival curves. There was significant survival difference between the clusters (P = 0.019). Cluster 1 (pink) had the least number of samples n=5 and poor survival. Cluster 7 (Violet) had few samples too but relatively better survival (n=8). Cluster 4 (Green) and 6 (Mint Green light) had 10 samples each while cluster 2 (brown) and 5 (blue) had 14 samples each. Cluster 3 (Olive-green) had the highest number of samples n=57 and most samples had poor survival too as compared to clusters 4 and 7.

I further conducted hierarchical clustering (Pearson coefficient correlation ward linkage) using the CNV value to determine clusters that may be used in grading). I was able to identify 7 clusters in the WHO grade II and grade III data set (Figure 3.6. 3b above). To identify if the

subgroups identified have differences in survival, I conducted survival analysis. I observed a significant difference in survival between the data sets (P < 0.05).

After clustering I conducted survival analysis within the clusters and observed a significant difference in survival between the clusters (Figure 3c above). There was a significant survival difference between the 7 subgroups (P = 0.019). Clusters 1 (pink) had the least number of samples n=5 and poor survival. Cluster 7 (purple) had few samples too but relatively better survival (n=8). Cluster 4 (Mint Green light) and 6 (Violet) had 10 samples each while Cluster 2 (Olive-green) and 5 (blue) had 14 samples each. Cluster 3(Green) had the highest number of samples n=57 and most samples had poor survival too as compared to cluster 4 and 7.

3.2.3 Identification of CNV frequencies in each astrocytoma sub-groups

To identify CNV patterns, I first determined the alteration profile in each group. I further re-grouped them based on the survival curves as some presented with better survival and some presented with poor survival figure (3.1 a and figure 3.1b)

Chromosome position

Chromosome position

Figure 3.7 Four clusters that had poor survival outcomes. Red bar indicates chromosomal gain i.e log 2 intensity > 0.1 while blue bar indicates chromosomal loss log2 value <-0.1.

Dotted line indicates the centromere i.e divide p and q arm of each chromosome respectively. The common deletion with >50% occurrence in at least 2 of the 4 groups were observed at Chr3q Chr4q, Chr5p/q, Chr7q and Chr13q. C2 was characterized by chr 6q deletion (98%) and Chr8p/q (50%) while C5 was characterized by 11p deletion (98%). Cluster3 was characterized by deletion at 6q (50%), Chr8p/q (25%) and Chr11p(25%). Though the frequency was low, the number of cases were considerably high to consider these CNVs in C3.

Chromosome position

Chromosome position

44

Figure 3. 8.Three clusters (C4, C6 and C7) that had poor and better survival mixed outcome among its subjects

The common CNV with >50% frequency in groups C6 and C7 included Chr4p and Chr13q co-deletions. C4 had relatively better survival though it had most alterations. It was characterised by Chr3q (60%), Chr4q (76%), Chr5p/q (50%), Chr11q (72%), Chr12q (72%) and Chr13q (60%) co-deletions. In-addition there was chr12p (77%) gain. C6 had Chr4q (55%) and Chr13q (60%) co-deletions and Chr8q gain (50%). G7 had Chr6q, Chr13q, Chr14q and Chr19q co-deletions.

3.2.4 Heatmap analysis of a selected poor and better survival subgroups

To identify the nature of the intensity values among the poor and better survival groups, we conducted heat map analysis. It was observed that the groups with poor survival had much alterations, mostly loss of chromosomal segments which appeared in higher frequencies than the groups with better survival (Figure 4.2.3). I observed severe deletion of chromosome 4 and 13 in C3 which had poor survival. In C7 which had better survival this region had 40-60% frequencies in deletions.

Limits (-0.1,0.1)

C3 cases with poor survival

Chromosome position

Limits (-0.1,0.1)

C4 cases with better survival

Chromosome position
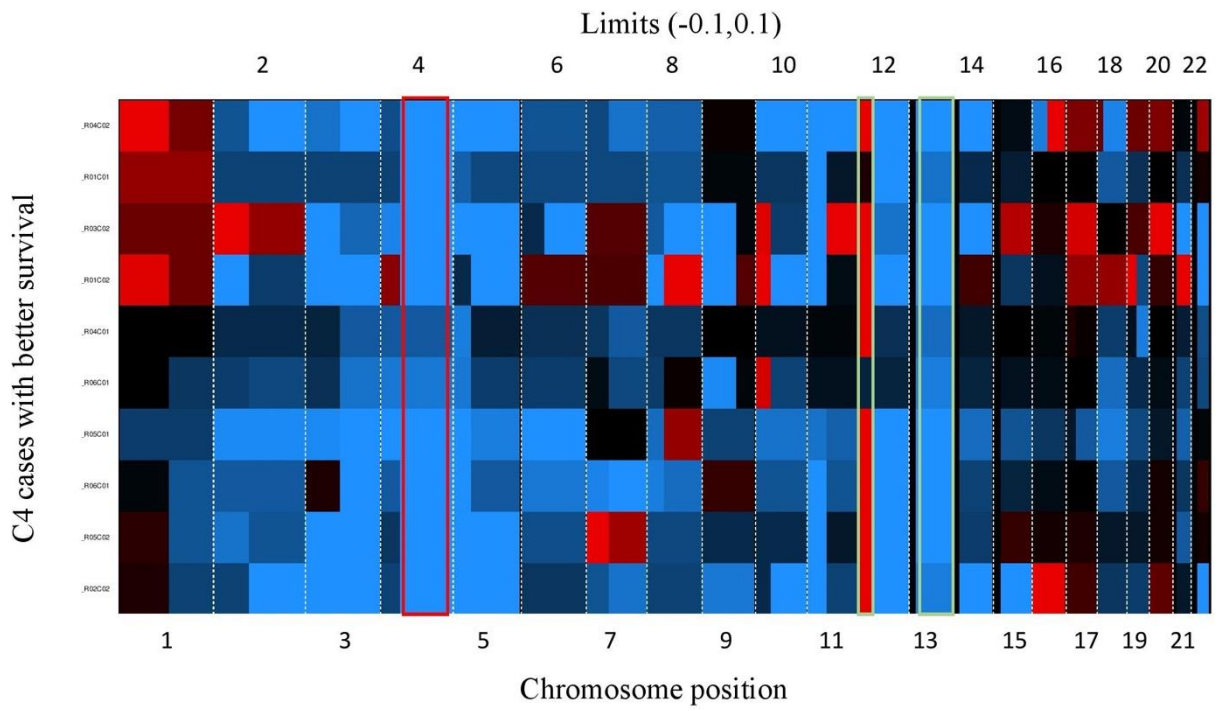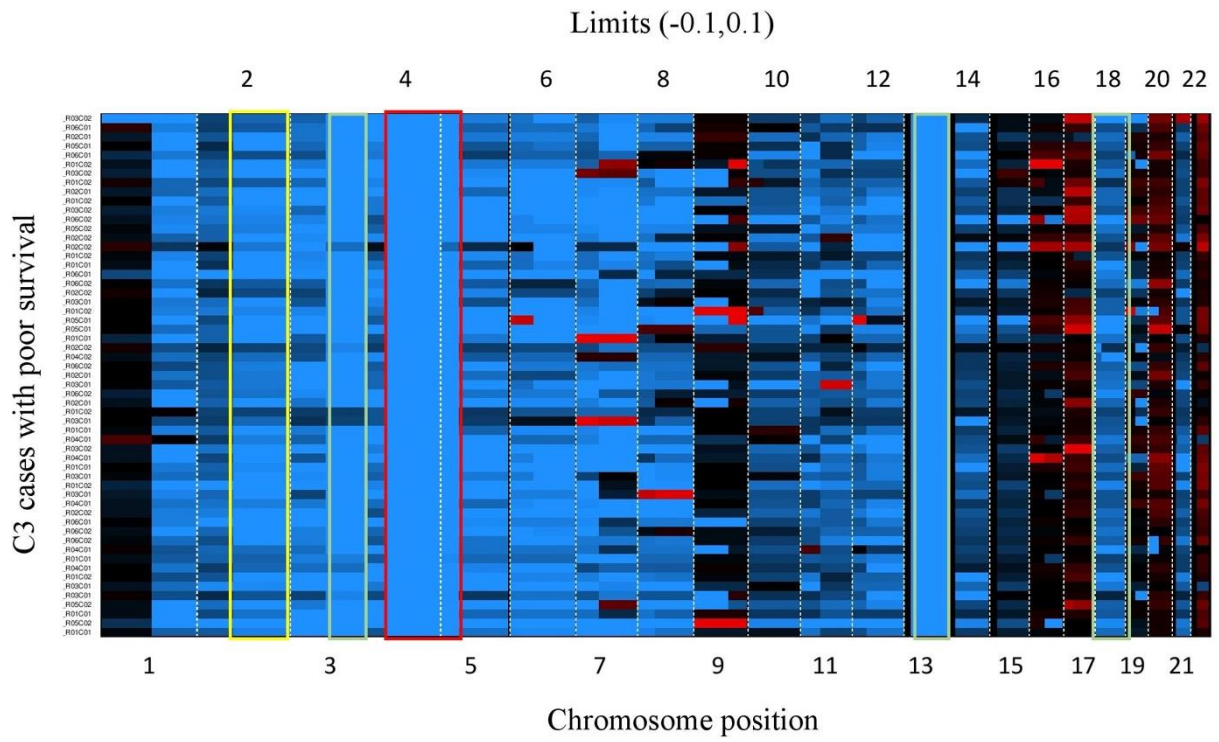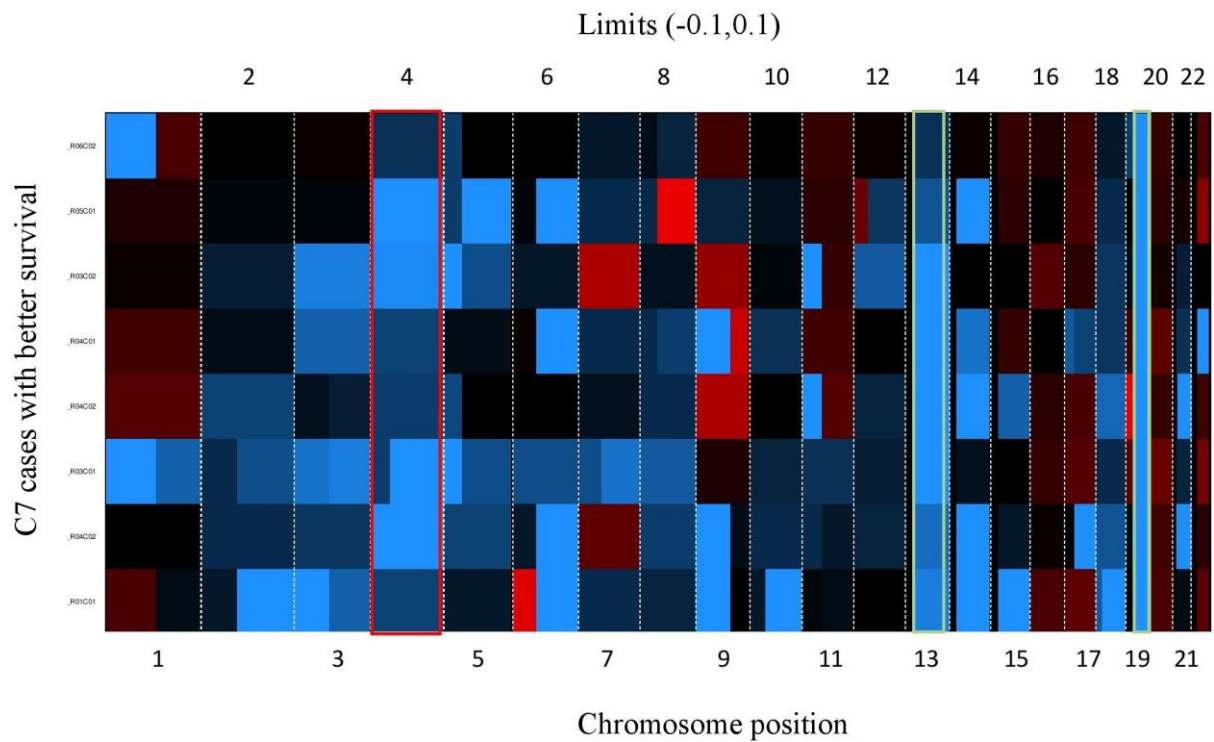
Figure 3.9. Heat maps showing the distribution of log2 intensity values in cluster 3 (C3), C4 and C7.

C3 had poor patient survival outcome whileC4 and C7 had better patient survival outcome. Dark nuisance means balance, blue (deletion) and red (gain) in all the autosomes. There were reduced alterations in C4 and C7 as compared to C3. General C3 had unique deletions at Chr 1q,2q,6q,14p,18p/q,21p/q and 14p While gain was at chr 17p/q, 20p and 22 q. Common deletion in C4 was on Chr3p/q, 4q, 5p/q, 12q,13p/q and 14q. In-addition Unique gain in group 4 was at Chr12p.C7 had chromosome deletions at 4p/q,13p/q and 19q (100% cases). C7 had 2 samples with Chr1p/19q co-deletions which could be oligodendroglioma tumor cases.
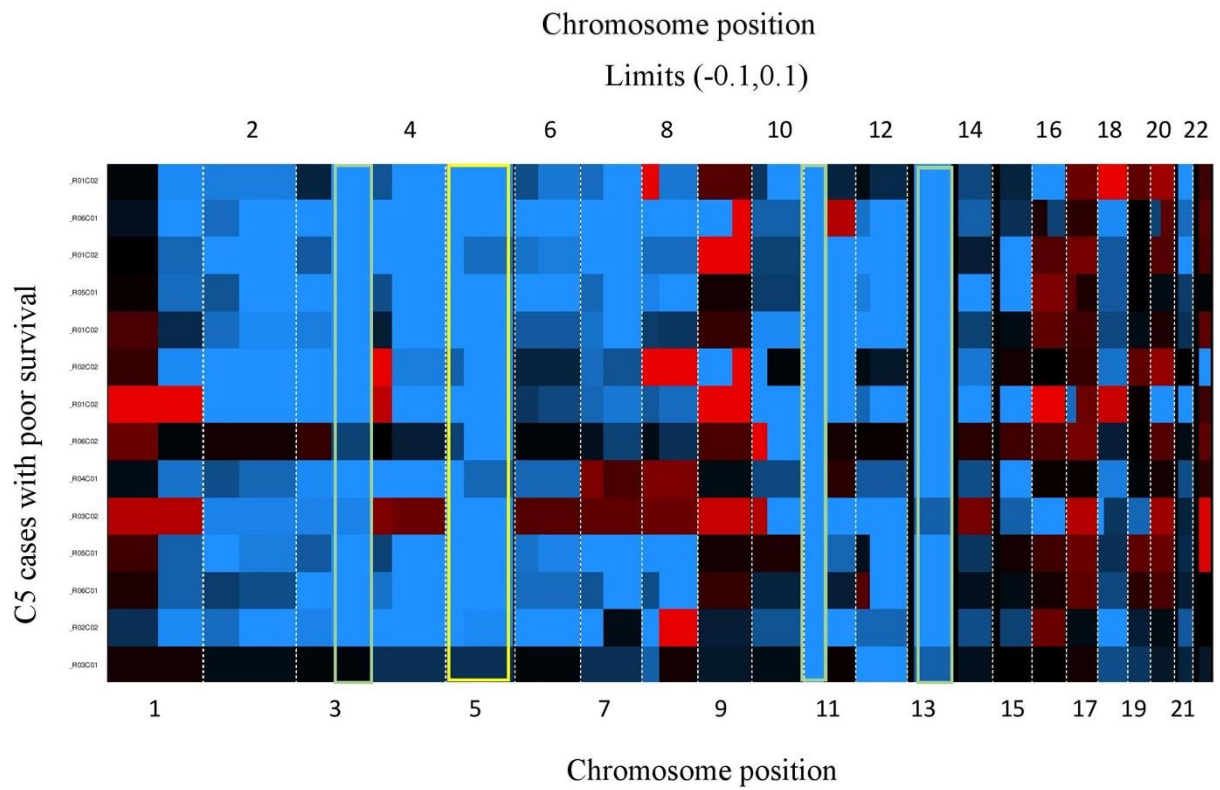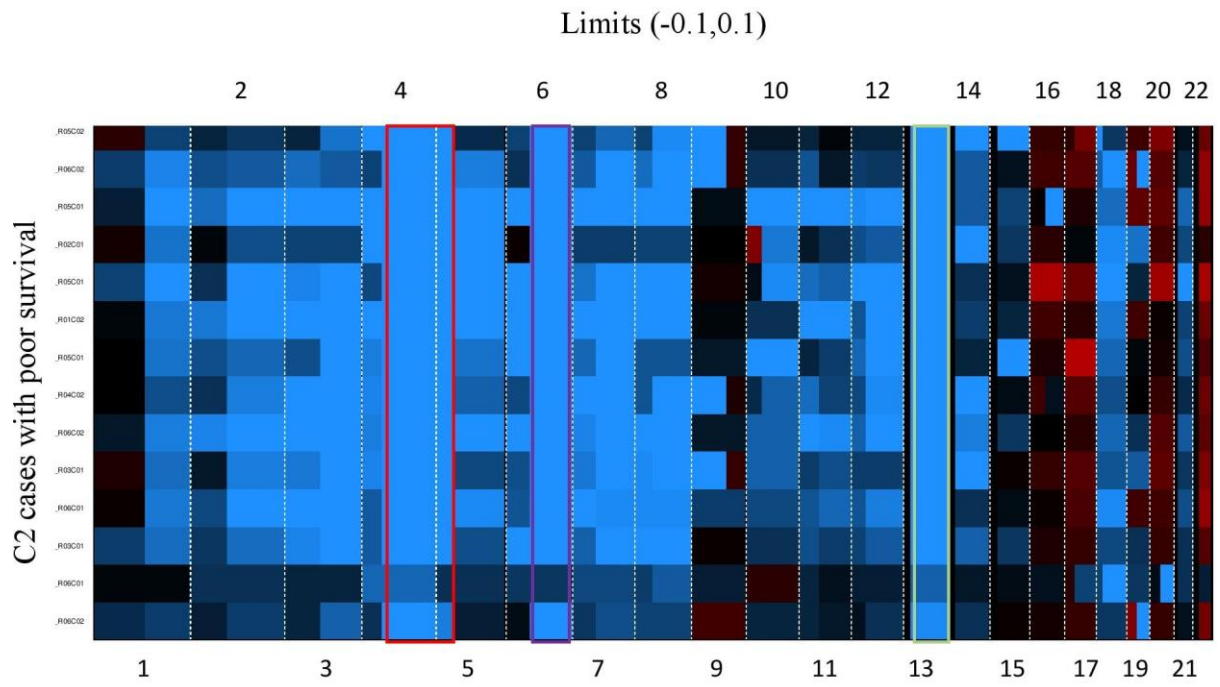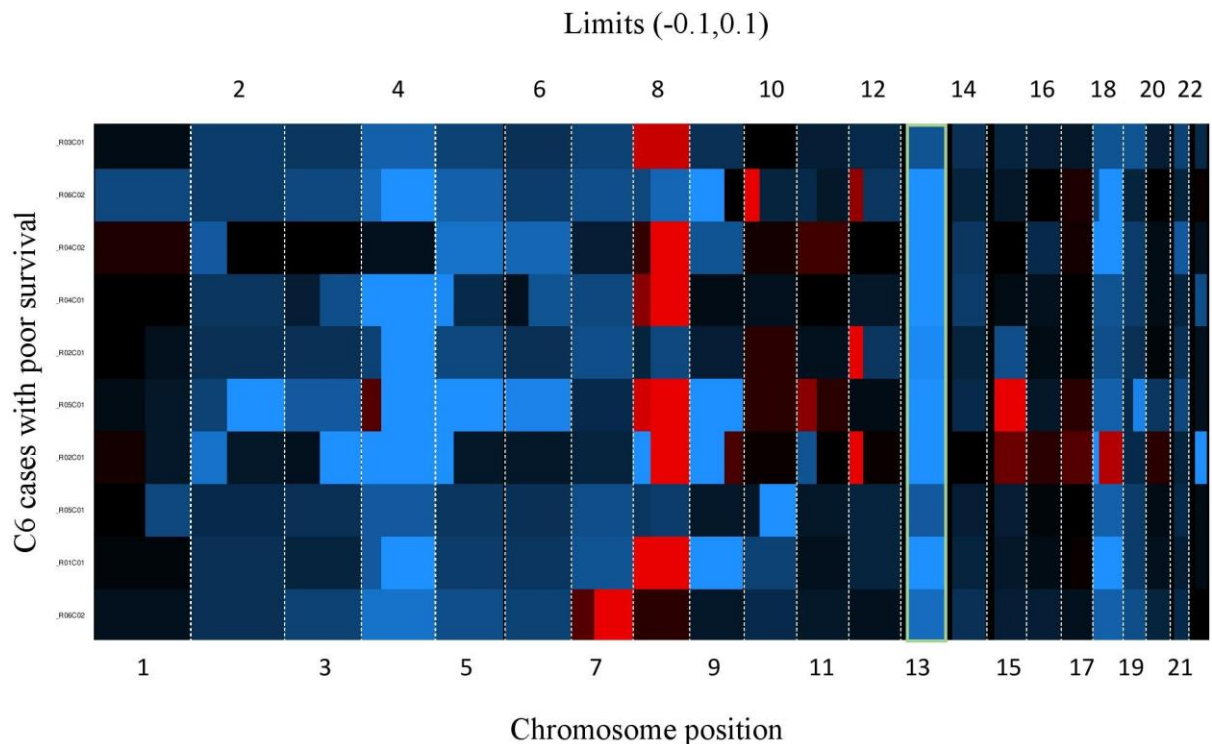
Figure 3.10. Heat maps showing the distribution of log2 intensity values in cluster 2 (C2), C5 and C6.

They all had poor patient survival outcomes i.e Overall median survival of 11 months, 8 months and 8 months respectively. Dark nuisance means balance, blue (deletion) and red (gain) in all the autosomes. There were reduced alterations in C6, C2 and C5 in reducing order respectively. Chromosome 13p/q was a common deletion. C5 which had the most poor outcome had unique deletion at Chr11p (100%) While C2 had Chr4p/q,6q and 18p/q deletions. In-addition Unique gain in C6 was at Chr8p (60%).

3.3 Methylation classes identified based on the classifier

To confirm the methylation classes in each cluster, we used the classifier. The principal output of the classifier was a list of predicted methylation classes per each group member as shown in figure 3.10 below. For each class, I considered the threshold probability calibrated score of $\geq$ 0.9 of the 91 classes. This implies that the remaining class probability will add to < 0.1 in the classification scheme. However, in some cases a cut-off value of ≥ 0.5 performed well. Details of the scores organised per cluster are shown in supplementary table 1 attached. After uploading unprocessed idat-files of Illumina Human Methylation 450 BeadChip arrays to the classifier, whereby in principle the classifier compares the data set of 12 in each load to a reference cohort of 2800 neuropathological tumours that were used to develop this classifier. I was able to get the methylation profiling of data along with checking Isocitrate dehydrogenase

(*IDH*) mutation status, and the 06- methylguanine DNA methyltransferase (*MGMT*) methylation promoter status as shown in selected set of samples in cluster 3 (figure 3.10 below).
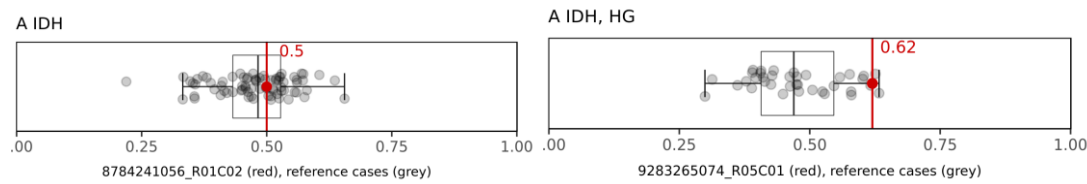
Figure 3.11 A_IDH (left) and A_IDH_HG (right) methylation status in group 3 samples.

Red dot shows sample methylation status against grey dots of reference methylated samples in a box plot with error bars for the reference. A total of 116 TCGA samples were successfully reclassified and methylation profiles along with CNV profiles and WHO grades are available.

3.3.1 Similar age distributions between clusters identified

Since age is one of the common confounders in cancer, especially to incidence, survival and severity of cancer, I further checked the age distribution between the groups before confirming the methylation classification of astrocytoma. There was no statistical difference in age distributions between the 7 clusters (Kruskali-wallis test P- value > 0.05). The average age between the groups ranged from 35-40 yrs.

3.4 Distribution of WHO grades in each cluster and methylation classes

I further observed differences in frequencies of both WHO grades in each cluster as shown in figure 3.11. All samples belonged to the family glioma *IDH m*utant. Majority of the samples belonged to subclass astrocytoma_ IDH mutant (A_IDH) with some classified as A_IDH_High grade glioma (A_IDH_HG) and also few were 1q/19p co-deletion oligodendroglioma based on methylation classifier. The A_IDH_HG had differential frequencies besides occurring as in mixed classification with astrocytoma within the clusters. The CNV profiles were also different but similar based on how close the clusters were in the survival curves. In summary, cluster 1 (C1) was dominated by astrocytoma sub-class (100%), in cluster 2 (C2), two samples (15.4%) had mixed classification with astrocytoma having higher methylation score of 0.62 and 0.72 against subclass high grade astrocytoma at 0.36 and 0.27 respectively. Astrocytoma subclass with >= 0.94 calibration score consisted of 74.6% of C2. Details of each individual sample's calibrated score are provided in supplementary table 1. Cluster 3 (C3) had the highest number of subject data (n=56). 8.9 % (n=5) were mixed with a higher astrocytoma score than high grade astrocytoma. 7.1% (n=4) were also mixed classification with high astrocytoma HG

calibration score while an additional sample was classified as A_IDH_HG with calibration score of 0.98. The rest 82.1% were classified as A_IDH. Cluster 4 (C4) had 40% (n= 4) high grade astrocytoma. Though 2 samples were mixed, they had A_IDH_HG higher calibration score than the astrocytoma subclass. We observed that the 60% subclass astrocytoma had a slightly low calibration score of >= 0.91. Cluster 5 (C5) was dominated by subclass A_IDH_HG with 64.3% (n=9) having been classified as A_IDH_HG or mixed. Two samples of the 9 had low A_IDH_HG subclass compared to astrocytoma class. The remaining 33.7% belonged to class astrocytoma. Cluster6 (C6) had 20% A_IDH_HG and 20% A_IDH_HG with bias to higher astrocytoma classes based on the score. The remaining 60% were all subclass astrocytoma majorly classified as WHO grade III. Cluster7 (C7) was the only class dominated by subclass 1q/19p codeleted oligodendroglioma at 50% (n=4). One sample belonged to A_IDH_HG subclass and 37.5% (n=3) were subclass astrocytoma. Two of which were WHO grade III astrocytoma and one grade II astrocytoma.
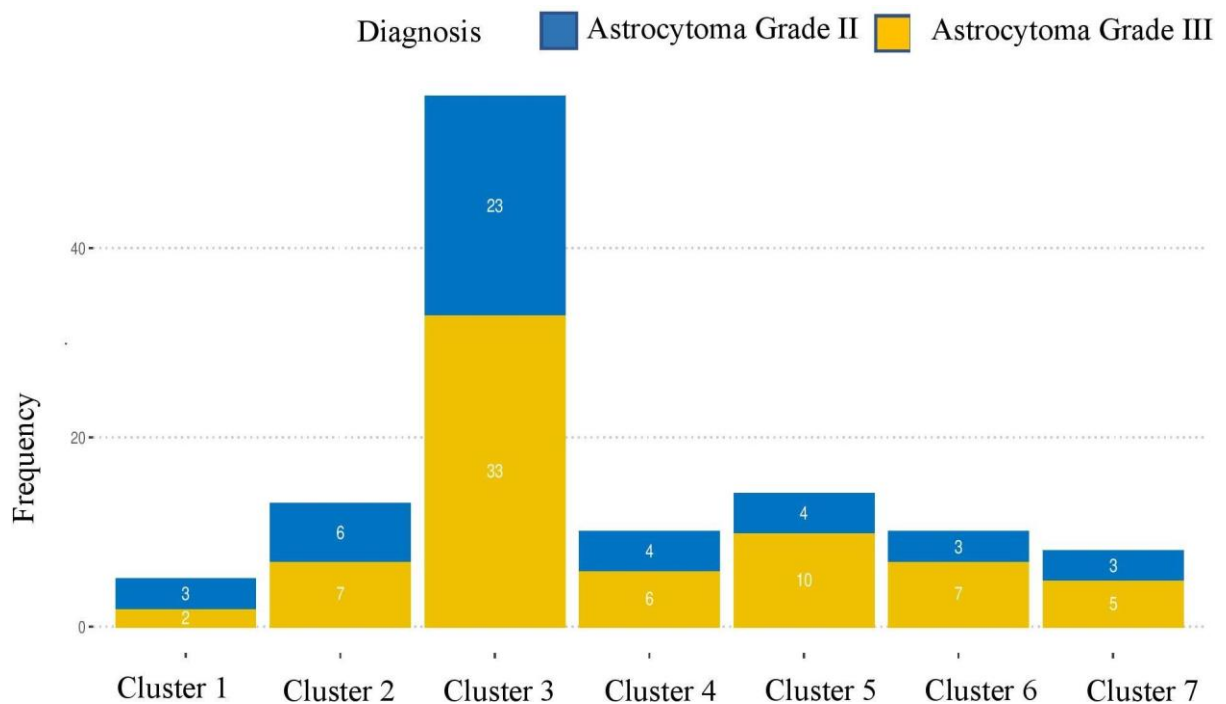
Figure 3.12 Distribution of the WHO grades after clustering the data based on copy number alteration log-intensity values.

Blue bars indicate grade II and orange bars indicate grade III in each cluster. Cluster 1 was dominated by WHO grade II (60%) with all samples being A_IDH subclass. Cluster 2 was dominated by WHO grade III (64%) with 2 of the 13 cases characterised as high-grade astrocytoma though at a methylation classifier score of 0.36 and 0.27. Cluster 3 was dominated

by grade III (59%) with 7 samples (13%) classified as being A_IDH_ HG. Cluster 4 was dominated by WHO grade III (60%). Cluster 5 was also dominated by WHO grade III (71%). Cluster 6 samples were A_IDH with 7 of 10 samples among this cluster being grade III (70%). Grade 7 was also dominated by WHO grade III (63%).

As compared to the slope there was no clear observable pattern between the survival slopes and the distribution of the WHO grade classes for example cluster 1 had a poor survival and I expected WHO grade III to dominate. However, the distribution of the methylation classes i.e. A_IDH, A_IDH-HG and codeletions at 1p/19q and oligodendroglioma classes were differential enriched in each group for example cluster 7 had 50% oligodendroglioma. It is therefore promising to have improved classification with a combination of methylome and copy number alterations.

3.5 Distribution of CNV between and within clusters in the class A_IDH_HG

There was a difference in CNV alterations among the A_IDH_HG within clusters (figure 3.12). we noticed common similar alterations and common deletion at cluster level. Cluster 5 had A_IDH_HG characterised by presence of Chr13p/q deletion (figure 3.12).
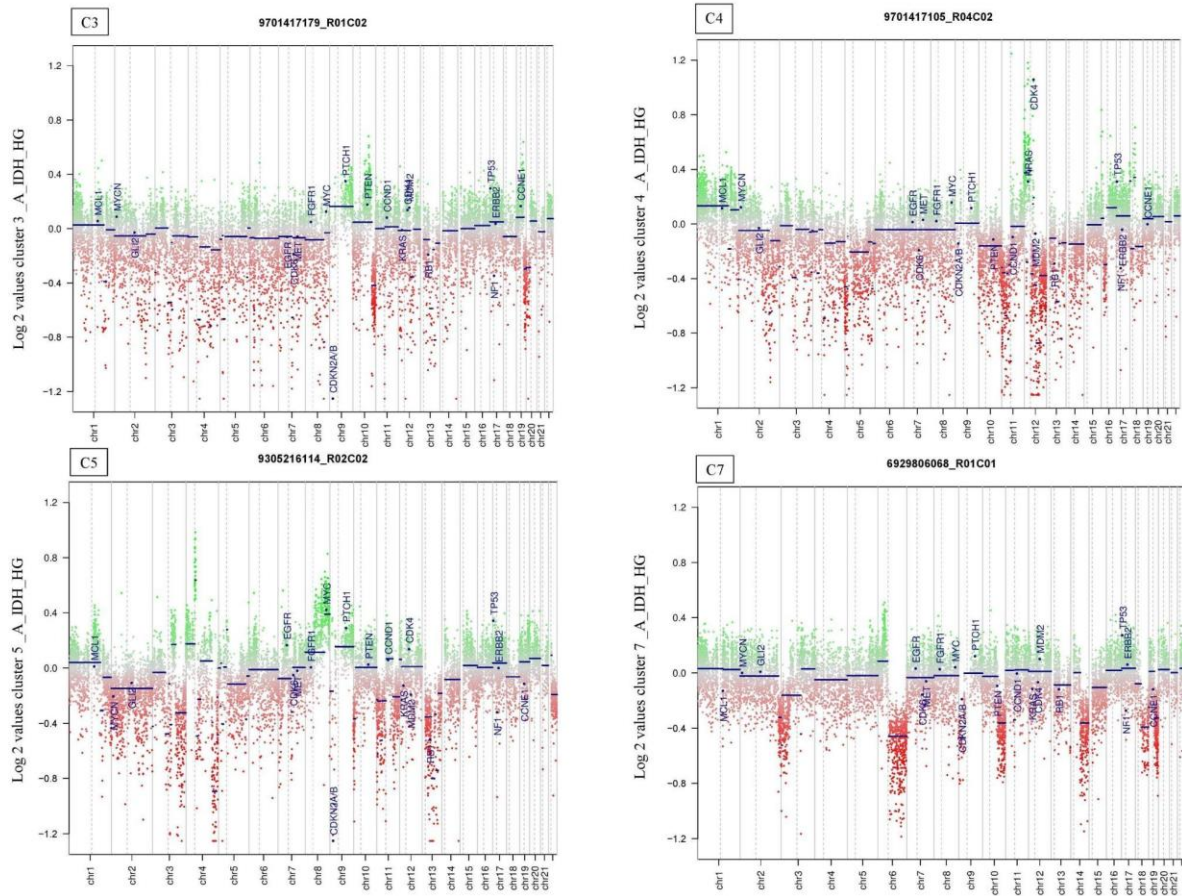
Figure 3.13 Difference in copy number alterations in selected clusters with blue text showing 29 genes commonly associated with brain tumour.

From the left top, cluster 3 had a clear gain in chr9 and a loss in chr19q while cluster 4 had gain in chr1 &11 with a loss in chr 12. Cluster 5 had a gain at chr4p with multiple consistent loss in chr3q chr4q and chr13. Cluster 7 had multiple loss at chr6q, chr10q, chr14q, chr18 and 19q
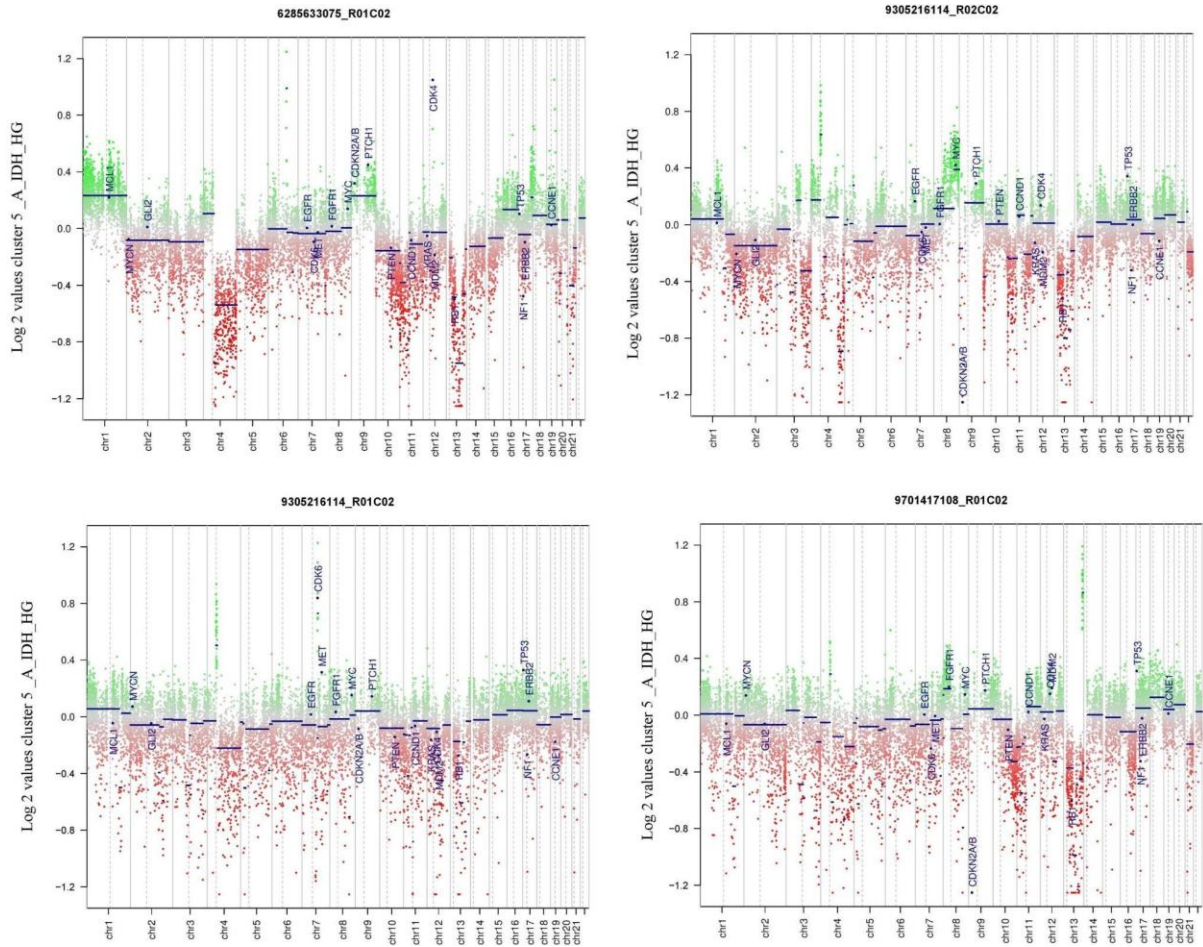
Figure 3.14 Difference in copy number alteration within cluster 5 with blue text showing 29 genes commonly associated with brain tumour

A unique common deletion at Chr13q/p was observed in the four-group member classified as A_IDH_HG.

3.6 Identification of CNV status in the 29-brain cancer associated genes per cluster

The groups with poor survival outcome which was determined earlier formed the clusters C1, C2, C3 and C5. They were characterised by co-deletion on *CDKND2A/B, MYB1* and *RB1* (figure 3.14; C1, C2, C3 and C4). The groups with better survival consisted of C4, C6 and C7. All had low frequencies of co-deletion on *CDKND2A/B* and *RB1* as compared to the group with poor survival. In addition, C4 and C6 had chromosomal gain in gene *CCND2*. However, C7 had a unique *C19MC* deletion beside having around 60% co-deletion MYB1 and RB1 gene .Supplementary figure 1 shows all the 29 genes selected per cluster.

C1. Log intensity values against 29 common brain cancer genes
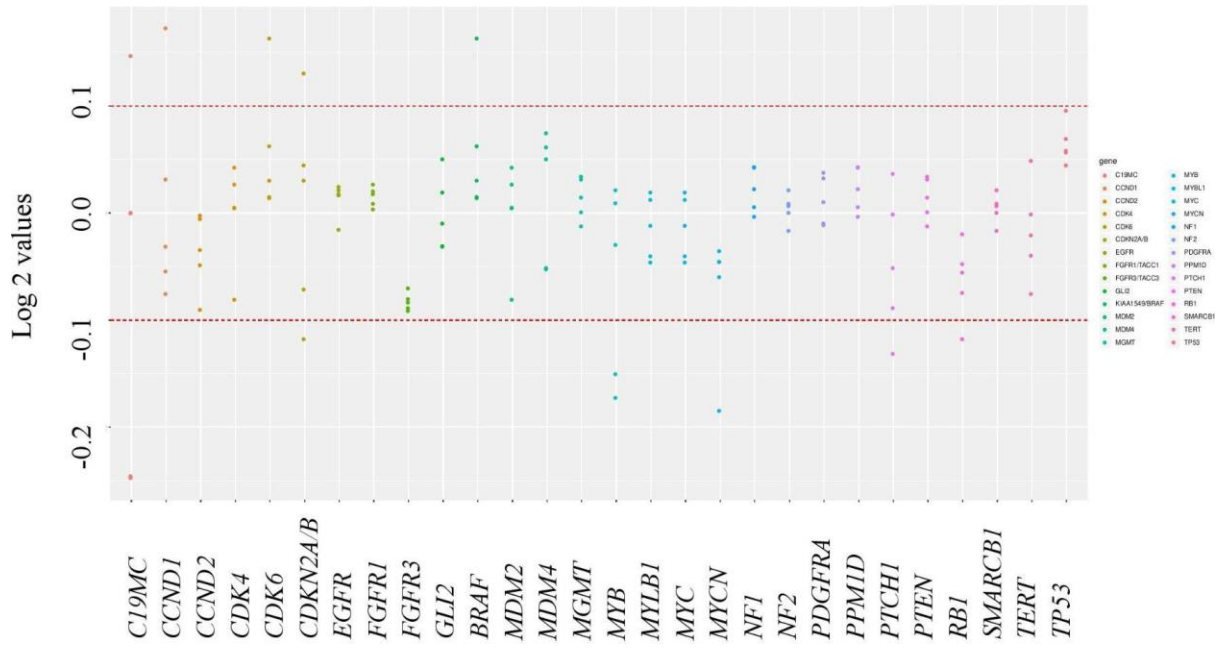
C2. Log intensity values against 29 common brain cancer genes
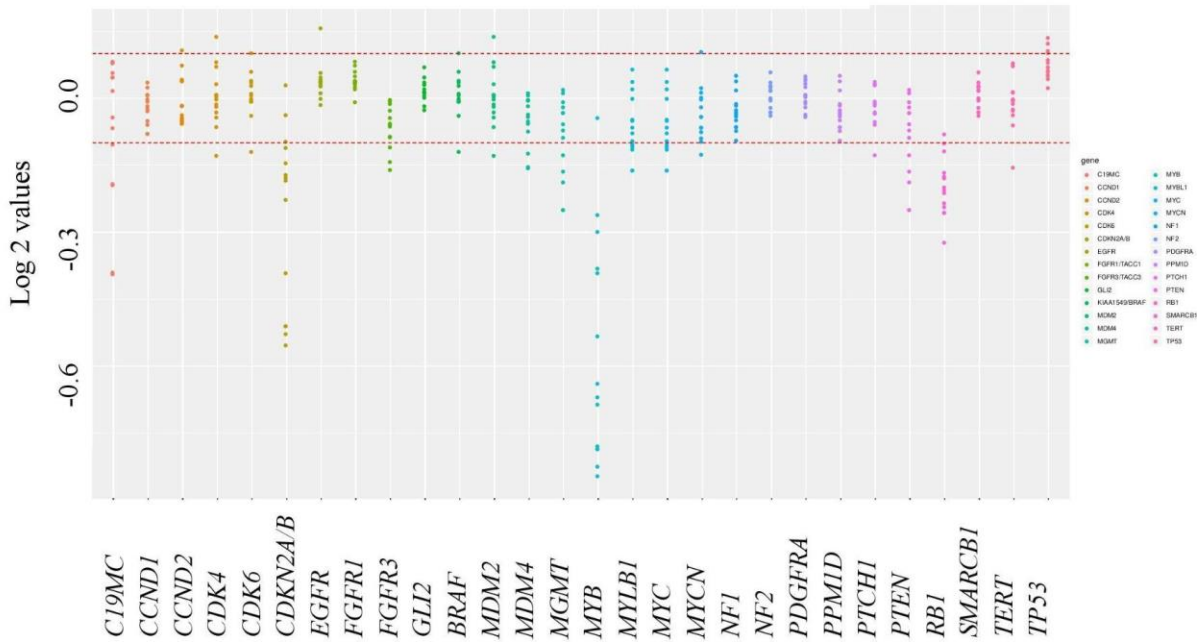
C3. Log intensity values against 29 common brain cancer genes
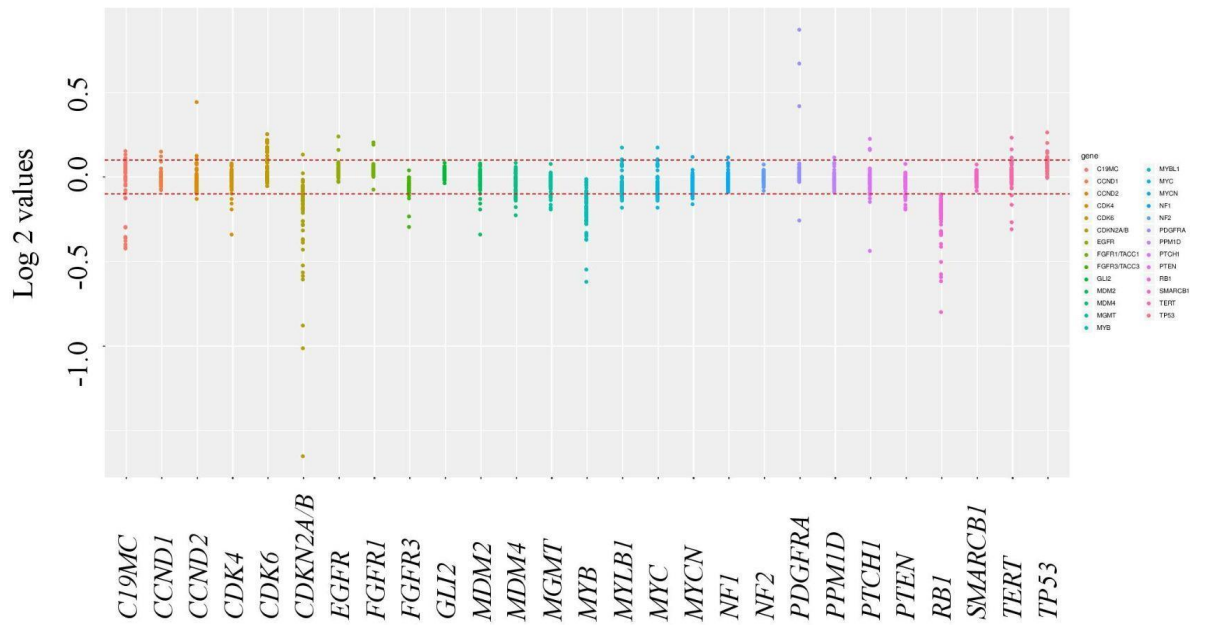
C4. Log intensity values against 29 common brain cancer genes
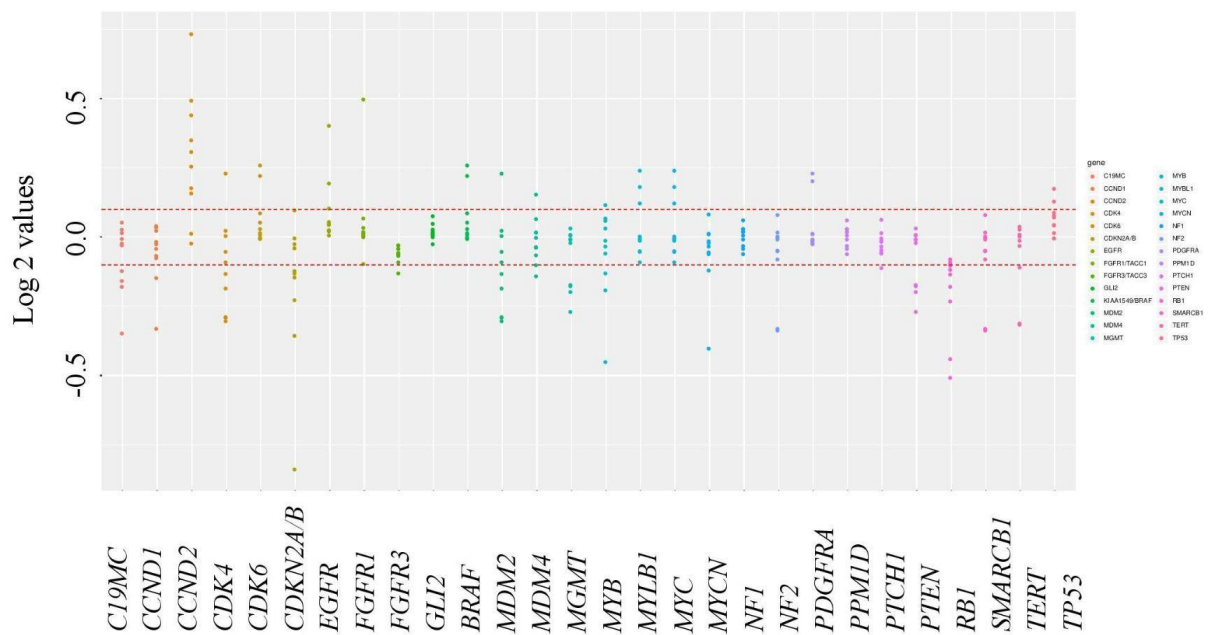
Figure 3.15 Distribution of 29 commonly brain tumour genes from for the first four clusters with their respective alteration.

The red dotted lines show the threshold for gain (0.1) the top red dotted line and loss threshold (-0.1) below 0 red dotted line. Patients in C1, C2, C3 and C5 had poorer survival outcomes than C4, C6 and C7. Consistent co-deletion were found in *CDKND2A/B, MYB1* and *RB1* genes in C1, C2, C3 and C5. This also occurred at C4 and C6 but at low frequencies beside the common gain on *CCND2* in these groups. C7 had a unique deletion on *C19MC, MYB1* and *RB1*.

3.6.1 Glioblastoma multiforme signalling pathways associated with the altered regions

Several pathways were associated with the altered regions. We therefore considered the most relevant pathways associated with glioma carcinogenesis based on literature, ratio of > 0.01 of enriched molecules in the pathway and overlap of pathway in at least 3 of the seven clustered altered regions. Moreover, we considered small P values of molecules in order to ensure that the molecules are not randomly enriched by chance. Using this approach, we identified a total of 35 pathways which were associated with angiogenesis, cell cycle, apoptosis regulation and growth factor receptor signaling for example Cell Cycle: G1/S Checkpoint Regulation, glioblastoma multiforme Signaling, ERK/MAPK signaling and VEGF signaling. The molecules are shown in figure 3.15. The pathways with high ratio of altered molecules in canonical pathways selected include; Retinoate Biosynthesis II (0.75), Spermidine Biosynthesis 1 (0.5), IL-3 signaling (0.3), Glioma Invasiveness signaling (0.3), p53 signaling (0.2) and axonal guidance signaling (0.2). Supplementary table 2 shows the enriched pathways and their associated molecules. In-addition, figure4 below shows the molecules enriched in glioblastoma multiforme. It was enriched in group 4.
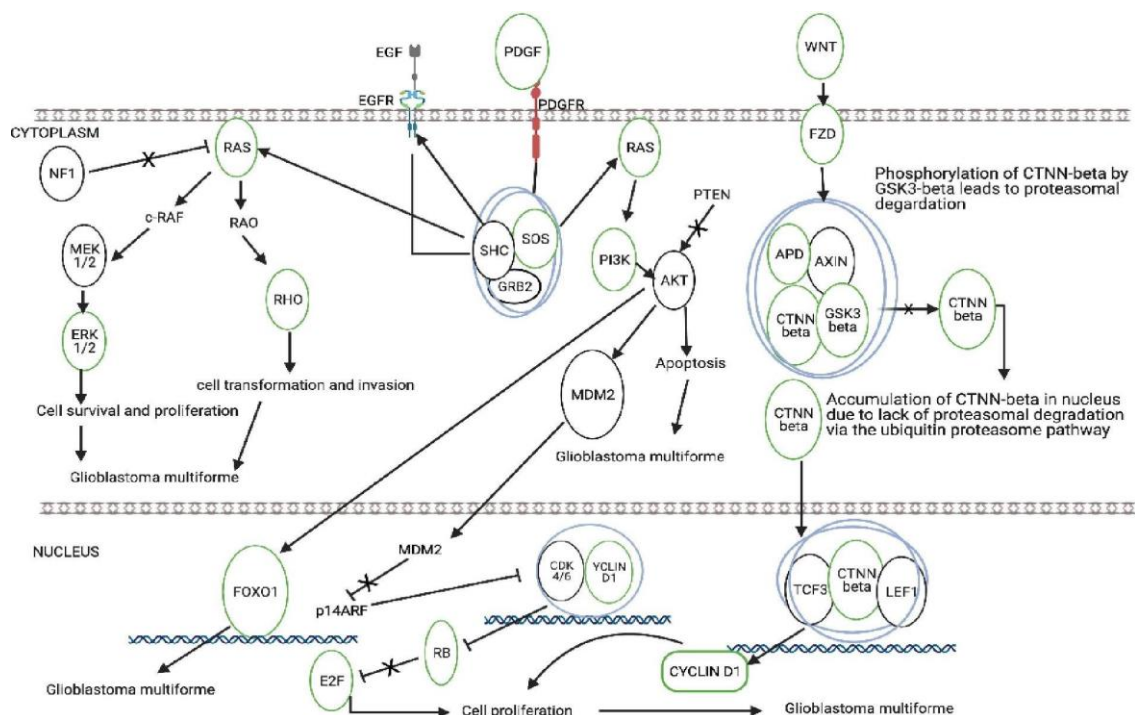
Figure 3.16 Molecules enriched in glioma multiforme signalling pathways in the majority of clusters.

The green circled genes were found in our data set. A complete list of altered genes and intensity values are attached as supplementary table 2.

3.7.1 Specific aims in phase III

In this phase, I proposed to Identify the genomic landscape of the brain tumour subgroups including copy number alterations and point mutations (missense, nonsense, frameshift deletion and frameshift insertions). I further aimed to Identify unique copy number patterns in primary, recurrent or metastasis tumours as compared to normal blood controls (BDN) in glioblastoma and lower grade glioma in large cohorts. Below are the results of the analysis.

3.7.2 Genomic landscape in glioblastoma and low-grade glioma

After testing my workflow, I analysed the genomic landscape in glioblastoma (n=597), as shown in figure 3.16 below. Based on our analysis, glioblastoma can be characterized by missense mutation in 98% of the chromosomes with chromosome 3, 7, 9 and 12 having only missense point mutations while Chromosome 8 was mainly dominated by nonsense mutations. Nonsense mutation with frameshift deletion and missense mutations was observed in chromosome 1, 10, 13, 15 and 17. Nonsense mutations were observed in Chromosomes 1, 2, 4, 6, 7, 8, 10, 11, 13, 14, 15, 16, 17, 19 and 20. Frameshift deletions were observed in chromosome 1, 10, 13, 15 and 17. Frameshift insertions were only present in chromosomes 3, 6 and 10. Unlike the low-grade glioma subtype, the glioblastoma landscape was observed to have gene amplification on chromosome
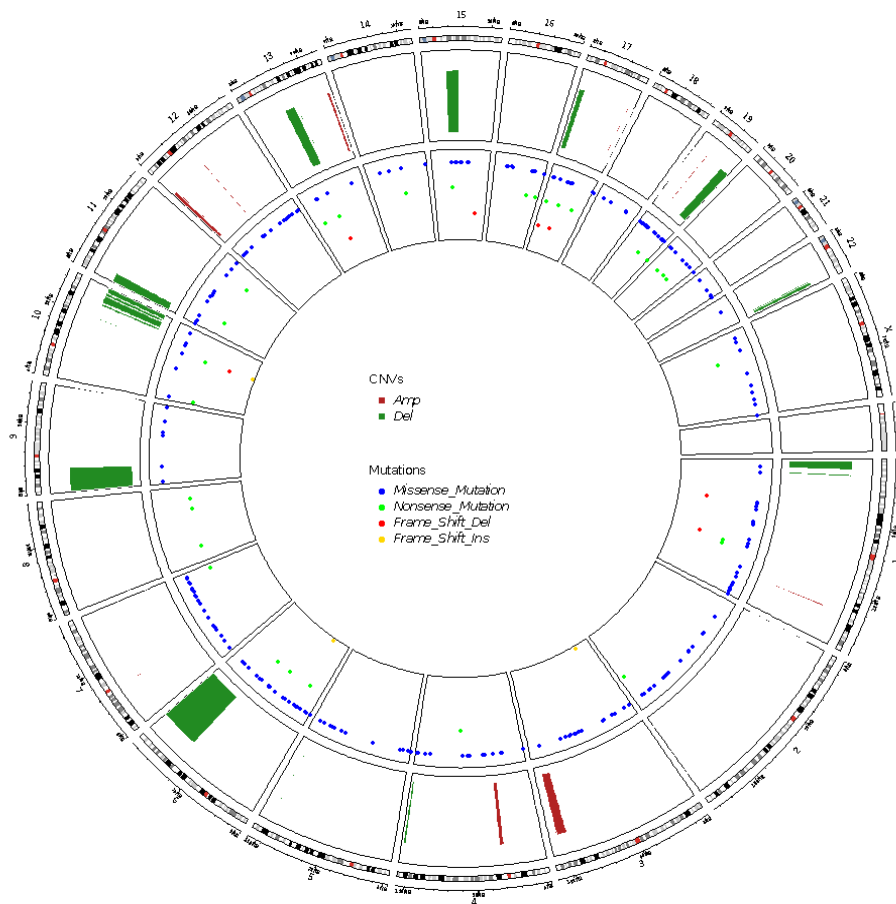
Figure 3.17 Genomic copy number variations in glioblastoma.

Red shades indicate the amplification and green shades stand for deletion. Dots indicate point mutation i.e blue dot is a missense, green is nonsense, red is a frameshift deletion and orange is a frameshift insertion.

I observed recurrent alterations in specific parts of the chromosomes especially chromosome 1, 3, 4, 7, 12, 13, 17 and 19 were characterized by amplification. I further observed deletions at chromosomes 1, 6, 9, 10, 11, 13, 15, 17 and 19.

Both analyses showed that LGG was characterised by gene amplification at chromosomes 1, 3, 4, 7, 12, 13, 17 and 19 while gene deletions were observed at chromosomes 1, 2, 4, 6, 9, 10, 11, 14, 19 and 22. Although such alterations are events, the mechanisms underlying the development and tumour progression is not well known. It is further important to determine the possible mechanisms that may explain the disease progression or malignancy. This could partly be shown through ingenuity pathway analysis of genes involved along with functional validation assay. I highly recommend such experiments.

3.7.3.1 Hierarchical clustering of low grade glioma primary tumour (LGG-PT)

We further conducted pearson- ward D2 hierarchical clustering as shown in the dendrogram in figure 3.17 below.
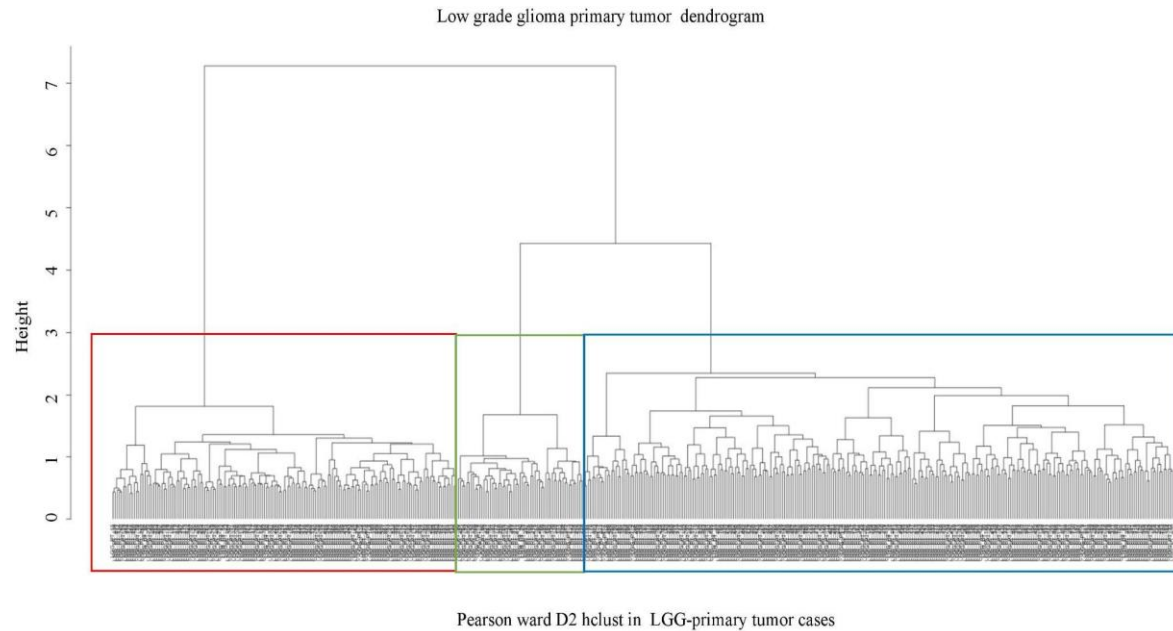
Figure 3.18.Three groups obtained after pearson-ward. D2 hierarchical clustering in low grade glioma -primary tumour (LGG-PT).

This was obtained after clustering n=586 samples of low-grade glioma.

3.7.3.2 Frequency copy number variations of the identified 3 cluster groups in LGG-PT

Low grade glioma_primary tumour C1; Thresholds=[-0.1,0.1]

Low grade glioma_primary tumour C2; Thresholds=[-0.1,0.1]
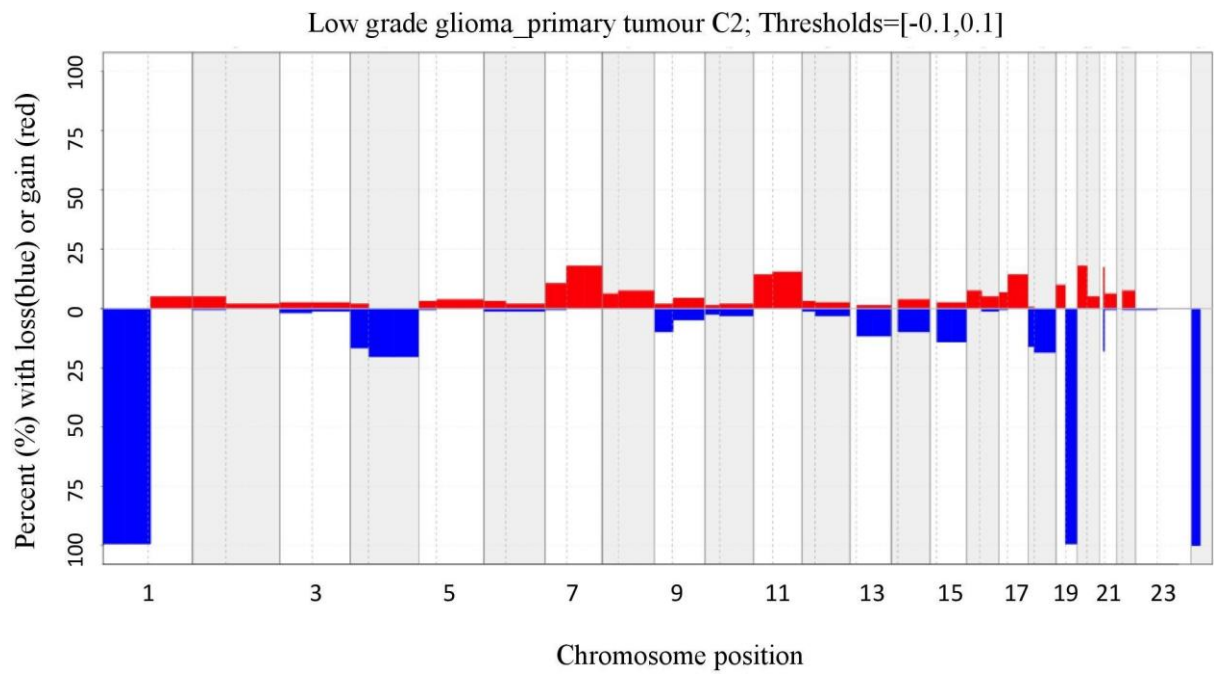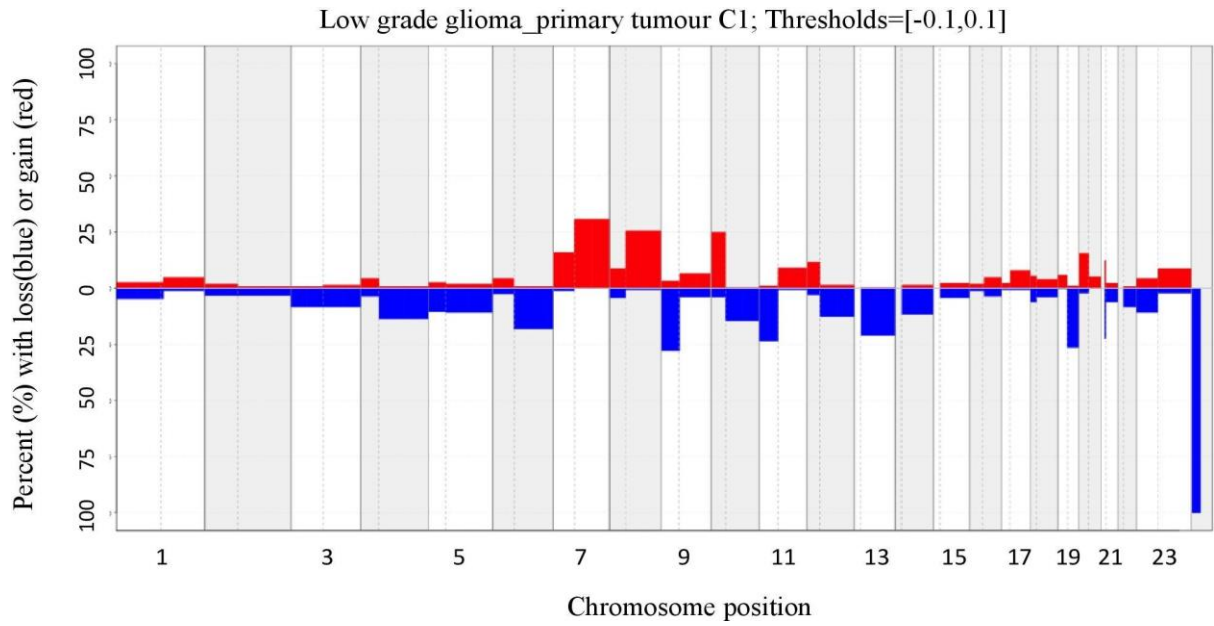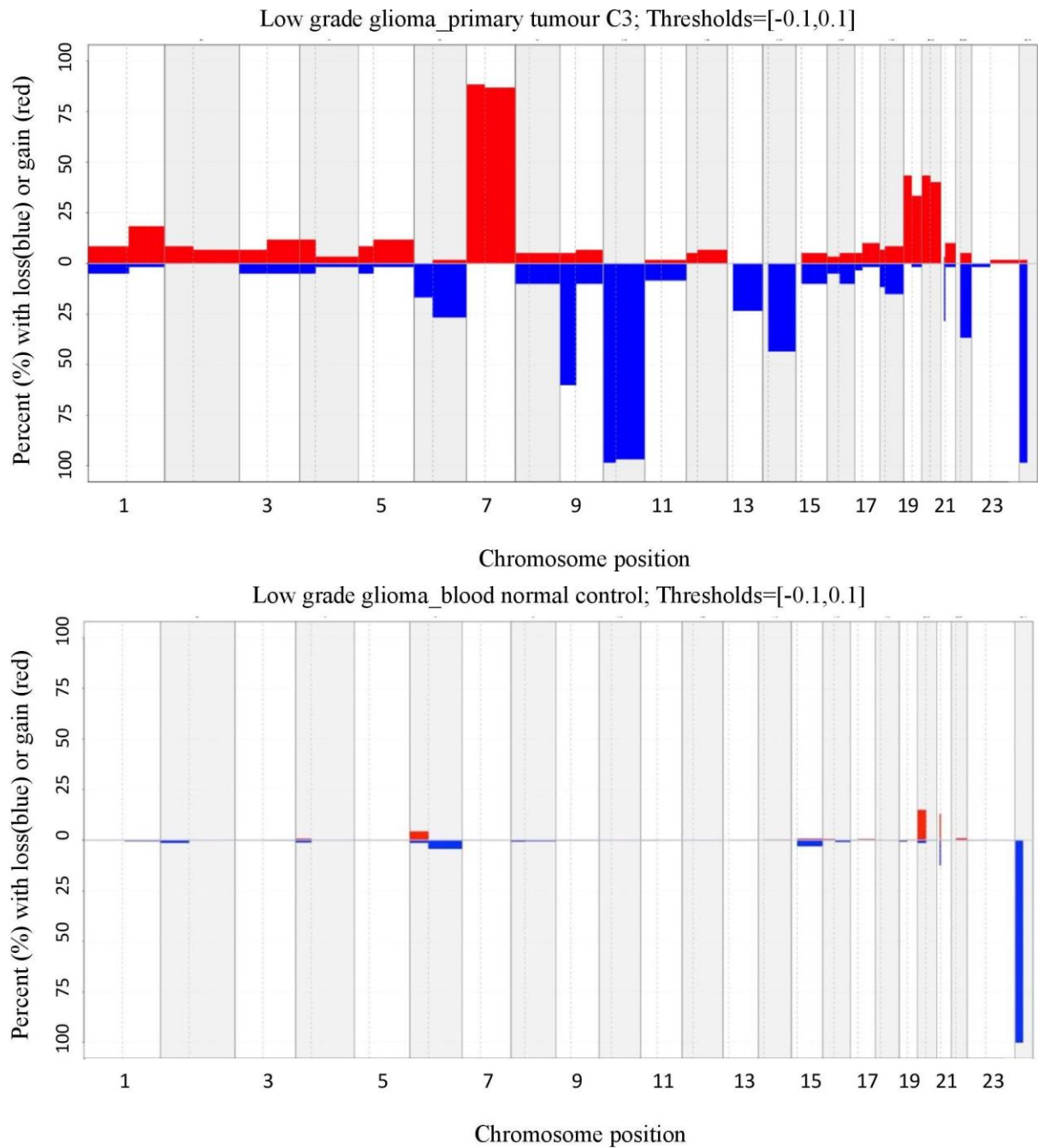
Figure 3.19 CNV frequencies in LGG-PT sub-groups identified.

Cluster 1, low grade glioma-primary tumour (LGG_PT1, n=277), cluster 2 (LGG_PT2, n=162), cluster 3(LGG_PT3, n=60) and low grade glioma normal blood control (LGG-BDN (n=472) from left to right respectively.

3.7.3.3 Heat- map analysis of the identified 3 cluster groups in LGG-PT

Low grade glioma_primary tumour C1; Thresholds=[-0.1,0.1]

Chromosome position

Low grade glioma_primary tumour C2; Thresholds=[-0.1,0.1]

Chromosome position
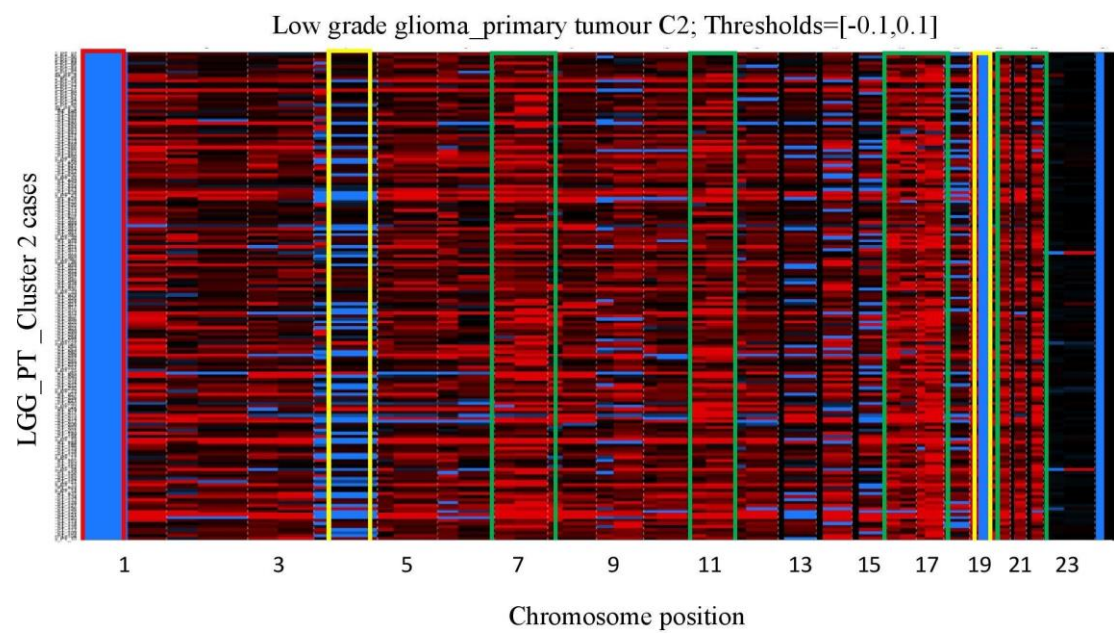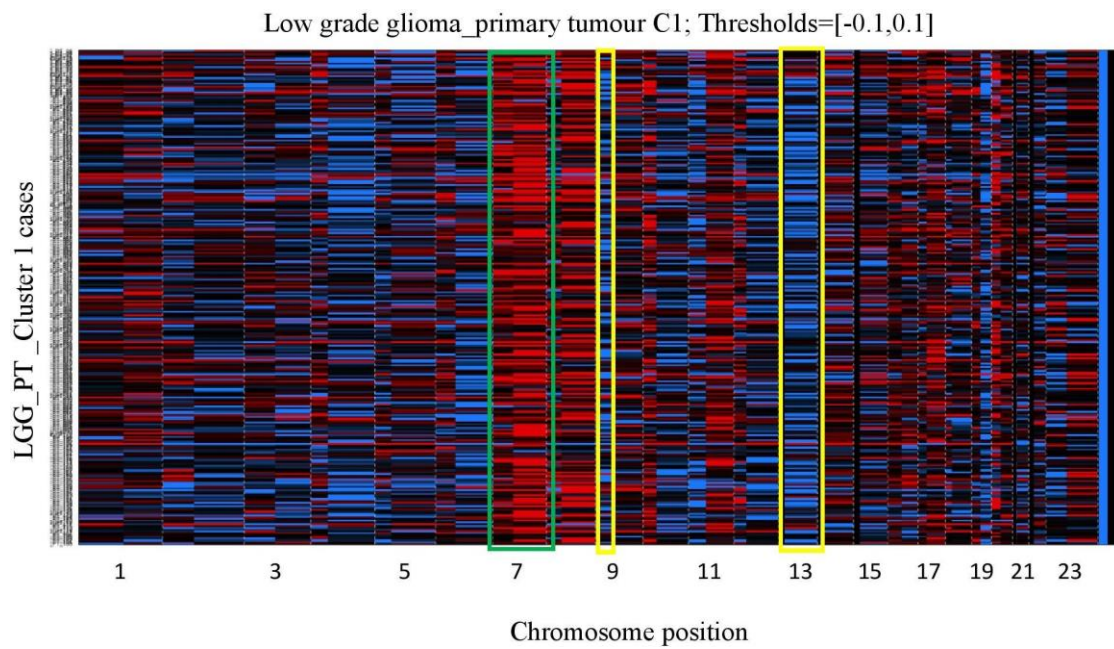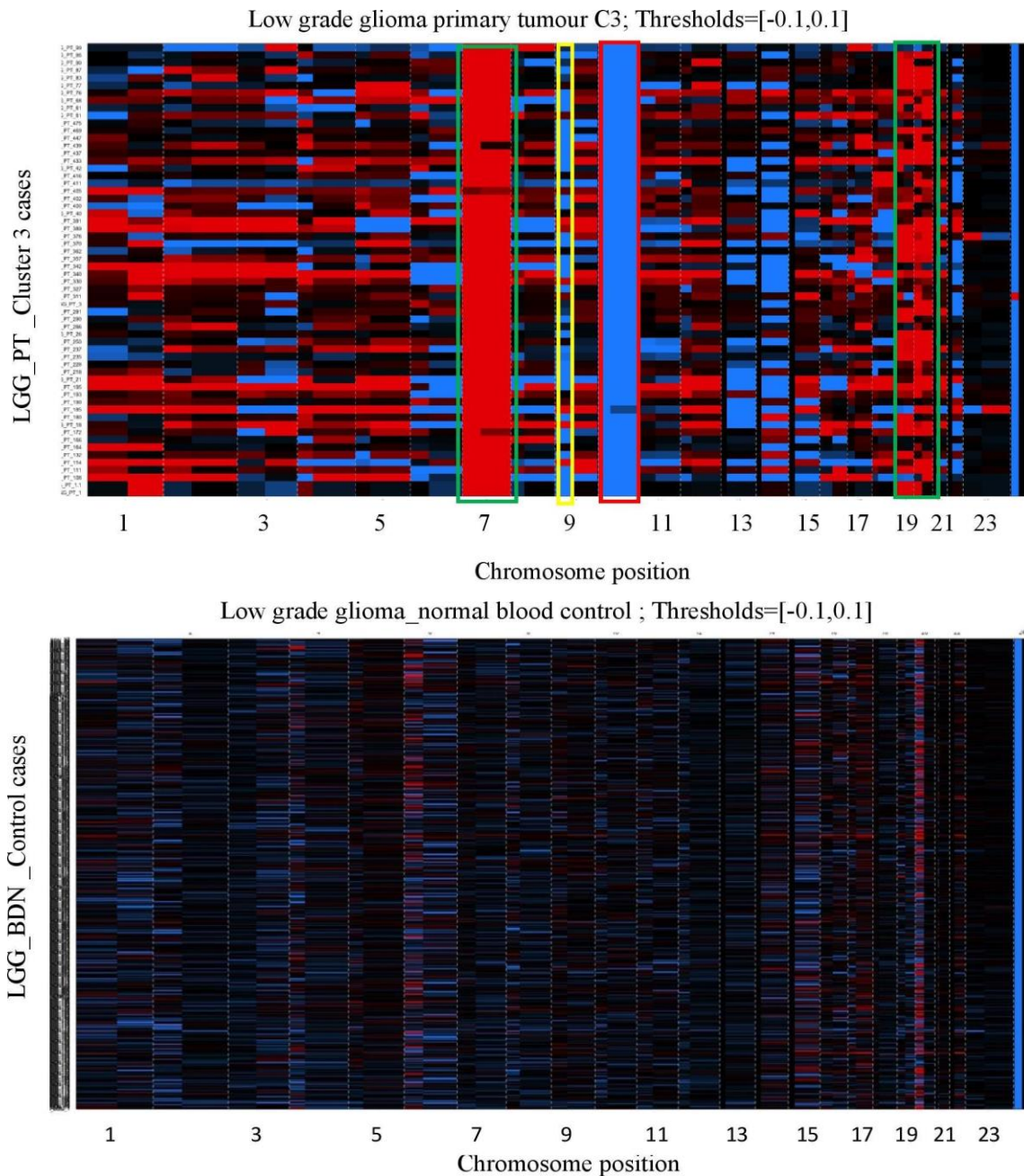
Figure 3.20 CNV profiles pattern in LGG-PT groups 1-3 (n=497) and LGG-BDN (n=479).

From left to right respectively. The LGG-PT group1 (n=277) would be characterized by chromosomal 7 gain. Group 2 (n=162) would be characterized by chromosomal 1 loss, chromosomal 4q and major chromosomal 19 loss. Group 3 was characterized by chromosomal gain at chromosome 7p/q and chromosomal loss at 10p/q and 9 q.

3.7.4 Glioblastoma primary tumour clusters (n=583)

For the 583 glioblastoma primary tumours (GBM-PT), we obtained 4 main sub-groups with different numbers of cases after clustering. GBM-PT group 1 (n=65), group 2 (n=264), group 3 (n=131) and group 4 (n=123). Figure 3.20 below shows the cluster groups.
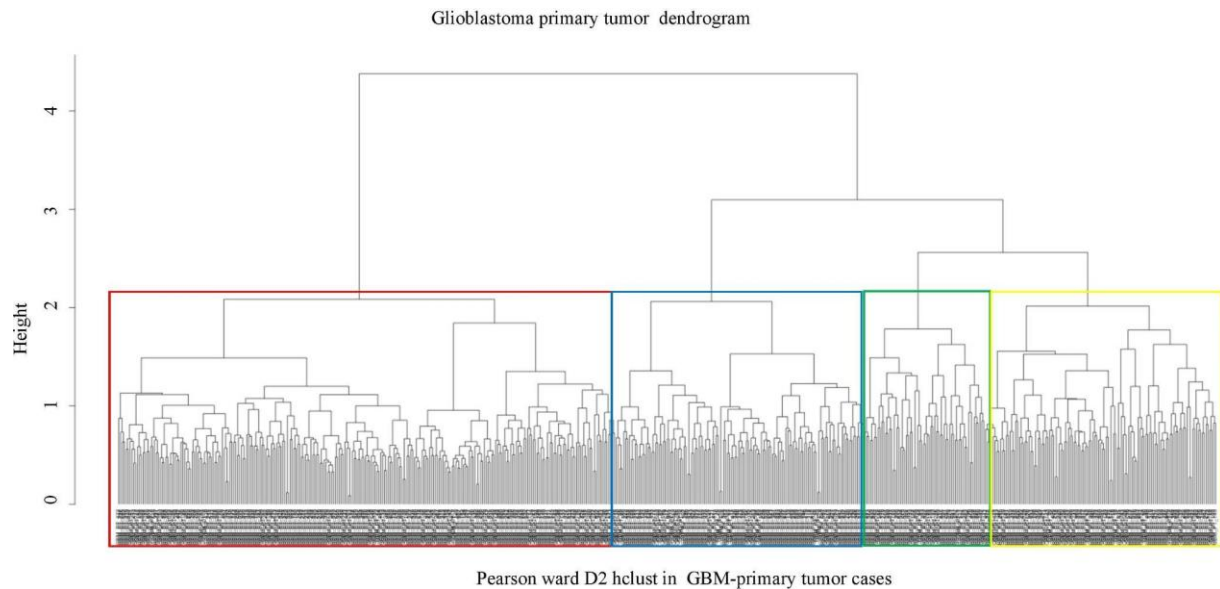
Figure 3.21.Hierarchical clustering (Pearson-ward. D2) of glioblastoma primary tumour n=583.

Four distinct groups were observed based on copy number variations log intensity values. The glioblastoma recurrent tumours (GBM-RT n=14) were divided into two subgroups. Due to the low number of samples in the GBM-RT cluster, I interpreted the data with cautions.

3.7.5 Copy number variations profiles in the five GBM-PT sub-groups

Glioblastoma_primary tumour C1; Thresholds=[-0.1,0.1]

Glioblastoma_primary tumour C2; Thresholds= [-0.1,0.1]

Glioblastoma_primary tumour C3; Thresholds= [-0.1,0.1]
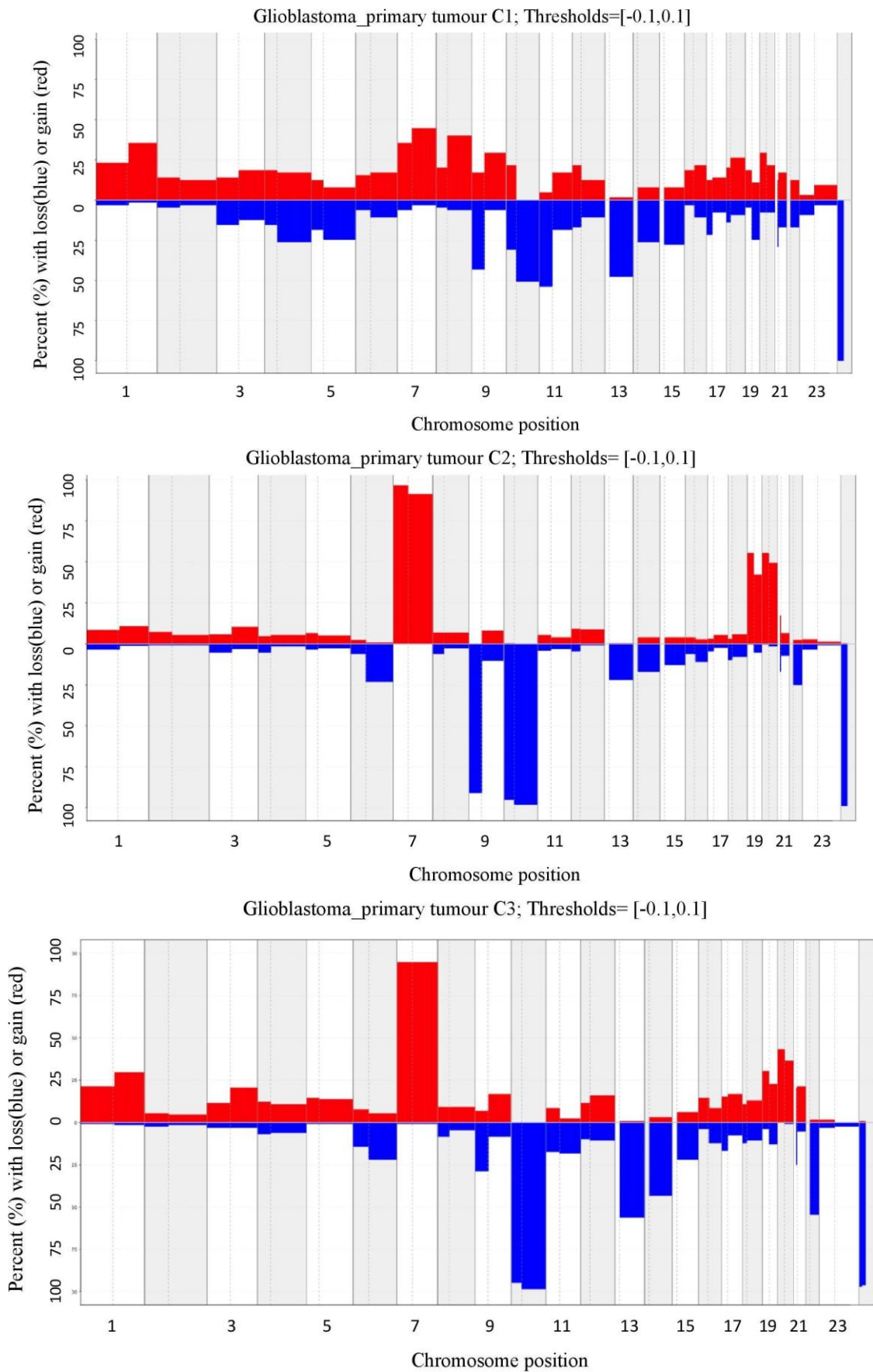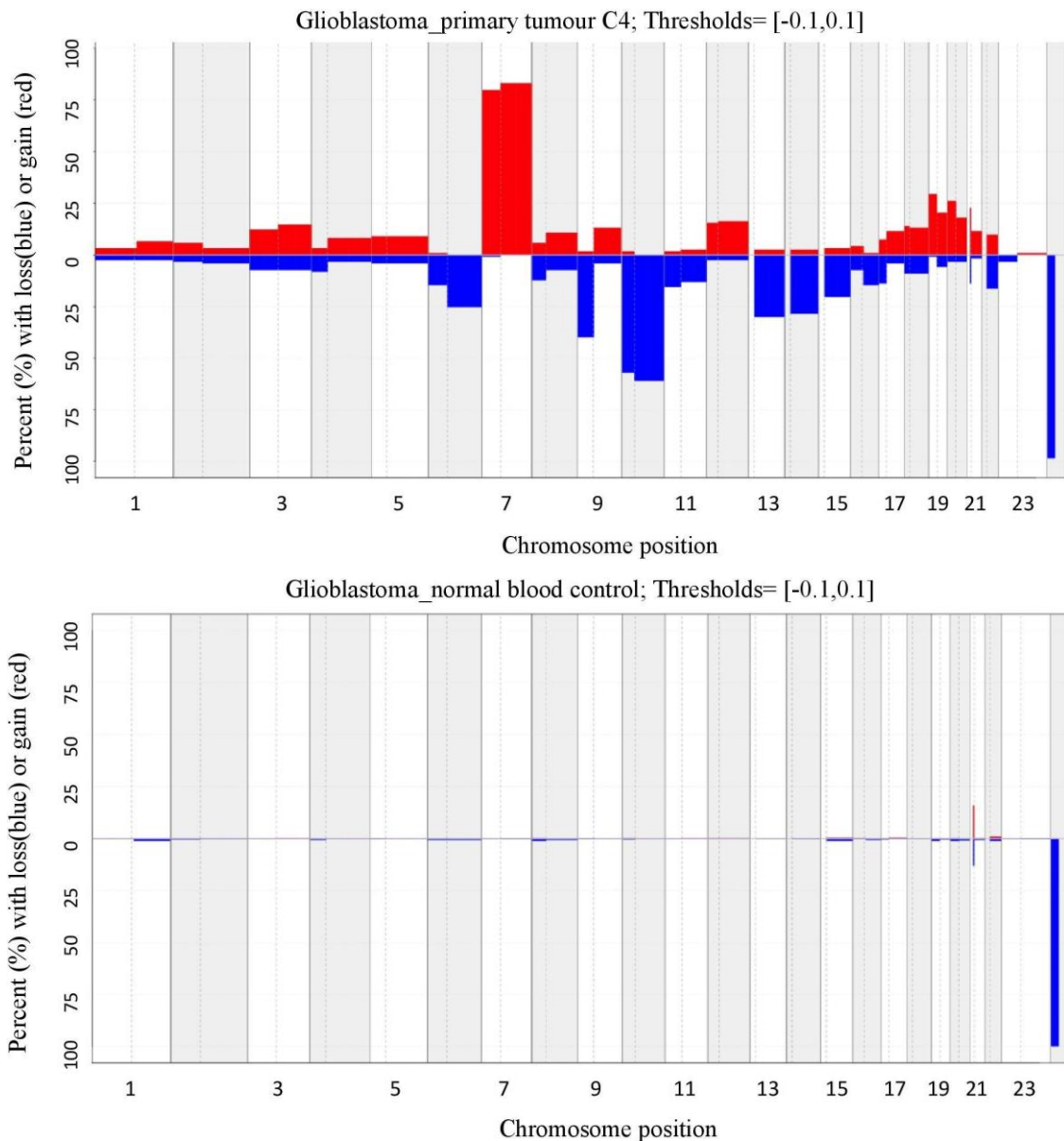
Figure 3.22 CNV frequencies in glioblastoma primary tumour sub-groups (GBM-PT).

Groups 1-4 from right to left respectively along the genome chromosome 1 to 24 in an alternate manner. The CNV frequencies were compared to glioblastoma blood normal control (GBM-BDN control). The primary tumours were characterised by common gain (red bar) with highest frequency at chr 7, deletion/loss (blue bar) at chr 9p, 10p/q and 22q. These regions are enriched with cell survival and apoptotic genes such as *CDK2A/B, MDM2, EGFR* and *PTEN* which are common in high grade glioma.

To identify common alterations, I checked for deletions or gains at high frequency in at least two of the clusters. We observed common amplifications at chromosomes 7p/q, 9q,12p/q, 18p/q, 19p/q and 20p/q. Common deletions were observed in chromosomes 9p, 10p/q, 13q,

14q, 15q, 16p/q and 22q. Deletions at chromosomes 6p/q were evident in group 1, 2 and 4 while gain at chromosome 1p/q was evident in group 2 and 3. Consequently, I investigated the GBM-RT two groups along with the 479 case GBM-blood normal control (BDN). Figure 5 shows the CNV frequency. We observed minor alterations at chromosome 4p, 6p , 13q, 19p in GBM-BDN but in <1% of the total samples analysed (n=479).
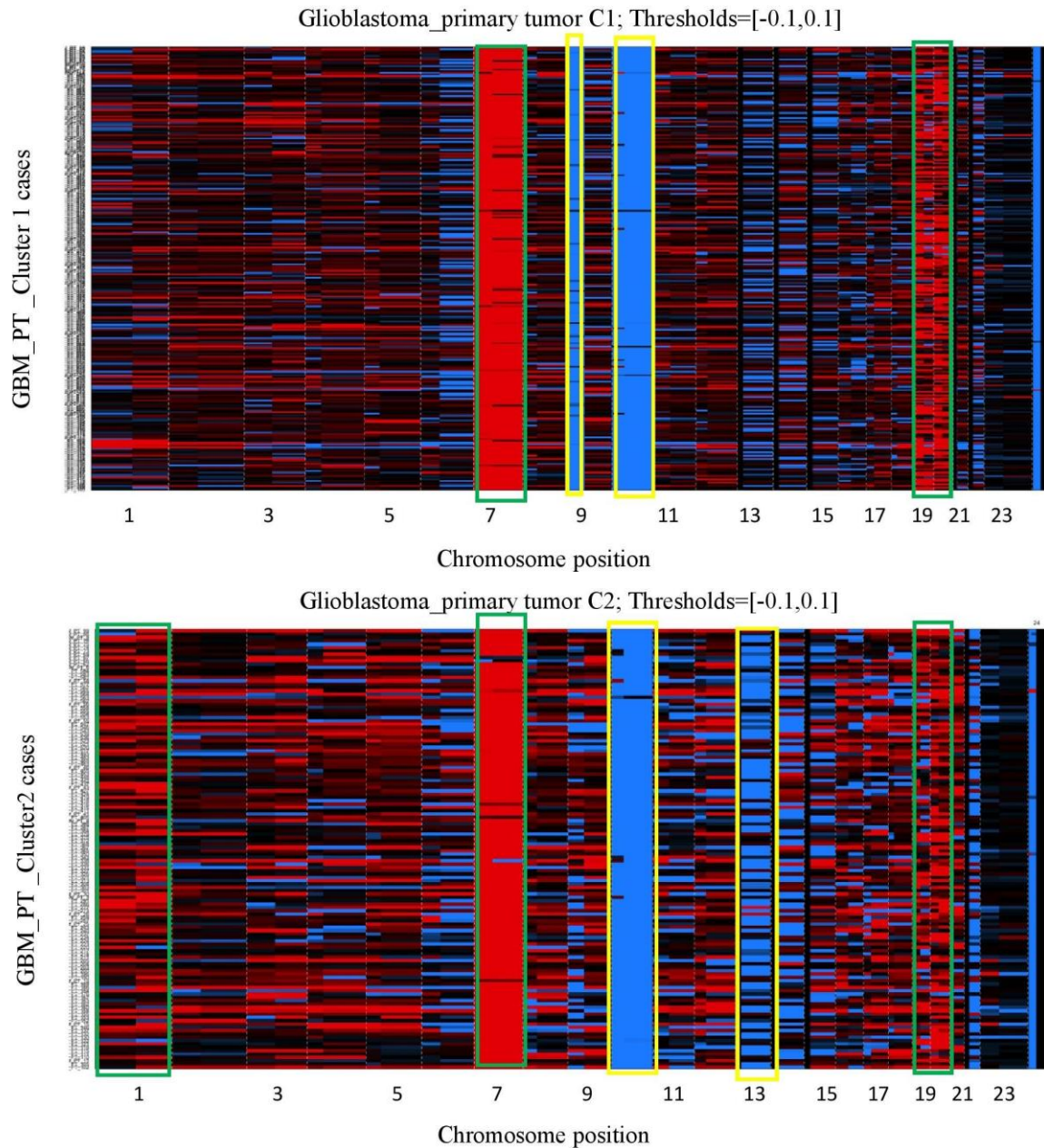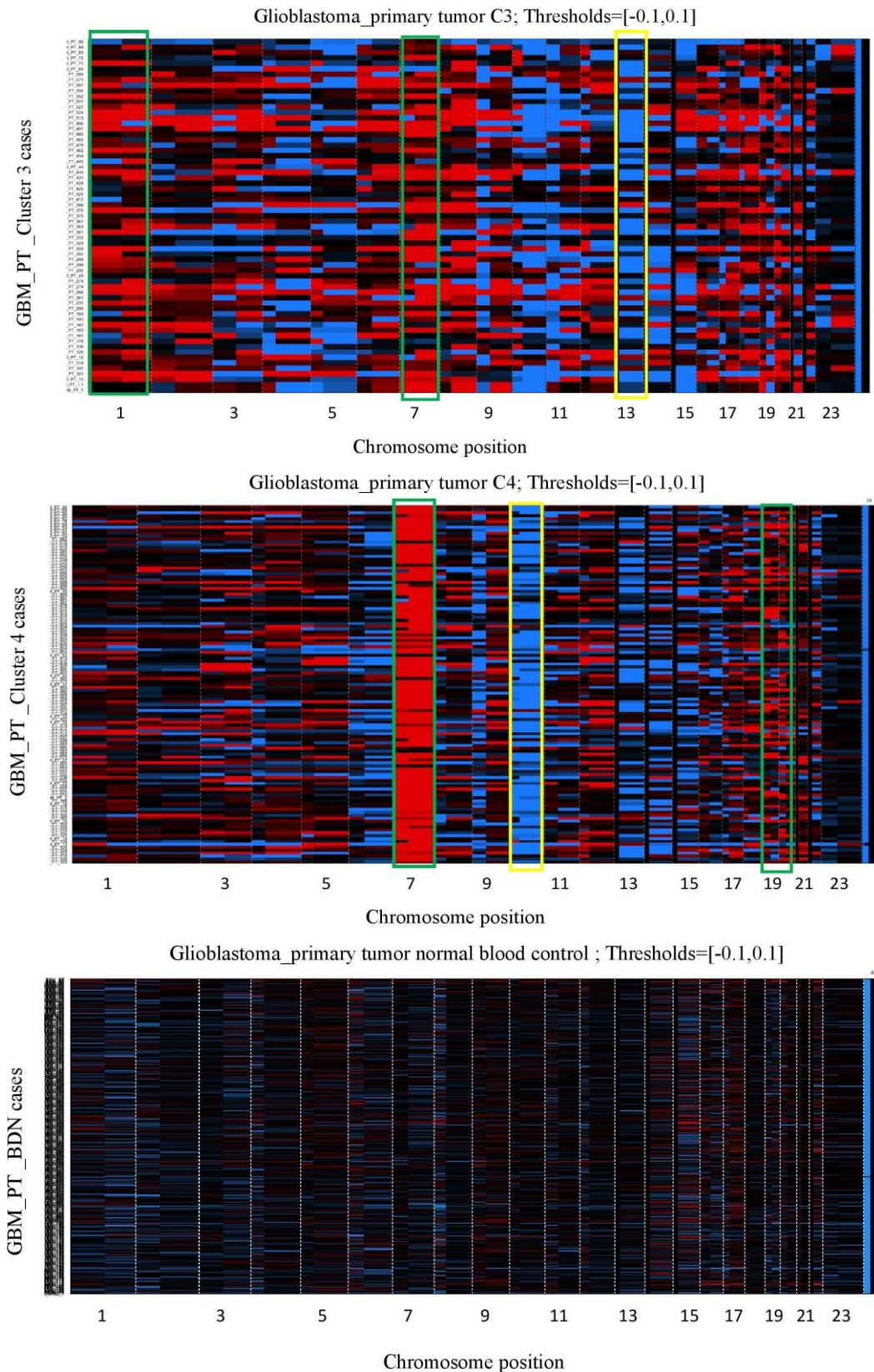
Figure 3.23 CNV profiles patterns following heat map analysis of each sample case in each of the GBM-PT four clusters.

Clustering was done using the log-intensity value cut-off (-0.1, 0.1) as loss and gain respectively. The profiles were compared to the GBM-BDN controls (n=479).

Using heat map analysis, we could establish key CNV profile patterns that occurred in high frequency beside the side of the log intensity value i.e dark blue means a deletion, dark red is a gain while a log intensity value close to loss or gain will have light blue and red respectively. Dark nuisance means a neutral log intensity value i.e log intensity value of equal or close to 0. Using this analysis, we could clearly see that clusters 1, 2 and 4 of glioblastoma multiforme would be characterized by gain of chromosome 7, major loss of chromosome 10 p/q with or without loss of chromosome 9q. Cluster 3 n=65 had a unique pattern characterized by chromosome 1q gain with or without loss of chromosome 13q. We also noted that group 2, 3 and 4 had high frequency of chromosomal 13 q losses. Note that the loss in chromosome 24 was seen in blood control as shown in the figure 3.23 below and thus we could not make significant conclusions with such loss.
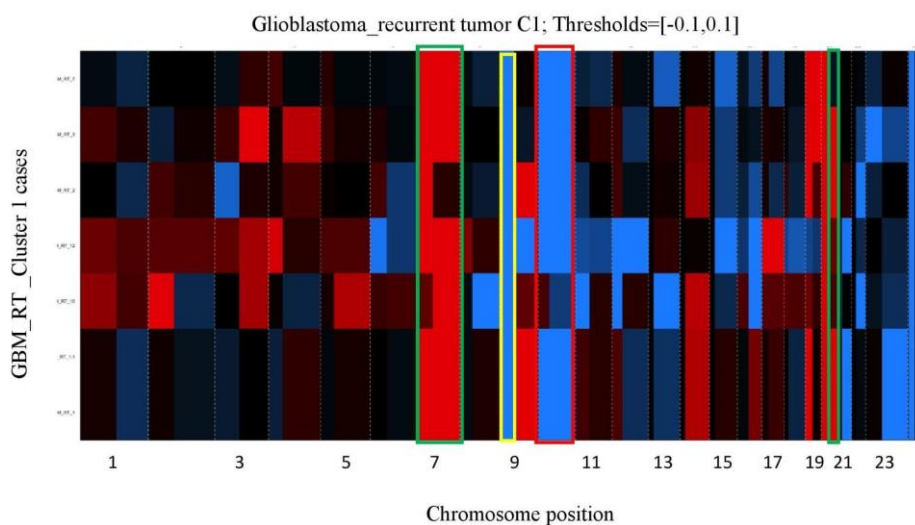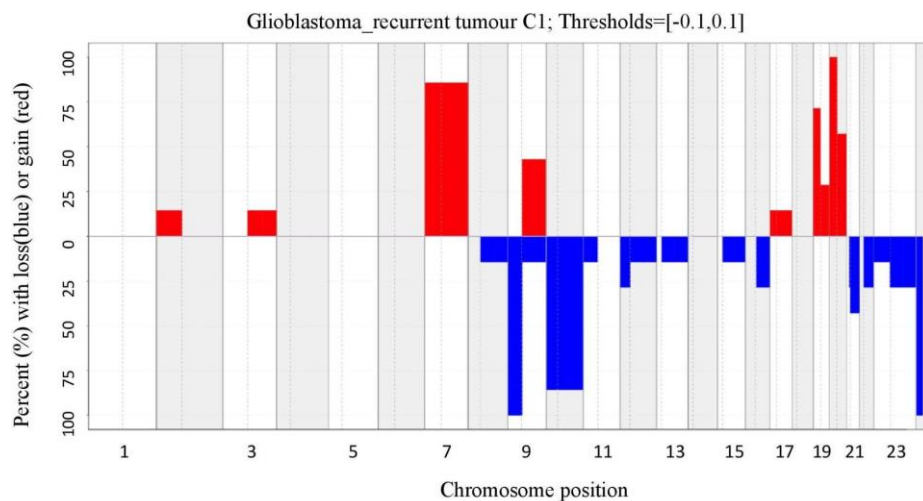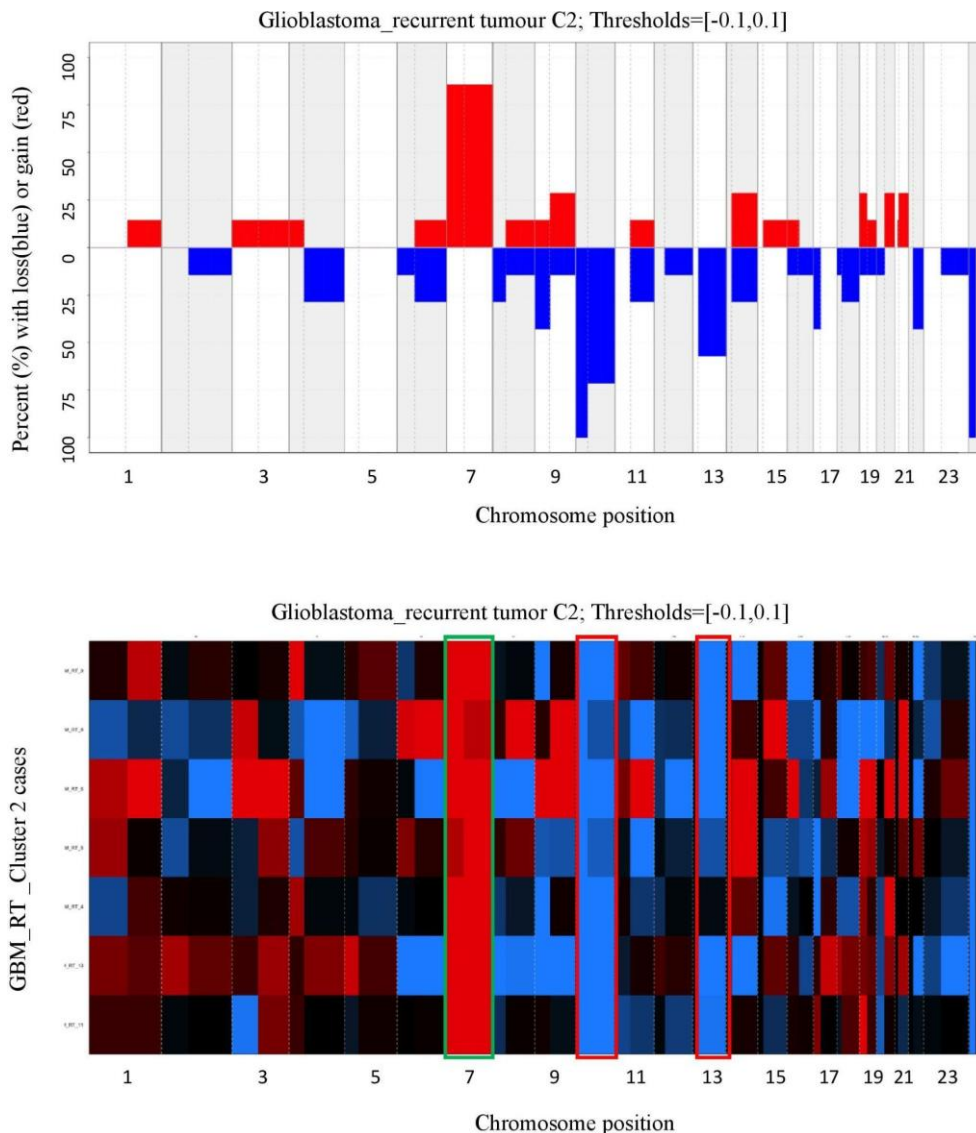
Figure 3.24 Alteration in GBM recurrent tumour (GBM-RT) cluster 1/2 (RT1/RT2) and their respective profile heatmap profiles in group1 and group2.

Cluster 1 is on the top and group 2 is on the bottom. We observed common gain of chromosomes 2p, 3q, 7p/q, 9q, 17p/q, 19p/q, 20q and 22 p while common loss of chromosome was highly observed at chromosomes 10p/q and 13q. Other common losses at low frequencies were observed at chromosome 9p/q, 8q, 15q, 16q and 22 q. Group 1 had unique gain at chromosomes 1pq, 3p, 4 p/q, 5q, 6p/q, 8q ,9p, 11q, 14q, 15q ,16p, 18p, 20q and 21q. The unique chromosome loss was found at 1p, 2p/q, 3p, 4p/q, 5p, 6p/q, 8p, 11q, 14q, 15q, 16p, 17p, 18p/q, 19p/q and 20p. Group 2 had unique chromosome loss at 12p and 21q while unique gain was only observed at chromosome 20p. Based on the heatmap profile, the GBM-RT groups could be characterized by gain of chromosome 7 and chromosome 10 deletion. Group 1q could be differentiated by having deletion at chromosome 13 while group 2 had unique deletion at chromosome 9.

4.0 Discussion

4.0.1 Epic 450K/850K data is reliable in CNV identification like aCGH

In this study, for the first time I demonstrated that copy number profiles could be determined from methylation data with moderate to substantial accuracy like the aCGH data. These findings are important since efforts of in-cooperating methylation data and copy number alterations in grading brain tumour entity is under way if supportive evidence comes to our knowledge.

I used a circular binary segmentation approach to determine change points in the aCGH data. This approach has been reported earlier by Olshen and Venkatraman (2002) [97]. The model detects change-points which corresponds to positions where the underlying DNA copy number has changed. Change points can therefore be used to detect regions of copy number alterations (gain and loss)[107]. Our approach to determine the cut-offs for the CGH data was highly efficient since 94% of balanced segments data were closer to log2 (R/T) of zero.

Overlaps in chromosomal segments were removed using the genomic range package and disjoin function. At the end of this analysis, I was able to come up with similar chromosomal regions within the two data set which we could compare. This approach ensured fewer false positives if any. To determine if the two methods produced essential similar results in calling the CNA, we calculated the percent agreement, and observed 85.8% agreement in CNA calls between the two methods. Usual percent agreements below 61% are seen as problematic and most clinical studies have recommend higher percent agreement. On average agreement above 79.5% (80) are highly recommended especially for clinical applicable routine methods[103,109,110]10 samples had a percent agreement of <61%. Our limitation of 4 of this set of samples was to establish the baseline cut-off values for calling the segments as gain or losses in either of the data sets and this could explain the high level of disagreement of around 40%. The remaining six could be attributed to failed bisulfite conversion during the methylation experiment. Moreover, I observed that the mean segment values of segments called balance were not close to zero as compared to the other samples plateau segments. However, this was reported as a common problem observed in calling segments as gain or losses[107].

To test extent of agreement observed, I computed kappa statistics in the three categories of responses (balance, gain and loss) as guided elsewhere[110,111]. I observed a high percent agreement of 80 and a lower k of 0.57 upon the kappa calculation. The discrepancy can be explained in the fact that kappa value is known to significantly decrease if response categories are more than two[93,114,115]. It should also be noted that the kappa value is squared for interpretation since the correlation coefficient cannot be directly interpreted but a squared correlation called coefficient of determination (COD) is used directly. It is therefore recommended to interpret the kappa value with caution. I accepted 0.40 to 0.60 kappa value as "moderate" which implied that, 0.40 is an adequate agreement in our case. A kappa value of 0.41 with a good percent agreement is acceptable in most health studies[103,110]. To confirm the test we further used intraclass correlation coefficient (ICC) which is known to be similar to weighted kappa especially when we have numerical data[109,110]. Interestingly both weighted kappa, unweighted kappa and the ICC values were similar k=0.574 with 95% confidence interval (C.I) of 0.54-0.61. These findings represent a widely applicable resource for the methylation data that can be used in copy number analysis. Search CNA results will find wide application not only in brain tumour but possible to other entities too[112]

4.1.0 WHO astrocytoma grade II and III had no significance difference in survival

I had grade II and III astrocytoma samples in our analysis. Grade II were considered as infiltrative tumours with ability to progress into chronic stage. Grade III are lesions with histological signs of malignancy. Grade III mainly included anaplastic astrocytoma which means intermediate however in reality it is maximal malignancy. Patients with grade III usually receive chemotherapy and often adjuvant radiations[113,114]. Grade III are usually associated with recurrence while grade IV are malignant neoplasm which are mitotically active and have fatal outcome[114]. In the glioma family, glioblastoma are grade IV neoplasm. Despite this classification in WHO grade, there was no difference in survival between the patients classified as grade II and III as shown in figure 4.2a (P> 0.05). This could be attributed to the fact that grade II astrocytoma which were mainly pilomyxoid astrocytoma, pleomorphic astrocytoma, xanthoastrocytoma and diffuse astrocytoma were revised and included in grade III[55,79]. In brief the revision led to diffuse gliomas to include both WHO grade II and III and leaving out both pilomyxoid astrocytoma and pleomorphic astrocytoma due to lack of *IDH* alteration but rather have *BRAF* alterations[112]. This group of astrocytoma are known to have similar phenotypes

and genetic similarities which may explain the similar survival outcome observed beside diffuse astrocytoma have ability to progress to anaplastic astrocytoma and glioblastoma[55,114]. The ability of the astrocytoma to progress to high grade tumour could explain the observed unique mixed methylation classes between astrocytoma and high-grade astrocytoma. Our data also provide evidence that the WHO grades may not accurately predict the severity of disease and more efforts are required to stratify patients with brain tumours for precise management.

4.1.1. Hierarchical cluster based on CNV log2 intensity value

Genomic DNA regions are frequently lost or gained during the tumour progression process. The lost regions have been shown to include some tumour suppressor genes while amplified regions sometimes harbour oncogenes[115]. This would ultimately contribute to pathogenesis and outcome of disease. I therefore opted to test whether the CNVs would cluster well based on severity of the WHO grade II and III. After conducting clustering, we obtained 7 distinct clusters (figure 4.2). I also found out that the group had a significant difference in survival (figure 4.2b). We closely observed the survival curve and discovered that they could be grouped as poor and better survival. I further checked the frequencies alteration in this group and indeed found increased alteration in the so-called poor survival group as compared to the better survival group as shown in figure 4.3a and 4.3b. This indicates that CNVs can aid as a prognostic factor. Similar findings were recently reported[116].

4.1.2 Classifier

All samples belonged to the glioma family, majority in Anaplastic (A)_IDH class, subclass astrocytoma and some clusters especially in clusters 3, 4, 5 and cluster 7 had some samples classified as A_IDH high grade (HG) subclass. I could not rule out the presence of oligodendroglioma due to presence of chr1p/19q co-deletion in some samples which were majorly enriched in cluster 7 (C7). Other samples were classified as A_IDH. Some samples could not be compared to reference well (score <= 0.9) however they were classified based on maximum score between the tumours as A_IDH. This is a common challenge as reported in earlier study[50].

I further found out that there was a mixed distribution of WHO grades. In specific clusters 1, 2, 4 and 7 were dominated almost equally by each grade. Cluster 3, 5 and 6 were dominated

by WHO grade III. Although WHO categorized the majority of samples as grade III, some were classified as methylation profiling subclass A_IDH_HG subclass and some as A_IDH. These results provide a possibility to use CNVs in reclassification of these WHO grades as more supportive clinical data becomes available.

Moreover, I also observed common and unique CNV alteration patterns among the 29 common genes. Genes differed in alterations between clusters but to an extent similar within each cluster sample. This indicates that the cases within a cluster have similar CNV profiles with minor differences in alterations. The common genes altered between clusters e.g *CDKN2A/B* in cluster 3 and 4 may indicate possible role of the gene leading to early events or progression of tumour to glioma but may or may not play a significant role in severity of the disease. We hypothesize that unique genes in each cluster may be usefully contributing to the outcome/severity of the disease of a specific cluster which may influence survival too. Detailed analysis in the distribution of the alterations and genes in each cohort will give more insight into how we could apply copy number alteration in further classification and grading of brain tumour, an aspect which I am working for at the moment.

4.1.3 *CDKN2A/B, RB1* and *MYB1* losses combination was associated with poor survival
4.1.3.1 Cyclin-dependent kinase inhibitor 2A/B (*CDKN2A/B*) loss

CDKN2A gene is located in Chr9p21. *CDKN2A* and *CDKN2B* encode p16 INK4a (cyclin-dependent kinase inhibitor 2A) and p15 INK4b which acts as tumour suppressors. p16 INKA inhibits cyclin-dependent kinase 4 (*CDK4*) and p15 cyclin-dependent kinase 5 (*CDK5*). *CDK4* and *CDK5* play an important role in different cell functions and regeneration[117]. Several genetic polymorphisms have been suggested in the upstream of *CDKN2A/B*, which might influence the expression of these genes and thereby cell cycle[118]. Moreover a deletion of *CDKN2A* gene may result to cell cycle acceleration and cancer through production of non-functional p16 which is key in suppressing multiple tumours[119]. The oncogenic process resulting from Chr9p21 deletion may also result due to dysregulation of apoptotic pathway[120]. WHO grade II-III have varied clinical presentations with anaplasia depending much on mitotic activity which is associated with C*DKN2A* loss. The loss of *CDKN2A* in-turn affects cell cycle and has also been implicated in poor survival in astrocytoma patients[121]. This could partly explain the poor survival observed in C1, C2, C3 and C5. In summary *CDKN2A* is associated with poor survival in gliomas but not in oligodendroglioma[121]. Astrocytoma are generally known to have less

mitotic activity than oligodendroglioma thus it is promising to use *CDKN2A* in molecular grading of WHO grade II and III. Further advantage of *CDKN2A* is the ability to predict poor survival in patients which would be helpful in-patient management too.

4.1.3.2 Retinoblastoma (*RB1*) loss

RB1 gene codes for RB1 protein which is a tumor suppressor used in determining therapeutic efficacy. It also plays a crucial role in the transition of cells from C1 to S phase. It has been shown that cyclin dependent kinases (*CDK4* and *CDK6*) and D-type cyclin leads to *RB1* phosphorylation which in turn releases the elongation factor 2 (*EF2*) transcription factor which results in activation of genes involved in G1 to S phase transition[122]. In our data set the deletion of this *RB1* gene was evident in C1, C2, C3 and C5 clusters.

The relative genetic simplicity of retinoblastoma, however, belies the significant functional complexity of the *RB1* encoded protein (pRb). The first cellular function described for pRb, and the most thoroughly studied, is as a negative regulator of the cell cycle [123]. Loss of pRb-mediated cell cycle control is frequently observed in cancer[124]. As cancer is a disease of abnormal cell proliferation, it makes intuitive sense that the key function underlying pRb mediated tumor suppression is cell cycle regulation. However, pRb loss also has profound effects on many other cellular processes relevant to cancer including differentiation, survival, senescence and genome stability[125]. This functional complexity is mirrored by the variety of molecular interactions involving pRb. Rb1 protein interacts with a large and steadily growing list of cellular proteins, and an even greater number of genes. Deriving satisfying general cancer principles from the study of pRb thus remains elusive[125–127]

4.1.3.3 *MYB* gain and loss

Studies of high-grade glioma in pediatrics also found no amplification of the *MYB* gene but rather a loss of *MYB* in a few cases (6/54) was detected in high grade gliomas[128]. However, in low grade glioma copy number gains in *MYB* at chromosome band 6q23 were observed in 2 of 14 samples studied[128]. This suggests that *MYB* upregulation may be modulated by alternate mechanisms, in addition to copy number alterations. These results partly correspond with our observation of *MYB* alteration in our data set. I recommend more studies of *MYB* gene distribution in glioma in large cohorts.

4.1.3.4 Pathways and molecules associated with astrocytoma sub-classes

To get more insight about the altered genes, I selected the genes that are located within the altered regions and conducted ingenuity pathway analysis. I was able to identify the canonical pathways associated with each group "altered genes". I selected the most relevant pathway such as the glioma signalling pathway to identify known and unknown genes that may be attributed to the risk of brain tumor. The list of this pathway and associated molecules are listed in supplementary table 2. I observed fairly unique patterns in each group with different pathways associated with the disease.

Notably, *PTEN, ERK/MAPK, P53, IL-3,* glioblastoma multiforme, glioma invasiveness and axonal guidance signaling which are associated with glioma formation featured in most clusters and were associated with altered adenomatous polyposis coli (*APC*) which is a tumor suppressor gene. Studies have shown that individuals with mutations in *APC* at chr 5q21 have increased risk of brain tumours[129]. The Glycogen synthase kinase 3 beta (*GSK-3β*) which affects cell proliferation was also altered within the pathways. Inhibition of *GSK-3β* has been associated with glioma death and poor proliferation of glioma cells[130]. Other molecules altered in the pathways was Retinoblastoma (Rb) codes for a tumor suppressor rb protein, Platelet-derived growth factor (*PDGF*) which do regulate cell growth, Phosphoinositide 3-kinases (*P13K*) which regulate cell growth and other cellular functions, Transcription factors that affect cell cycle (E2F) were also lost, a proto-oncogene Rat sarcoma (*Ras*), WNT, Son of sevenless (*SOS*), Auditory processing deficit (*APD*) and Beta catenin (*CTNNB1*a).

Over activation of Wnt/beta-catenin/Tcf pathway affects several cancer types however constitutive activation of Wnt/beta-catenin/Tcf signaling pathway in specific was strongly associated with astrocytic tumors[131].

Cluster 4 was associated with poor survival. The presence of the GSK3 beta and *AXIN* complexes alteration may indicate the ability of the tumour to progress to astrocytoma. *GSK3 beta* has been shown to regulate different cellular activity ranging from cell growth to apoptosis[132].

Low expression of *BDNF* would be associated with alzheimer disease while mature *BDNF* has been shown *in vitro* to inhibit C6 glioma cell apoptosis and leads to cell growth[132,133].

Moreover, upregulation of *BDNF* is usually evident in gliomas[134]. The presence of altered *BDNF* in our data set may therefore signify the ability of the astrocytoma cases to progress to other chronic glioma malignant forms.

PPFIA2 gene is known to encodes a member of LAR protein-tyrosine phosphatase-interacting protein (liprin) family, which plays a role in axon guidance and neuronal synapse development by recruiting LAR protein-tyrosine phosphatases to the plasma membrane beside known to be regulated by cannabis to bring about neuropsychological functions[135]. Alterations may interfere with axonal guidance hence may contribute to neurological sequelae observed.

The other pathways that had a high ratio of molecules from our data (> 0.2) in the canonical pathways include axonal guidance signaling. Molecules such as netrins and semaphorins found in axonal guidance pathways have been associated with neuronal migration and survival control through repulsion or attractions towards growing axons[136–138]. Other studies have also associated the axonal guidance molecules such as ephrin with pathological processes including cancer by enhancing angiogenesis process[137,139–141].

Molecules on MAPK/ ERK pathways were also altered at high ratio. The MAP/ ERK signaling molecules under normal circumstances are highly expressed in brain and aid in memory formation, pain perception, induction of cortical neurogenesis, midbrain and cerebellum development[140,142–145]. Beside this role altered ERK has been associated with pathogenesis of higher-grade astrocytoma[143]. In our data set we observed gain of *PDGFR* and *EGFR* among cluster members though at an average frequency of 28%. Previous studies have associated amplification of this genes to the constitutive activation of MAPK/ERK thus contribute to the pathogenesis[143,146–149]. In specific deletion of 1 Mb microdeletion on chromosome 22q11.2 which harbour *MAPK1/ERK2* genes was associated with proliferation and differentiation anomalies due to reduced level of ERK2 protein levels[143,146–149]. Moreover, the reduced ERK2 level was also associated with clinical exhibition of microcephaly, impaired cognition, and developmental delay observed[143,146–149].

4.2.0 Common and unique profiles in identified sub-groups of low grade glioblastoma primary tumour and high grade glioblastoma primary tumour

Chromosome 7p/q amplifications involves *EGFR* gene amplification at chr7p11.2. *EGFR* overexpression is commonly observed in glioblastoma patients and it's a current target in precision medicine approaches[150,151]. Chromosomes 9 deletions involved loss of *CDKN2A/B* at chr9p21.3, *MTAP* gene at chr9p21.3 and *SMARCA2* at chr9p24.3. The alterations of the genes has been previously associated with primary glioma and likely absent in *IDH*-mutation glioblastoma[18]. Combined with chromosomal 10 loss, this alteration would be relevant to distinguish lower grade glioma and high-grade glioma in patients.

Hierarchical clustering of the samples resulted in identification of clusters with distinct copy number alterations. Similar studies also showed reliability of hierarchical clustering in classification of glioma, however they did not use hclust and pearson ward correlations in the analysis as we did[18,152]. I hereby showed that Pearson-ward D2 agglomerative approach as one of the best methods in clustering of tumours based on copy number alterations.

We further noted that C3 LGG- PT (n=60) had similar chromosomal patterns like GBM-PT three groups. Some secondary glioblastoma cases which progress from low grade glioma are often encountered in misdiagnosis and we could not rule out this possibility[153] .. The common alterations observed were chromosomal gain at chromosome 7p/q and chromosomal loss at 10p/q and 9 q. This would be a possible misclassification of these patients into LGG glioma during diagnosis which would have been captured if combined CNV and other parameters were in place. Mis-diagnosis would probably result in mismanagement of patients especially during drug administration therefore different phenotypic makers and approaches would be helpful in glioblastoma management[153]. Following the chromosomal alteration profiles, it is promising to precisely classify patients into subgroups for better management. This would possibly increase the cure rate of the glioblastoma patients.

5. 0 Summary

Brain tumour's range from benign neoplasm such as pilocytic astrocytoma to malignant ones e.g glioblastoma. Histopathological diagnosis of these entities is frequently challenged with inter-observer variability. Moreover, the used genome wide methylation patterns cannot grade tumour severity which is key in patient management. Although specific copy number variation (CNV) profiles such as 1p/19q co-deletions is known to characterise oligodendroglioma and joint gain of chr 7 and loss of chr 10 characterise glioblastoma, other CNV profiles have not been well integrated in brain tumour diagnosis. Therefore, it seems promising to achieve improvements in methylation-based diagnostics and disease prognosis by establishing an approach to systematically include CNV information in classification of brain tumours. With the aim of addressing this issue, **in the first phase** of my study, I evaluated whether methylation data (450K and 850K epic) could inform about the presence of CNVs. I used 61 paired data sets processed from microarray based comparative genomic hybridization (aCGH) and Epic 450K/850K methylation arrays respectively. Copy number plots of the methylation data set were generated from the "conumee" R-package while aCGH data set plots were inferred from the "DNA copy" package. I observed >80% percent agreement between the two methods. To rule out chance agreement and check the extent agreement, I calculated Kappa statistics. I observed moderate (0.54) to substantial (0.61) Kappa statistic values. In conclusion I provided evidence that the methylation data is reliable in determining CNVs.

In the second phase, I evaluated the CNV profiles and survival times using Kaplan Meier analysis between WHO classified astrocytoma grade II and III data (n=117) obtained from the cancer genome atlas (TCGA). Before clustering, I observed no significant difference in survival in WHO grade II and III. After hierarchical clustering (Pearson coefficient correlation ward linkage) using the log2 CNV values, I was able to identify 7 clusters which had different survival rates. The clusters had both unique and shared alteration between them. For example, cluster 4 (n=10) showed better survival with deletions at Chr3q, 4q, 5p/q, 11p, 12q, 13q and gain in Chr12p. These regions carry genes such as *ANO2, CD4, LRRC23, VWF* and *GALNT8* genes. Cluster 3 had poor survival and increased deletions at chr 1q, 2q, 3q, 4q, 5p/q, 6q, 7q, 11p, 13q and chr gain at 9p (n=54). Some key genes altered in these loci, included *C2orf88, CDKN2A/B, RB1, SORBS2, POLD1, MYBPC2* and *TP63*. These genes play critical roles in

cell cycle regulation, growth and tumour suppressions. Cluster 7 had losses at chr 4p/q, 13p/q and 19q (n=8) which contained genes like *LRBA, FBXW7, MARCHF1, SPOCK3, MTUS2* and *RFC3*. Moreover, *CDH12* gene and Long noncoding RNA (*LINC005*) regulating *CCND2* at 5p and 13q respectively were also deleted in >75% of samples. I further noticed that glioblastoma recurrent cases and primary tumor could be differentiated by presence of chr7p/q gain, 9p, 10p/q and 13p/q deletions using a total of n= 1500 cases and n= 1400 controls data set retrieved from TCGA. The 9p and 10p/q loci are already known to encode cell survival and apoptotic genes such as *CDK2A/B, MDM2, EGFR* and *PTEN* which are common in high grade glioma. These results therefore promise better tumour diagnosis and patients stratification approach which would help in both patient management and treatment outcomes predictions by use of CNV profiles.

In the third phase of my study, I evaluated the methylation classes and pathways associated with genes in the altered regions. I observed a different frequency in the distribution of Isocitrate dehydrogenase (*IDH*) mutation and the 06- methylguanine DNA methyltransferase (*MGMT*) in the 7 clusters. In specific clusters 1 and 6 were A_IDH 100% and 70% respectively. A_IDH_HG dominated the other clusters as follows: cluster 5 (50%), cluster 4 (33%), cluster 3 (13%) and cluster 7 (12%). This indicates that methylome classes can be aligned with the CNV profiles. Using ingenuity pathway-based knowledge, I was able to identify canonical pathways associated with altered genes per group. I observed that fairly unique signaling pathways were associated with the disease. Notably, PTEN, ERK/MAPK, P53, IL-3, Glioblastoma multiforme, glioma invasiveness and axonal guidance signaling which are associated with glioma formation are featured in most clusters. Key altered genes included adenomatous polyposis coli which is a tumor suppressor, Glycogen synthase kinase 3 beta (*GSK-3β*) which affects cell proliferation, retinoblastoma (*Rb,*) which codes a tumor suppressor rb protein while Platelet-derived growth factor (*PDGF*) and Phosphoinositide 3-kinases (*P13K*) both regulate cell growth and other cellular functions. Proto-oncogen Rat sarcoma (*Ras*), WNT, Son of Sevenless (*SOS*), Auditory processing deficit (APD) and beta catenin (*CTNNB1*) were also lost. The WNT pathway activation aids in cellular differentiation which promotes brain tumour formation while Ras /PI3K/RTK pathway contributes to tumour growth deregulation. These findings show that multiple pathways dysregulated by CNVs can help in establishing novel brain tumour stratification, diagnostics and consequently identification of novel drug targets.

6.0 Zusammenfassung

Hirntumore reichen von gutartigen Neubildungen wie dem pilozytischen Astrozytom bis hin zu bösartigen Neubildungen wie dem Glioblastom. Die histopathologische Diagnose dieser Entitäten wird häufig durch die Variabilität zwischen den Beobachtern erschwert. Darüber hinaus können die verwendeten genomweiten Methylierungsmuster nicht den Schweregrad des Tumors einstufen, der für das Patientenmanagement entscheidend ist. Obwohl spezifische Kopienzahlvariationen (CNV), wie z.B. 1p/19q Co-Deletionen, bekannt sind, um das Oligodendrogliom zu charakterisieren, und der gemeinsame Gewinn von chr 7 und der Verlust von chr 10 das Glioblastom kennzeichnen, sind andere CNV-Profile nicht gut in die Hirntumordiagnose integriert worden. Daher scheint es vielversprechend, Verbesserungen in der methylierungsbasierten Diagnostik und Krankheitsprognose zu erreichen, indem ein Ansatz zur systematischen Einbeziehung von CNV-Informationen in die Klassifikation von Hirntumoren etabliert wird. Auf dieses Ziel habe ich hingearbeitet. In der ersten Phase meiner Studie evaluierte ich, ob Methylierungsdaten (450K und 850K epic) über das Vorhandensein von CNV informieren können. Ich verwendete 61 gepaarte Datensätze, die aus der Mikroarray-basierten vergleichenden genomischen Hybridisierung (aCGH) bzw. dem Epic 450K/850K Methylierungsarray verarbeitet wurden. Die Kopienzahlplots der Methylierungsdatensätze wurden aus dem "conumee" R-Paket in bioconductor mit kleinen Modifikationen generiert. Für den aCGH-Datensatz wurden CNV-Plots aus dem "DNA copy"-Paket abgeleitet. Ich beobachtete >80% prozentuale Übereinstimmung zwischen den beiden Methoden. Um eine zufällige Übereinstimmung auszuschließen und das Ausmaß der Übereinstimmung zu überprüfen, habe ich die Kappa-Statistik berechnet. Ich beobachtete mäßige (0,54) bis erhebliche (0,61) Kappa-Statistikwerte. Zusammenfassend habe ich den Nachweis erbracht, dass die Methylierungsdaten bei der Bestimmung von CNV zuverlässig sind.

In der zweiten Phase bewertete ich die CNV-Profile und Überlebenszeiten mittels Kaplan-Meier-Analyse zwischen den WHO-klassifizierten Astrozytomen Grad II und III der Krebsgenom-Atlas (TCGA)-Daten (n=117). Vor dem Clustering beobachtete ich keinen signifikanten Unterschied im Überleben zwischen WHO-Grad II und III. Nach dem hierarchischen Clustering (Pearson coefficient correlation ward linkage) unter Verwendung der log2 CNV-Werte konnte ich 7 Cluster/Untergruppen identifizieren, die ein unterschiedliches

Überleben hatten. Die Cluster hatten sowohl einzigartige als auch gemeinsame Veränderungen zwischen ihnen. Zum Beispiel zeigte Cluster 4 (n=10) ein besseres Überleben mit Deletionen an 3q,4q,5p/q,11p,12q,13q und Gain in chr 12p. Diese Regionen tragen Gene wie ANO2, CD4, LRRC23, VWF und GALNT8 Gene. Cluster 3 hatte ein schlechtes Überleben und erhöhte Deletionen in chr 1q,2q,3q,4q,5p/q,6q,7q,11p,13q und Gain in chr 9p (n=54). Einige Schlüsselgene, die an diesen Loci verändert waren, waren C2orf88, CDKN2A/B, RB1, SORBS2, POLD1, MYBPC2 und TP63. Diese Gene spielen eine entscheidende Rolle bei der Zellzyklusregulation, dem Wachstum und der Tumorsuppression. Cluster 7 hatte Verlust an chr 4p/q, 13p/q und 19q (n=8), die Gene wie LRBA, FBXW7, MARCHF1, SPOCK3, MTUS2 und RFC3 enthielten. Darüber hinaus waren auch das CDH12-Gen und die lange nicht-kodierende RNA (LINC005), die CCND2 auf 5p bzw. 13q reguliert, in >75% der Proben deletiert. Des Weiteren konnte ich feststellen, dass Glioblastom-Rezidivfälle und Primärtumor durch das Vorhandensein von chr7p/q gain, 9p, 10p/q und 13p/q-Deletionen unterschieden werden konnten, wobei ich insgesamt n= 1500 Fälle und n= 1400 Kontrollen aus dem TCGA-Datensatz verwendete. Die 9p- und 10p/q-Loci sind bereits dafür bekannt, dass sie für Zellüberlebens- und apoptotische Gene wie CDK2A/B, MDM2, EGFR und PTEN kodieren, die bei hochgradigen Gliomen häufig vorkommen. Diese Ergebnisse versprechen daher eine bessere Tumordiagnose und einen Ansatz zur Patientenstratifizierung, der sowohl beim Patientenmanagement als auch bei der Vorhersage von Behandlungsergebnissen durch die Verwendung von CNV-Profilen helfen würde.

In der dritten Phase meiner Studie untersuchte ich die Methylierungsklassen und -pfade, die mit Genen in den veränderten Regionen assoziiert sind. Ich beobachtete eine unterschiedliche Häufigkeit in der Verteilung der Isocitrat-Dehydrogenase (IDH)-Mutation und der 06-Methyl-Guanin-DNA-Methyl-Transferase (MGMT) in den 7 Untergruppen. In den spezifischen Clustern 1 und 6 waren A_IDH 100% bzw. 70%. In den anderen Clustern dominierte A_IDH_HG wie folgt: Cluster 5 (50%), Cluster 4 (33%), Cluster 3 (13%) und Cluster 7 (12%). Dies zeigt, dass Methylom-Klassen mit den CNV-Profilen abgeglichen werden können. Mit Hilfe von ingenuity pathway-based knowledge konnte ich kanonische Signalwege identifizieren, die mit veränderten Genen pro Gruppe assoziiert sind. Ich beobachtete, dass ziemlich einzigartige Signalwege mit der Krankheit assoziiert waren. Insbesondere PTEN, ERK/MAPK, P53, IL-3, Glioblastoma multiforme, Gliom-Invasivität und Axonal Guidance Signaling, die mit der Gliombildung assoziiert sind, traten in den meisten Gruppen auf. Zu den wichtigsten veränderten Genen gehörten die adenomatöse Polyposis coli, die ein

Tumorsuppressor ist, die Glykogensynthasekinase 3 beta (GSK-3β), die die Zellproliferation beeinflusst, das Retinoblastom (Rb,), das ein Tumorsuppressor-Rb-Protein kodiert, während der Platelet-derived growth factor (PDGF) und die Phosphoinositid-3-Kinasen (P13K) beide das Zellwachstum und andere zelluläre Funktionen regulieren. Proto-Onkogen Ratten-Sarkom (Ras), WNT, Son of Sevenless (SOS), Auditory processing deficit und beta-Catenin wurden ebenfalls verloren. Die Aktivierung des WNT-Signalwegs hilft bei der zellulären Differenzierung, die die Bildung von Hirntumoren fördert, während der Ras /PI3K/RTK-Signalweg zur Deregulation des Tumorwachstums beiträgt. Diese Ergebnisse zeigen, dass mehrere durch CNV dysregulierte Signalwege bei der Etablierung einer neuartigen Stratifizierung von Hirntumoren, bei der Diagnostik und folglich bei der Identifizierung neuer medikamentöser Angriffspunkte helfen können.

7.0 References

1.   Veldhuijzen, D. S., Greenspan, J. D., Kim, J. H. & Lenz, F. A. Altered pain and thermal sensation in subjects with isolated parietal and insular cortical lesions. *Eur. J. Pain* **14**, 535.e1-535.11 (2010).

2.   Aldape, K. *et al.* Challenges to curing primary brain tumours. *Nat. Rev. Clin. Oncol.* **16**, 509–520 (2019).

3.   Zong, H., Parada, L. F. & Baker, S. J. Cell of origin for malignant gliomas and its implication in therapeutic development. *Cold Spring Harb. Perspect. Biol.* **7**, 1–13 (2015).

4.   -Galvan *et al.* Molecular diagnostics and pathology of Major brain tumour. *Intech* **i**, 13 (2016).

5.   Maher, E. A. & Bachoo, R. M. *Glioblastoma. Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease: Fifth Edition* (2014). doi:10.1016/B978-0-12-410529-4.00078-4

6.   Wulczyn, E. *et al.* Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS One* **15**, (2020).

7.   Olar, A. & Aldape, K. D. Using the molecular classification of glioblastoma to inform personalized treatment. *J. Pathol.* **232**, 165–177 (2014).

8.   Ryall, S., Tabori, U. & Hawkins, C. Pediatric low-grade glioma in the era of molecular diagnostics. *Acta Neuropathol. Commun.* **8**, 30 (2020).

9.   Fults, D., Brockmeyer, D., Tullous, M. W. & Pedone, C. A. p53 Mutation and Loss of Heterozygosity on Chromosomes 17 and 10 during Human Astrocytoma Progression. *Cancer Res.* **52**, 674–679 (1992).

10.   Ryland, G. L. *et al.* Loss of heterozygosity: what is it good for? *BMC Med. Genomics* **8**, 45 (2015).

11.   Brown, K. W. & Malik, K. T. A. The molecular biology of Wilms' tumour. *Expert Rev. Mol. Med.* **3**, 1–16 (2001).

12.   Bignell, G. R. *et al.* High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* **14**, 287–295 (2004).

13.   Rodriguez, F. J., Vizcaino, M. A. & Lin, M. T. Recent Advances on the Molecular Pathology of Glial Neoplasms in Children and Adults. *J. Mol. Diagnostics* **18**, 620–634 (2016).

14.   Molenaar, R. J., Maciejewski, J. P., Wilmink, J. W. & van Noorden, C. J. F. Wild-type and mutated IDH1/2 enzymes and therapy responses. *Oncogene* **37**, 1949–1960 (2018).

15.   Monga, V., Jones, K. & Chang, S. Clinical Relevance of Molecular Markers in Gliomas. *Rev. Médica Clínica Las Condes* **28**, 343–351 (2017).

16.   Clark, O., Yen, K. & Mellinghoff, I. K. Molecular Pathways: Isocitrate Dehydrogenase Mutations in Cancer. *Clin. Cancer Res.* **22**, 1837 LP – 1842 (2016).

17.   Cohen, A. *et al.* DNA copy number analysis of Grade II-III and Grade IV gliomas reveals differences in molecular ontogeny including chromothripsis associated with IDH mutation status. *Acta Neuropathol. Commun.* **3**, 34 (2015).

18.   Cohen, A. L., Holmen, S. L. & Colman, H. IDH1 and IDH2 Mutations in Gliomas. *Curr. Neurol. Neurosci. Rep.* **13**, 345 (2013).

19.   Pietsch, E. C., Sykes, S. M., McMahon, S. B. & Murphy, M. E. The p53 family and programmed cell death. *Oncogene* **27**, 6507–6521 (2008).

20.   Rivlin, N., Brosh, R., Oren, M. & Rotter, V. Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis. *Genes Cancer* **2**, 466–474 (2011).

21.   Zhang, X. & Sjöblom, T. Targeting Loss of Heterozygosity: A Novel Paradigm for Cancer Therapy. *Pharmaceuticals (Basel).* **14**, 57 (2021).

22.   Guan, Y., Li, Y. & Li, Y. The change of MGMT gene expression in glioma patients was affected by methylation regulation and in the treatment of alkylation agent. *Int. J. Clin. Exp. Med.* **12**, 8725–8731 (2019).

23.   Xu, H. *et al.* Epidermal growth factor receptor in glioblastoma (Review). *Oncol. Lett.* **14**, 512–516 (2017).

24.   Wee, P. & Wang, Z. Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. *Cancers (Basel).* **9**, 52 (2017).

25.   Karpel-Massler, G., Wirtz, C. R. & Halatsch, M. E. Ribozyme-mediated inhibition of 801-bp deletion-mutant epidermal growth factor receptor mrna expression in glioblastoma multiforme. *Molecules* **15**, 4670–4678 (2010).

26.   Pfenninger, C. V *et al.* CD133 Is Not Present on Neurogenic Astrocytes in the Adult Subventricular Zone, but on Embryonic Neural Stem Cells, Ependymal Cells, and Glioblastoma Cells. *Cancer Res.* **67**, 5727 LP – 5736 (2007).

27.   Ahmed, S. I. *et al.* CD133 Expression in Glioblastoma Multiforme: A Literature Review. *Cureus* **10**, (2018).

28. Ma, Y.-H. *et al.* Expression of stem cell markers in human astrocytomas of different WHO grades. *J. Neurooncol.* **86**, 31–45 (2008).

29. Pallini, R. *et al.* Cancer Stem Cell Analysis and Clinical Outcome in Patients with Glioblastoma Multiforme. *Clin. Cancer Res.* **14**, 8205 LP – 8212 (2008).

30. Zeppernick, F. *et al.* Stem Cell Marker CD133 Affects Clinical Outcome in Glioma Patients. *Clin. Cancer Res.* **14**, 123 LP – 129 (2008).

31. Wang, J. *et al.* CD133 negative glioma cells form tumors in nude rats and give rise to CD133 positive cells. *Int. J. Cancer* **122**, 761–768 (2008).

32. Galli, R. *et al.* Isolation and Characterization of Tumorigenic, Stem-like Neural Precursors from Human Glioblastoma. *Cancer Res.* **64**, 7011 LP – 7021 (2004).

33. Ehrmann, J., Kolář, Z. & Mokrý, J. Nestin as a diagnostic and prognostic marker: immunohistochemical analysis of its expression in different tumours. *J. Clin. Pathol.* **58**, 222 LP – 223 (2005).

34. Chinnaiyan, P. *et al.* The prognostic value of nestin expression in newly diagnosed glioblastoma: Report from the Radiation Therapy Oncology Group. *Radiat. Oncol.* **3**, 32 (2008).

35. Dahlstrand, J., Collins, V. P. & Lendahl, U. Expression of the Class VI Intermediate Filament Nestin in Human Central Nervous System Tumors. *Cancer Res.* **52**, 5334 LP – 5341 (1992).

36. Yang, X. H. *et al.* Nestin expression in different tumours and its relevance to malignant grade. *J. Clin. Pathol.* **61**, 467 LP – 473 (2008).

37. Dahlrot, R. H., Hermansen, S. K., Hansen, S. & Kristensen, B. W. 2013 - what is the clinical value of cancer stem cell markers in gliomas - int j clin exp pathol - Dahlrot.pdf. **6**, 334–348 (2013).

38. Shibahara, J., Kashima, T., Kikuchi, Y., Kunita, A. & Fukayama, M. Podoplanin is expressed in subsets of tumors of the central nervous system. *Virchows Arch.* **448**, 493–499 (2006).

39. Ogasawara, S., Kaneko, M. K., Price, J. E. & Kato, Y. Characterization of Anti-podoplanin Monoclonal Antibodies: Critical Epitopes for Neutralizing the Interaction Between Podoplanin and CLEC-2. *Hybridoma* **27**, 259–267 (2008).

40. Mishima, K. *et al.* Increased expression of podoplanin in malignant astrocytic tumors as a novel molecular marker of malignant progression. *Acta Neuropathol.* **111**, 483 (2006).

41. Ordóñez, N. G. Podoplanin: A Novel Diagnostic Immunohistochemical Marker. *Adv. Anat. Pathol.* **13**, (2006).

42.   Son, M. J., Woolard, K., Nam, D.-H., Lee, J. & Fine, H. A. SSEA-1 Is an Enrichment Marker for Tumor-Initiating Cells in Human Glioblastoma. *Cell Stem Cell* **4**, 440–452 (2009).

43.   Kenney-Herbert, E. *et al.* CD15 Expression Does Not Identify a Phenotypically or Genetically Distinct Glioblastoma Population. *Stem Cells Transl. Med.* **4**, 822–831 (2015).

44.   Nunes, M. C. *et al.* Identification and isolation of multipotential neural progenitor cells from the subcortical white matter of the adult human brain. *Nat. Med.* **9**, 439–447 (2003).

45.   Colin, C. *et al.* In vitro identification and functional characterization of glial precursor cells in human gliomas. *Neuropathol. Appl. Neurobiol.* **32**, 189–202 (2006).

46.   Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, (2019).

47.   Bejjani, B. A. & Shaffer, L. G. Application of array-based comparative genomic hybridization to clinical diagnostics. *Journal of Molecular Diagnostics* **8**, (2006).

48.   Patel, A. & Cheung, S. W. Application of DNA microarray to clinical diagnostics. in *Methods in Molecular Biology* **1368**, (2016).

49.   Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* (1998). doi:10.1038/2524

50.   Capper, D. *et al.* Practical implementation of DNA methylation and copy-number-based CNS tumor diagnostics: the Heidelberg experience. *Acta Neuropathol.* (2018). doi:10.1007/s00401-018-1879-y

51.   Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).

52.   James, L. A. Comparative genomic hybridization as a tool in tumour cytogenetics. *Journal of Pathology* (1999). doi:10.1002/(SICI)1096-9896(199903)187:4<385::AID-PATH290>3.0.CO;2-5

53.   Haraksingh, R. R., Abyzov, A. & Urban, A. E. Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. *BMC Genomics* **18**, (2017).

54.   Chen, J. & Wang, Y. P. A statistical change point model approach for the detection of DNA copy number variations in array CGH data. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* (2009). doi:10.1109/TCBB.2008.129

55.   Hupé, P., Stransky, N., Thiery, J. P., Radvanyi, F. & Barillot, E. Analysis of array CGH data: From signal ratio to gain and loss of DNA regions. *Bioinformatics* (2004). doi:10.1093/bioinformatics/bth418

56.   Solomon, O. *et al.* Comparison of DNA methylation measured by Illumina 450K and EPIC BeadChips in blood of newborns and 14-year-old children. *Epigenetics* **13**, (2018).

57.   Kilaru, V. *et al.* Critical evaluation of copy number variant calling methods using DNA methylation. *Genet. Epidemiol.* **44**, (2020).

58.   McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, (2008).

59.   Brat, D. J. *et al.* Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* **372**, (2015).

60.   Snijders, A. M. *et al.* Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.* (2001). doi:10.1038/ng754

61.   Bishop, R. Applications of fluorescence in situ hybridization (FISH) in detecting genetic aberrations of medical significance. *Biosci. Horizons Int. J. Student Res.* **3**, 85–95 (2010).

62.   McNamara, M. G., Sahebjam, S. & Mason, W. P. Emerging Biomarkers in Glioblastoma. *Cancers (Basel).* **5**, 1103–1119 (2013).

63.   Chen, C.-P. *et al.* Application of interphase fluorescence in situ hybridization to uncultured amniocytes for differential diagnosis of pseudomosaicism from true mosaicism in mosaic isochromosome 20q detected at amniocentesis. *Taiwan. J. Obstet. Gynecol.* **52**, 450–453 (2013).

64.   Korshunov, A., Sycheva, R., Gorelyshev, S. & Golanov, A. Clinical utility of fluorescence in situ hybridization (FISH) in nonbrainstem glioblastomas of childhood. *Mod. Pathol.* **18**, 1258–1263 (2005).

65.   Ziegler, A., Koch, A., Krockenberger, K. & Großhennig, A. Personalized medicine using DNA biomarkers: a review. *Hum. Genet.* **131**, 1627–1638 (2012).

66.   Tanboon, J., Williams, E. A. & Louis, D. N. The Diagnostic Use of Immunohistochemical Surrogates for Signature Molecular Genetic Alterations in Gliomas. *J. Neuropathol. Exp. Neurol.* **75**, 4–18 (2016).

67.   Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science (80-.).* **321**, 1807–1812 (2008).

68.   Horbinski, C., Kofler, J., Kelly, L. M., Murdoch, G. H. & Nikiforova, M. N. Diagnostic Use of IDH1/2 Mutation Analysis in Routine Clinical Testing of Formalin-Fixed, Paraffin-Embedded Glioma Tissues. *J. Neuropathol. Exp. Neurol.* **68**, 1319–1325 (2009).

69.   Reuss, D. E. *et al.* IDH mutant diffuse and anaplastic astrocytomas have similar age at presentation and little difference in survival: a grading problem for WHO. *Acta Neuropathol.* (2015). doi:10.1007/s00401-015-1438-8

70.   Gondim, D. D. *et al.* Determining IDH-Mutational Status in Gliomas Using IDH1-R132H Antibody and Polymerase Chain Reaction. *Appl. Immunohistochem. Mol. Morphol.* **27**, (2019).

71.   Biterge, B. & Schneider, R. Histone variants: key players of chromatin. *Cell Tissue Res.* **356**, 457–466 (2014).

72.   Takami, H. *et al.* Revisiting TP53 Mutations and Immunohistochemistry—A Comparative Study in 157 Diffuse Gliomas. *Brain Pathol.* **25**, 256–265 (2015).

73.   Clynes, D., Higgs, D. R. & Gibbons, R. J. The chromatin remodeller ATRX: a repeat offender in human disease. *Trends Biochem. Sci.* **38**, 461–466 (2013).

74.   Cuykendall, T. N., Rubin, M. A. & Khurana, E. Non-coding genetic variation in cancer. *Curr. Opin. Syst. Biol.* **1**, 9–15 (2017).

75.   Kim, H. *et al.* Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Res.* **25**, 316–327 (2015).

76.   Körber, V. *et al.* Evolutionary Trajectories of IDH^{WT} Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis. *Cancer Cell* **35**, 692-704.e12 (2019).

77.   The RNAcentral Consortium. RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res.* **47**, D221–D229 (2019).

78.   Vlassenbroeck, I. *et al.* Validation of real-time methylation-specific PCR to determine O 6-methylguanine-DNA methyltransferase gene promoter methylation in glioma. *J. Mol. Diagnostics* (2008). doi:10.2353/jmoldx.2008.070169

79.   Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathologica* (2016). doi:10.1007/s00401-016-1545-1

80.   Masui, K., Mischel, P. S. & Reifenberger, G. Molecular classification of gliomas. in *Handbook of Clinical Neurology* (2016). doi:10.1016/B978-0-12-802997-8.00006-2

81.   Fernandez, A. F. *et al.* A DNA methylation fingerprint of 1628 human samples. *Genome Res.* **22**, 407–419 (2012).

82.   Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, (2016).

83.   Karimi, S. *et al.* The central nervous system tumor methylation classifier changes neuro-oncology practice for challenging brain tumor diagnoses and directly impacts patient care. *Clin. Epigenetics* **11**, 1–10 (2019).

84.   Jaunmuktane, Z. *et al.* Methylation array profiling of adult brain tumours: diagnostic outcomes in a large, single centre. *Acta Neuropathol. Commun.* **7**, 24 (2019).

85.   Pajtler, K. W. *et al.* Molecular Classification of Ependymal Tumors across All CNS Compartments, Histopathological Grades, and Age Groups. *Cancer Cell* (2015). doi:10.1016/j.ccell.2015.04.002

86.   Capper, D. *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* (2018). doi:10.1038/nature26000

87.   Wesseling, P., Kros, J. M. & Jeuken, J. W. M. The pathological diagnosis of diffuse gliomas: Towards a smart synthesis of microscopic and molecular information in a multidisciplinary context. *Diagnostic Histopathol.* **17**, 486–494 (2011).

88.   MacDonald, T. J. *et al.* Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. *Nat. Genet.* **29**, 143–152 (2001).

89.   Kleihues, P. & Ohgaki, H. Primary and secondary glioblastomas: From concept to clinical diagnosis. *Neuro. Oncol.* **1**, 44–51 (1999).

90.   Giannini, C. *et al.* Oligodendrogliomas: Reproducibility and prognostic value of histologic diagnosis and grading. *J. Neuropathol. Exp. Neurol.* **60**, (2001).

91.   Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* (2016). doi:10.1016/j.jcm.2016.02.012

92.   Boon, K., Edwards, J. B., Eberhart, C. G. & Riggins, G. J. Identification of astrocytoma associated genes including cell surface markers. *BMC Cancer* **4**, 1–8 (2004).

93.   Mao, C. X. *et al.* The molecular classification of astrocytic tumors. *Oncotarget* **8**, 96340–96350 (2017).

94.   Xavier-Magalhães, A., Nandhabalan, M., Jones, C. & Costa, B. M. Molecular prognostic factors in glioblastoma: state of the art and future challenges. *CNS Oncol.* **2**, 495–510 (2013).

95.   Verhaak, R. G. W. *et al.* Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).

96.   Churpek, J. E. *et al.* Identification and molecular characterization of a novel 3′ mutation in RUNX1 in a family with familial platelet disorder. *Leukemia and Lymphoma* **51**, 1931–1935 (2010).

97.   Seshan, V. E. & Olshen, A. B. DNAcopy : A Package for Analyzing DNA Copy Data. *Bioconductor Vignette* 1–7 (2014).

98.   Snijders, A. M. *et al.* Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.* **29**, 263–264 (2001).

99.   Feber, A. *et al.* Using high-density DNA methylation arrays to profile copy number alterations. *Genome Biol.* **15**, R30–R30 (2014).

100.   Aryee, M. J. *et al.* Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, (2014).

101.   Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **9**, (2013).

102.   McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. medica* (2012).

103.   Bewick, V., Cheek, L., Ball, J. & Meier, K. Statistics review 12 : Survival analysis Estimating the survival curve using the. (2004). doi:10.1186/cc2955

104.   Kishore, J., Goel, M. & Khanna, P. Understanding survival analysis: Kaplan-Meier estimate. *Int. J. Ayurveda Res.* **1**, (2010).

105.   Layton, D. M. Understanding Kaplan-Meier and Survival Statistics. *Int. J. Prosthodont.* **26**, (2013).

106.   Meyniel, J. P. *et al.* A genomic and transcriptomic approach for a differential diagnosis between primary and secondary ovarian carcinomas in patients with a previous history of breast cancer. *BMC Cancer* **10**, (2010).

107.   Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature Reviews Genetics* **7**, (2006).

108.   Analysis, I. P. Ingenuity Pathway Analysis Knowledge Base. *Website* 5020 (2015).

109.   Watson, P. F. & Petrie, A. Method agreement analysis: A review of correct methodology. *Theriogenology* **73**, 1167–1179 (2010).

110.   Oskam, I. C. *et al.* Testicular germ cell development in relation to 5α-Androstenone levels in pubertal entire male pigs. *Theriogenology* **69**, 967–976 (2008).

111.   Coughlin, C. R., Scharer, G. H. & Shaikh, T. H. Clinical impact of copy number variation analysis using high-resolution microarray technologies: advantages, limitations and concerns. *Genome Medicine* (2012). doi:10.1186/gm381

112. Louis, D. N. *et al.* The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* **114**, 97–109 (2007).

113. Diamandis, P. & Aldape, K. World Health Organization 2016 Classification of Central Nervous System Tumors. *Neurologic Clinics* (2018). doi:10.1016/j.ncl.2018.04.003

114. Louis, D. N. *et al.* The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathologica* (2007). doi:10.1007/s00401-007-0243-4

115. Mirchia, K. *et al.* Total copy number variation as a prognostic factor in adult astrocytoma subtypes. *Acta Neuropathol. Commun.* **7**, 8 (2019).

116. Hannou, S. A., Wouters, K., Paumelle, R. & Staels, B. Functional genomics of the CDKN2A/B locus in cardiovascular and metabolic disease: What have we learned from GWASs? *Trends in Endocrinology and Metabolism* **26**, (2015).

117. ShahidSales, S. *et al.* A genetic variant in CDKN2A/B gene is associated with the increased risk of breast cancer. *J. Clin. Lab. Anal.* **32**, (2018).

118. Nobori, T. *et al.* Deletions of the cyclin-dependent kinase-4 inhibitor gene in multiple human cancers. *Nature* **368**, 753–756 (1994).

119. Ruas, M. & Peters, G. The p16(INK4a)/CDKN2A tumor suppressor and its relatives. *Biochimica et Biophysica Acta - Reviews on Cancer* **1378**, (1998).

120. Reis, G. F. *et al.* CDKN2A Loss Is Associated with Shortened Overall Survival in Lower-Grade (World Health Organization Grades II-III) Astrocytomas. *J. Neuropathol. Exp. Neurol.* **74**, (2015).

121. Weinberg, R. A. The retinoblastoma protein and cell cycle control. *Cell* **81**, (1995).

122. Goodrich, D. W. The retinoblastoma tumor-suppressor gene, the exception that proves the rule. *Oncogene* **25**, 5233–5243 (2006).

123. Giacinti, C. & Giordano, A. RB and cell cycle progression. *Oncogene* **25**, 5220–5227 (2006).

124. Chinnam, M. & Goodrich, D. W. *RB1, Development, and Cancer*. *Current Topics in Developmental Biology* **94**, (2011).

125. Indovina, P., Pentimalli, F., Casini, N., Vocca, I. & Giordano, A. RB1 dual role in proliferation and apoptosis: Cell fate control and implications for cancer therapy. *Oncotarget* **6**, (2015).

126. Dyson, N. J. RB1: A prototype tumor suppressor and an enigma. *Genes and Development* **30**, (2016).

127. Tatevossian, R. G. *et al.* MYB upregulation and genetic aberrations in a subset of pediatric low-grade gliomas. *Acta Neuropathol.* (2010). doi:10.1007/s00401-010-0763-1

128.   Attard, T. M., Giglio, P., Koppula, S., Snyder, C. & Lynch, H. T. Brain tumors in individuals with Familial Adenomatous Polyposis: A cancer registry experience and pooled case report analysis. *Cancer* **109**, 761–766 (2007).

129.   Kotliarova, S. *et al.* Glycogen synthase kinase-3 inhibition induces glioma cell death through c-MYC, nuclear factor-kappaB, and glucose regulation. *Cancer Res.* **68**, 6643–6651 (2008).

130.   Sareddy, G. R., Panigrahi, M., Challa, S., Mahadevan, A. & Babu, P. P. Activation of Wnt/β-catenin/Tcf signaling pathway in human astrocytomas. *Neurochem. Int.* **55**, 307–317 (2009).

131.   Mills, C. N., Nowsheen, S., Bonner, J. A. & Yang, E. S. Emerging Roles of Glycogen Synthase Kinase 3 in the Treatment of Brain Tumors. *Front. Mol. Neurosci.* **4**, 1–8 (2011).

132.   Xiong, J. *et al.* Mature BDNF promotes the growth of glioma cells in vitro. *Oncol. Rep.* **30**, 2719–2724 (2013).

133.   Buchman, A. S. *et al.* Higher brain BDNF gene expression is associated with slower cognitive decline in older adults. *Neurology* **86**, 735–741 (2016).

134.   He, Y. *et al.* Liprin alfa 2 gene expression is increased by cannabis use and associated with neuropsychological function. *Eur. Neuropsychopharmacol.* **29**, (2019).

135.   Chédotal, A., Kerjan, G. & Moreau-Fauvarque, C. The brain within the tumor: New roles for axon guidance molecules in cancers. *Cell Death Differ.* **12**, 1044–1056 (2005).

136.   Surawska, H., Ma, P. C. & Salgia, R. The role of ephrins and Eph receptors in cancer. *Cytokine Growth Factor Rev.* **15**, 419–433 (2004).

137.   Kolodkin, A. L., Matthes, D. J. & Goodman, C. S. The semaphorin genes encode a family of transmembrane and secreted growth cone guidance molecules. *Cell* **75**, 1389–1399 (1993).

138.   Hanahan, D. & Folkman, J. Patterns and Emerging Mechanisms of the Angiogenic Switch during Tumorigenesis. *Cell* **86**, 353–364 (1996).

139.   Jones, S. *et al.* Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene. *Science* **324**, 217 (2009).

140.   Ji, R.-R., Gereau, R. W., Malcangio, M. & Strichartz, G. R. MAP kinase and pain. *Brain Res. Rev.* **60**, 135–148 (2009).

141.   Jones, D. T. W., Gronych, J., Lichter, P., Witt, O. & Pfister, S. M. MAPK pathway activation in pilocytic astrocytoma. doi:10.1007/s00018-011-0898-9

142. Samuels, I. S. *et al.* Development/Plasticity/Repair Deletion of ERK2 Mitogen-Activated Protein Kinase Identifies Its Key Roles in Cortical Neurogenesis and Cognitive Function. (2008). doi:10.1523/JNEUROSCI.0679-08.2008

143. David Sweatt, J. The neuronal MAP kinase cascade: A biochemical signal integration system subserving synaptic plasticity and memory. *Journal of Neurochemistry* **76**, 1–10 (2001).

144. Sato, T. & Nakamura, H. The Fgf8 signal causes cerebellar differentiation by activating the Ras-ERK signaling pathway. *Development* **131**, 4275–4285 (2004).

145. Furnari, F. B. *et al.* Malignant astrocytic glioma: genetics, biology, and paths to treatment. (2007). doi:10.1101/gad.1596707

146. Ichimura, K., Ohgaki, H., Kleihues, P. & Collins, V. P. Molecular pathogenesis of astrocytic tumours. *J. Neurooncol.* **70**, 137–160 (2004).

147. Ae, J. J. *et al.* RAS/RAF pathway activation in gliomas: the result of copy number gains rather than activating mutations. doi:10.1007/s00401-007-0239-0

148. Network, C. G. A. R. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).

149. French, P. J. *et al.* Defining EGFR amplification status for clinical trial inclusion. *Neuro. Oncol.* **21**, 1263–1272 (2019).

150. Lopez-Gines, C. *et al.* New pattern of EGFR amplification in glioblastoma and the relationship of gene copy number with gene expression profile. *Mod. Pathol.* **23**, 856–865 (2010).

151. Brennan, C. *et al.* Glioblastoma subclasses can be defined by activity among signal transduction pathways and associated genomic alterations. *PLoS One* **4**, (2009).

152. Liu, C., Xu, W., Liu, P. & Wei, Y. A Mistaken Diagnosis of Secondary Glioblastoma as Parasitosis. *Front. Neurol.* **10**, 952 (2019).

153. Du, J. *et al.* Identification of Prognostic Model and Biomarkers for Cancer Stem Cell Characteristics in Glioblastoma by Network Analysis of Multi-Omics Data and Stemness Indices. *Front. Cell Dev. Biol.* **8**, 1116 (2020).

8.0 List of manuscripts and publications

1. Ngalah Bidii Stephen, Prashanth Suravajhala and Obul Reddy Bandapalli (2021).Copy number variation profiles as a potential tumor grading marker and risk factor predictor across WHO Astrocytoma grades II and III .(**Manuscript written**).

2. Ngalah Bidii Stephen , Dhatri Madduru, Pranathi Pappu, Urvashi Vijay, Prashanth Suravajhala and Obul Reddy Bandapalli (2021). Hepato-carcinogenesis and the role of next generation sequencing in liver cancer.(**Book chapter under review**)

3. Ngalah Bidii Stephen, Corinna Herz and Evelyn Lamy (2021). Anti-inflammatory activity of non-psychoactive constituents of *cannabis sativa* and other plants compounds in cancer prevention and treatment with a focus on cyclooxygenase inhibition.(**written review**)

4. Ngalah Bidii Stephen and Obul Reddy Bandapalli (2020). An overview of clinical significant germline variants associated with haematological cancers (**written review**)

5. Ngalah Bidii Stephen, Daniel Schrimpf, Andrey Korshunov, Stefan Pusch, Andreas von Deimling et al., Validation of DNA copy number variations using Methylation and Comparative genomic hybridization microarrays data (**Manuscript written-unpublished**)

Published manuscripts in peer reviewed Journals

1. Kayla M hardwick,., Bidii Stephen Ngalah , Dawn E Gunderson-Rindal , Sarah Schaack et al.,2020. Comprehensive transcriptome of the maize stalk borer, Busseola fusca, from multiple tissue types,developmental stages, and parasitoid wasp exposure.Genome Biol Evol 2020 Sep 18;evaa195. doi: 10.1093/gbe/evaa195.

2. Peninah Muiruri, Denis W. Juma, Luicer A. Ingasia, Lorna J. Chebon, Benjamin Opot, Bidii S. Ngalah, Jelagat Cheruiyot, Ben Andagalu, Hoseah M. Akala, Venny C. S. Nyambati, Joseph K. Ng'ang'a,and Edwin Kamau.Selective sweeps and   genetic lineages of *Plasmodium falciparum* multi-drug resistance (*pfmdr1*) gene in Kenya. Malar J . 2018; 17: 398.doi: 10.1 186/s12936-018-2534-8

3.Bidii S. N, Luiser A. I, Zipporah N, Edwin K et al., Analysis of Major Genome Loci Underlying Artemisinin Resistance and pfmdr1 Copy Number in pre- and post-ACTs in Western Kenya. Sci. Rep. 5, 8308; DOI: 10.1038/srep08308 (2015).

4.Lorna J. C, Bidii S. N, Luiser A. Edwin K et al., Genetically determined response to Artemisinin Treatment in western Kenya, plasmodium falciparum parasite (submitted to Scientific reports journal Jan 2015).

5. Angela O.Achieng, Penina Muiruri, Bidii S. Ngalah, Bernhards R. Ogutu, Edwin Kamau et al.,.Temporal trends in prevalence of Plasmodium falciparum molecular markers selected for by artemether-lumefantrine treatment in pre- ACT and post-ACT parasites in Western Kenya. International Journal for Parasitology: Drugs and Drug Resistance DOI: 1 0.1016/j.ijpddr.2015.05.005.

6. Cheruiyot, J, Dennis W. Bidii S. N, Wallace D. Edwin K.et al., Polymorphisms in Pfmdr1, Pfcrt, and Pfnhe1 genes are associated with reduced in vitro activities of quinine in *Plasmodium falciparum* isolates from western Kenya. *Antimicrob Agents Chemother* 58, 3737-43 (2014).

7. Dennis W. J, Angela A. O, Luiser I, Bidii S. N, Lorna J.C, Fredrick E, Jacob J, Edwin K. et al., Trends in drug resistance codons in *Plasmodium falciparum* dihydrofolate reductase and dihydropteroate synthase genes in Kenyan parasites from 2008 to 2012.*Malaria Journal 2014, 13:250*.

9.0 Appendix

C1. Log intensity values against 29 common brain cancer genes
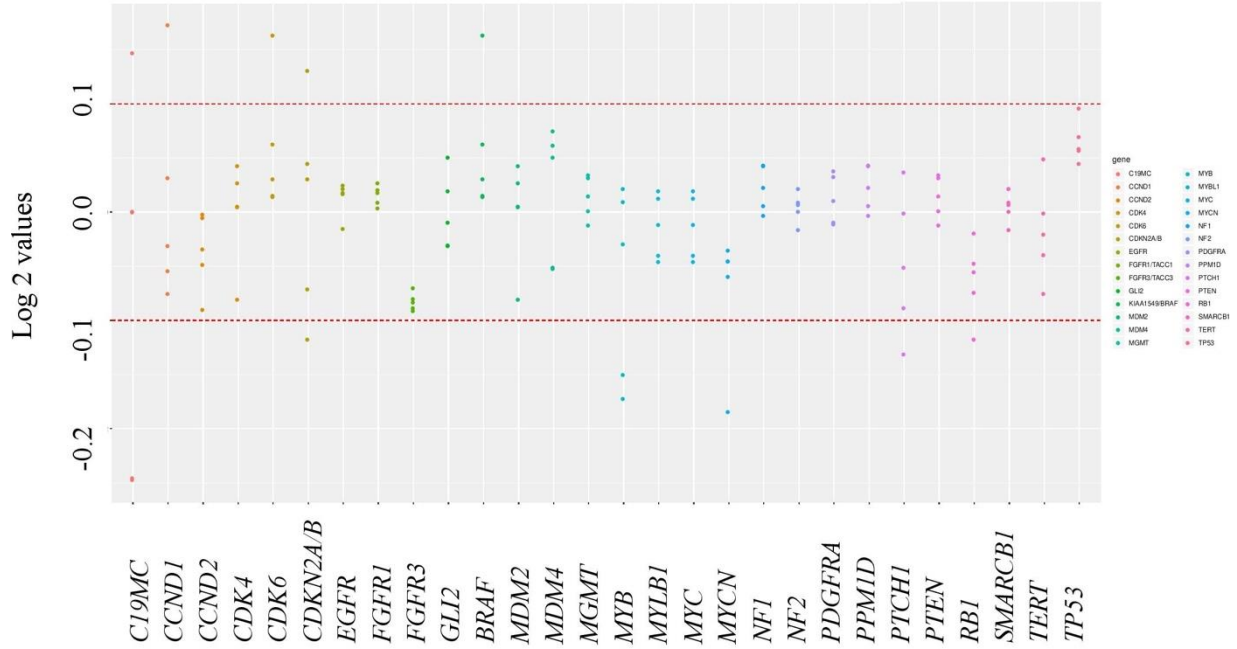
C2. Log intensity values against 29 common brain cancer genes
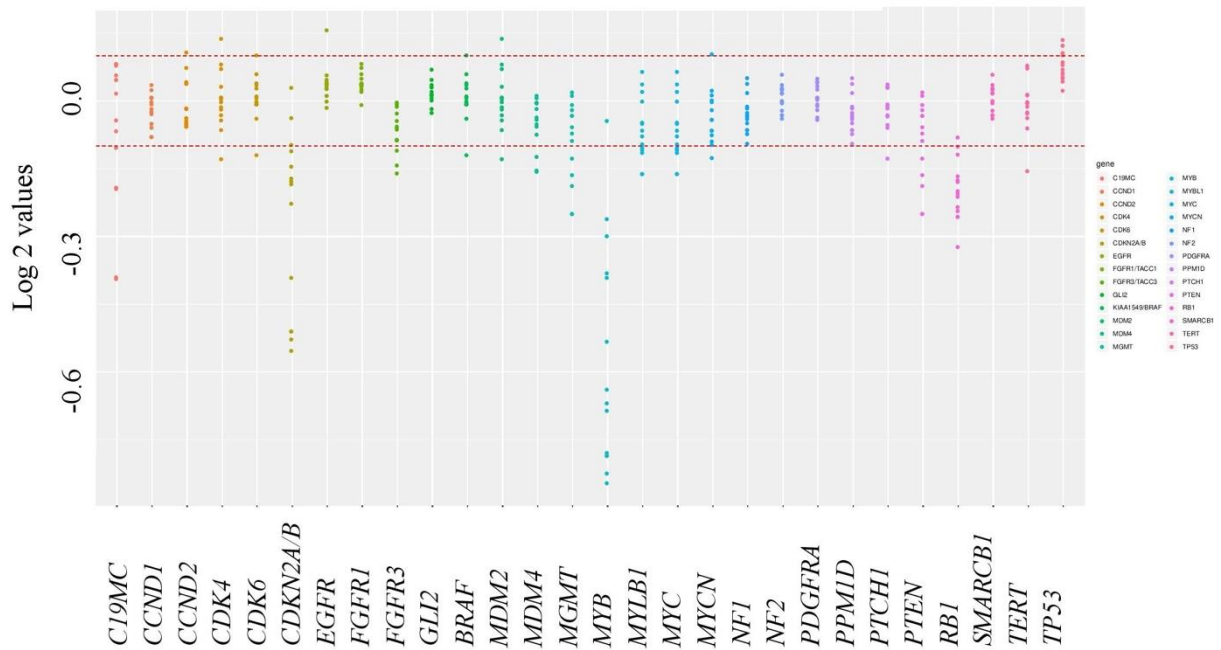
C3. Log intensity values against 29 common brain cancer genes
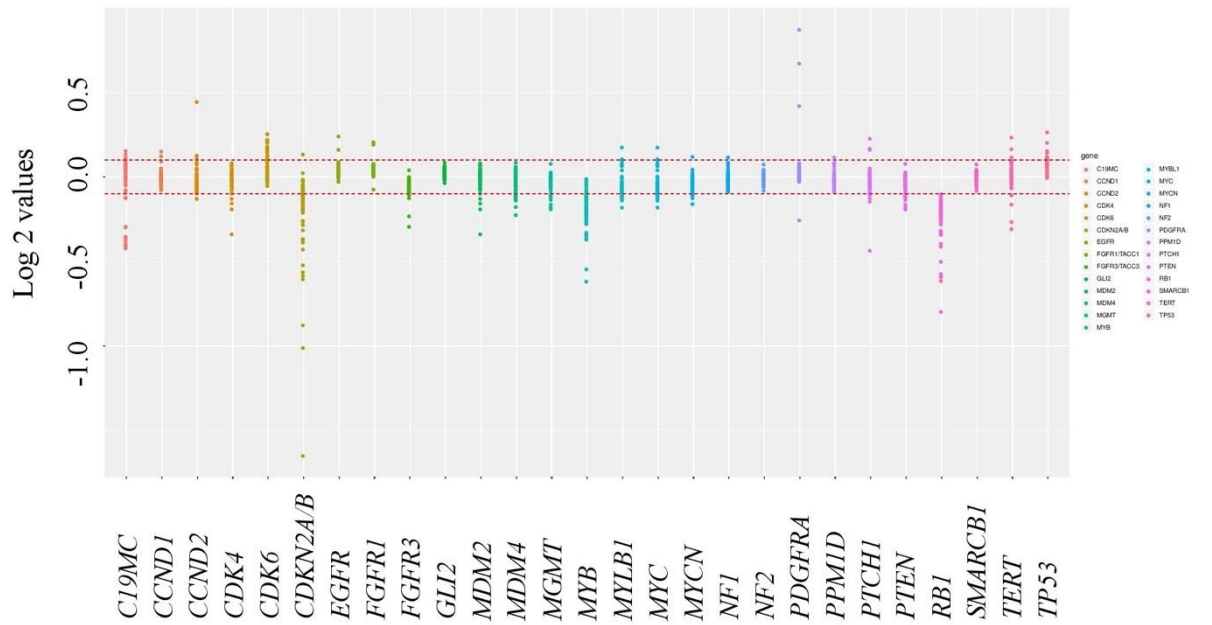
C4. Log intensity values against 29 common brain cancer genes
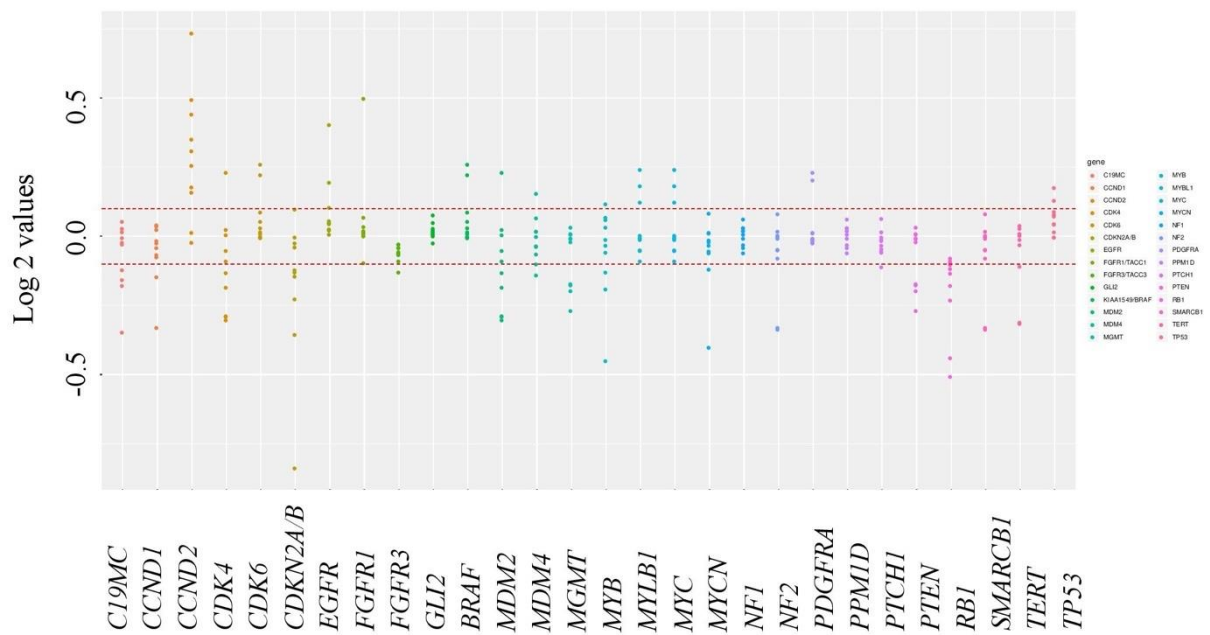
C5. Log intensity values against 29 common brain cancer genes
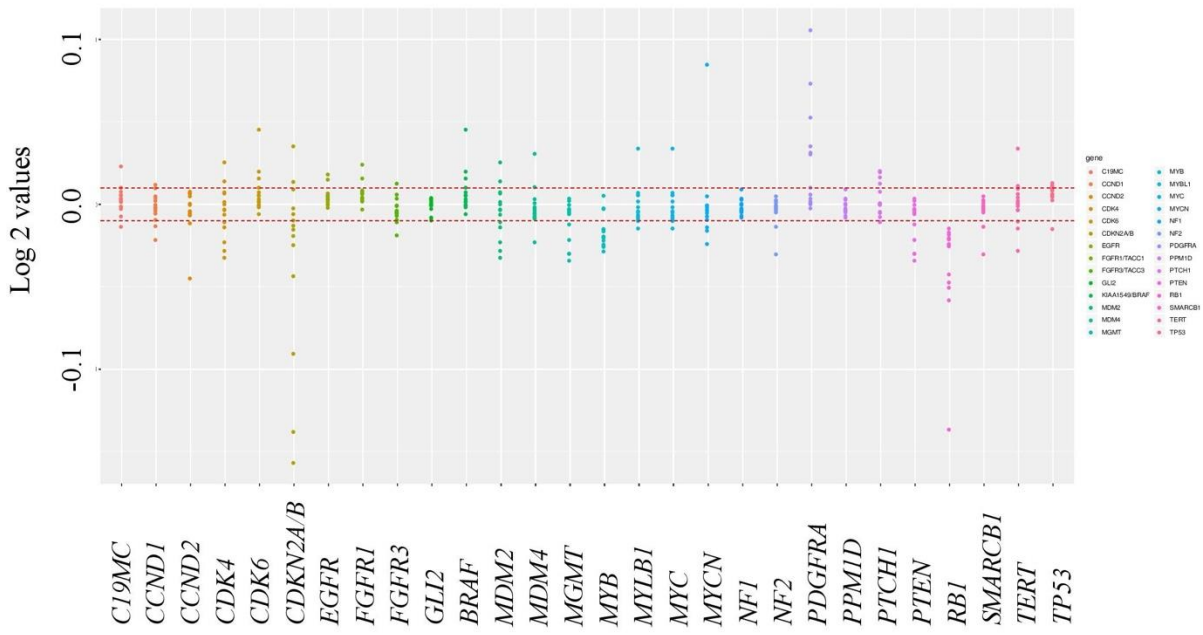
C6. Log intensity values against 29 common brain cancer genes
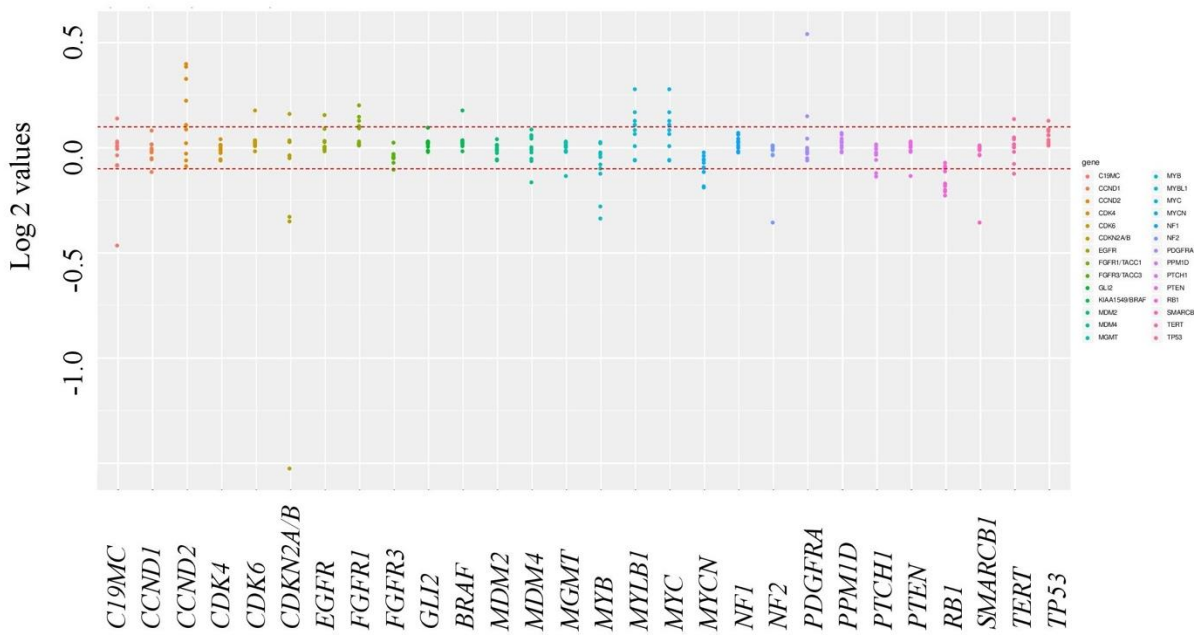
C7. Log intensity values against 29 common brain cancer genes
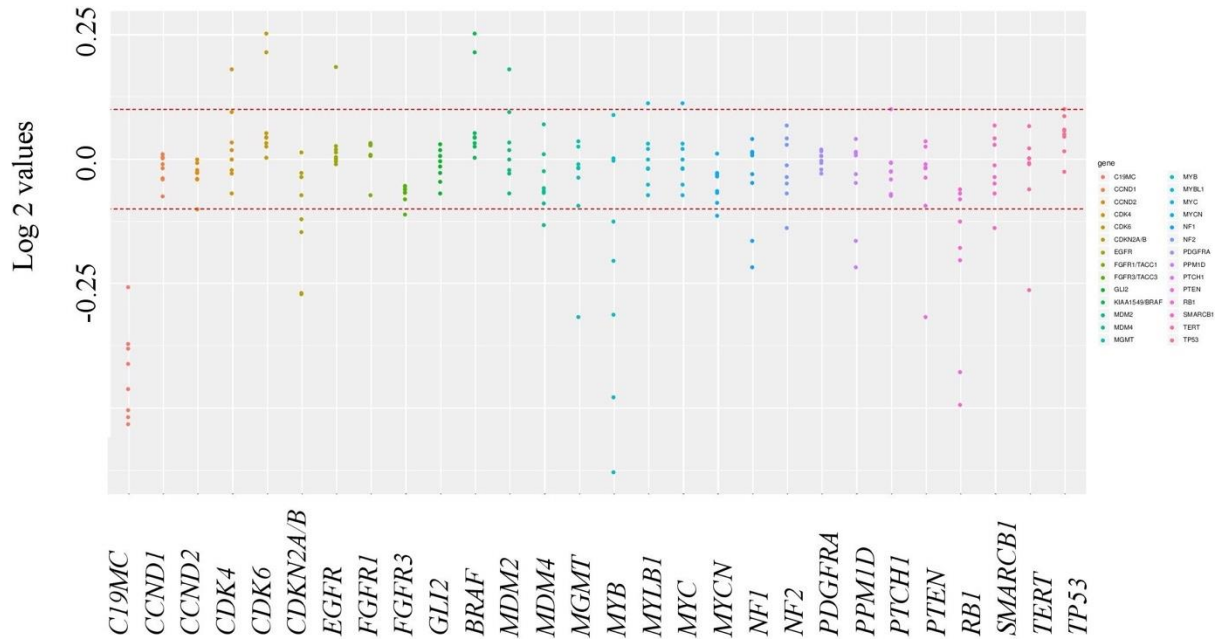
Supplementary figure 1 Distribution of the 29 common brain tumour genes from cluster1-7.

The Red dotted line shows the Thresh-hold for gain (0.1) the top (above 0) red dotted line and loss threshold (-0.1) below 0 red dotted line. Patients in cluster; figure C1, C2, C3 and C5 had poorer survival outcomes than C4, C6 and C7. The groups with poor survival were characterised by co-deletion on CDKND2A/B, MYB1 and RB1. The groups with better survival had around 60% deletion on CDKND2A/B and RB1. In addition, the group had chromosomal gain in gene CCND2 (Figure C4 and C6). However, C7 had a unique C19MC deletion beside having around 60% co-deletion MYB1 and RB1 gene.

Supplementary table 1. Distribution of identified cluster along with Methylation classes, WHO grades and calibration scores

Cluster ID	TCGA ID	Methylation class	Methylation subclass	WHO Grade	class cal-score	subclass cal-score	Gender
C1	TCGA-CS-4938-01B-11D-1894-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.98	F
C1	TCGA-HW-7490-01A-11D-2025-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C1	TCGA-HW-7493-01A-11D-2025-06	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	F
C1	TCGA-FG-7636-01A-11D-2087-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M
C1	TCGA-HT-7601-01A-11D-2087-06	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	F
C2	TCGA-CS-4944-01A-01D-1467-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C2	TCGA-CS-6667-01A-11D-1894-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	F
C2	TCGA-P5-A5EZ-01A-11D-A27L-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.62/0.36	M
C2	TCGA-DB-A4XF-01A-11D-A27L-06	Glioma IDH mutant	Astrocytoma	III	0.99	0.98	F
C2	TCGA-DH-A66B-01A-11D-A29T-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.96	M
C2	TCGA-S9-A6TZ-01A-21D-A32C-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.97	F
C2	TCGA-HT-A5RB-01A-11D-A28N-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.96	F
C2	TCGA-S9-A6U6-01A-12-D-A33U-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.94	M
C2	TCGA-S9-A7R8-01A-11D-A34K-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.98	F
C2	TCGA-DU-A7TC-01A-21D-A34K-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C2	TCGA-S9-A7R3-01A-11D-A34K-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	F
C2	TCGA-VV-A86M-01A-11D-A368-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.96	F
C2	TCGA-DB-A64X-01A-11D-A29T-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.72/0.27	F

C3	TCGA-CS-6290-01A-11D-1706-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C3	TCGA-CS-4942-01A-01D-1467-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	F
C3	TCGA-HT-7884-01B-11D-2399-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.94	F
C3	TCGA-HT-8563-01A-11D-2399-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.59/0.4	F
C3	TCGA-HT-7858-01A-11D-2399-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C3	TCGA-HT-7855-01A-11D-2399-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.93	M
C3	TCGA-HW-8320-01A-11D-2399-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.98	M
C3	TCGA-HW-8321-01A-11D-2399-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.97	M
C3	TCGA-HT-8106-01A-11D-2399-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.22/0.76	M
C3	TCGA-DB-A4XB-01A-11D-A26N-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.98	M
C3	TCGA-FG-A60L-01A-12D-A31M-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	F
C3	TCGA-FG-A6J3-01A-11D-A31M-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	F
C3	TCGA-DH-A66D-01A-11D-A31M-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	F
C3	TCGA-DB-A4XD-01A-11D-A27L-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M
C3	TCGA-HW-A5KL-01A-11D-A27L-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C3	TCGA-P5-A5EU-01A-11D-A27L-05	Glioma IDH mutant	Subclass HG Astrocytoma	III	0.99	0.99	M
C3	TCGA-HW-A5KM-01A-11D-A27L-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C3	TCGA-P5-A5EW-01A-11D-A27L-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	F
C3	TCGA-TM-A7C4-01A-11DA32C-05	Glioma IDH mutant	Astrocytoma	II	0.93	0.92	F
C3	TCGA-S9-A6TU-01A-12D-A32C-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C3	TCGA-P5-A72X-01A-11D-A32C-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M
C3	TCGA-P5-A733-01A-11D-A32C-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.97	F

C3	TCGA-TM-A7CF-02A-11D-A32C-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.97	F
C3	TCGA-DB-A750-01A-11D-A32C-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M
C3	TCGA-DU-A6S7-01A-21D-A32C-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.88/0.1	F
C3	TCGA-DU-A76O-01A-11D-A32C-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C3	TCGA-S9-A6U9-01A-11D-A32C-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M
C3	TCGA-HT-A74O-01A-11D-A32C-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.97	M
C3	TCGA-P5-A72W-01A-11D-A32C-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.98	M
C3	TCGA-DB-A75L-01A-11D-A32C-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.92	F
C3	TCGA-DU-A5TP-01A-11D-A28N-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.98	M
C3	TCGA-P5-A5F1-01A-11D-A28N-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.96	M
C3	TCGA-S9-A6WL-01A-21D-A33U-05	Glioma IDH mutant	Subclass Astro/1q/19p codeleted oligodendroglioma	III	0.99	0.99	M
C3	TCGA-S9-A6TS-01A-12D-A33U-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.16/0.83	F
C3	TCGA-TM-A7CA-01A-21D-A33U-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.89	M
C3	TCGA-S9-A6U1-01A-21D-A33U-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	F
C3	TCGA-S9-A6U8-01A-21D-A33U-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M
C3	TCGA-S9-A6WG-01A-11D-A33U-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.98	M
C3	TCGA-S9-A7QW-01A-11D-A34D-05	Glioma IDH mutant	Astrocytoma	III	0.98	0.9	F
C3	TCGA-S9-A7QX-01A-11D-A34D-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.97	F
C3	TCGA-DH-A7UV-01A-12D-A34D-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.94	M
C3	TCGA-S9-A71S-01A-11D-A34D-05	Glioma IDH mutant	Subclass HG Astrocytoma	III	0.99	0.99	F
C3	TCGA-S9-A6WO-01A-21D-A34D-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C3	TCGA-S9-A71Z-01A-11D-A34D-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.63/0.36	F

C3	TCGA-S9-A7R4-01A-12D-A34K-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M
C3	TCGA-E1-A7Z6-01A-12D-A34K-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	F
C3	TCGA-WH-A86K-01A-11D-A368-05	Glioma IDH mutant	Astrocytoma/HG Astro	II	0.99	0.79/0.2	M
C3	TCGA-F6-A804-01A-11D-A368-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.96	F
C3	TCGA-VM-A8CH-01A-12D-A368-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.98	F
C3	TCGA-TM-A84Q-01A-12D-A368-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.98	M
C3	TCGA-WY-A85A-01A-21D-A368-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.98	M
C3	TCGA-TM-A84F-01A-11D-A368-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.98	0.43/0.53	M
C3	TCGA-WY-A858-01A-11D-A368-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.86/0.13	F
C3	TCGA-DU-A5TU-01A-11D-A28N-05	Glioma IDH mutant	Astrocytoma/HG Astro	II	0.99	0.39/0.6	F
C3	TCGA-DB-A75M-01A-11D-A32C-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C3	TCGA-FG-A4MX-01A-11D-A26N-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.98	M
C4	TCGA-CS-5393-01A-01D-1467-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.92	M
C4	TCGA-DB-5277-01A-01D-1467-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.12/0.86	M
C4	TCGA-E1-5304-01A-01D-1467-05	Glioma IDH mutant	Subclass HG Astrocytoma	III	0.99	0.97	M
C4	TCGA-DB-5273-01A-01D-1467-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M
C4	TCGA-HT-7686-01A-11D-2254-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.91	F
C4	TCGA-DU-7007-01A-11D-2025-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.98	M
C4	TCGA-HT-7478-01A-11D-2025-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C4	TCGA-HT-7604-01A-11D-2087-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.93	F
C4	TCGA-DU-A5TU-01A-11D-A28N-05	Glioma IDH mutant	Astrocytoma/HG Astro	II	0.99	0.39/0.6	F
C4	TCGA-TM-A841-01A-11D-A368-05	Glioma IDH mutant	Subclass HG Astrocytoma	III	0.99	0.99	M

C5	TCGA-FG-6689-01A-11D-1894-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C5	TCGA-DU-7010-01A-11D-2025-05	Glioma IDH mutant	Subclass HG Astrocytoma	III	0.99	0.99	F
C5	TCGA-CS-6665-01A-11D-1894-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.25/0.74	F
C5	TCGA-HW-8319-01A-11D-2399-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.32/0.67	F
C5	TCGA-HT-A61B-01A-11D-A29T-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.76/0.23	M
C5	TCGA-HT-A616-01A-11D-A29T-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.98	F
C5	TCGA-HT-A618-01A-11D-A29T-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.49/0.5	F
C5	TCGA-HT-A5R7-01A-11D-A28N-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.17/0.15	M
C5	TCGA-HT-7477-01B-11D-A28N-05	Glioma IDH mutant	Subclass HG Astrocytoma	III	0.99	0.98	M
C5	TCGA-DH-A7UT-01A-12D-A34D-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M
C5	TCGA-S9-A7R7-01A-11D-A34K-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C5	TCGA-DH-A7UU-01A-12D-A34D-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.53/0.45	M
C5	TCGA-WY-A859-01A-12D-A368-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.96	F
C5	TCGA-S9-A89Z-01A-11D-A368-05	Glioma IDH mutant	Subclass HG Astrocytoma	III	0.99	0.94	M
C6	TCGA-E1-5303-01A-01D-1467-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M
C6	TCGA-CS-4943-01A-01D-1467-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.79/0.15	M
C6	TCGA-FG-6691-01A-11D-1894-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.94	F
C6	TCGA-E1-5307-01A-01D-1894-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.96	F
C6	TCGA-CS-6666-01A-11D-1894-05	Glioma IDH mutant	Astrocytoma/HG Astro	III	0.99	0.38/0.6	M
C6	TCGA-FG-8185-01A-11D-2254-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.98	M
C6	TCGA-HT-7479-01A-11D-2025-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M

C6	TCGA-HT-7478-01A-11D-2025-05	Glioma IDH mutant	Astrocytoma/HG Astro	II	0.98	0.64/0.31	M
C6	TCGA-HT-7485-01A-11D-2025-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.99	M
C6	TCGA-DU-7299-01A-21D-2025-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M
C7	TCGA-E1-5302-01A-01D-1467-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.99	M
C7	TCGA-DH-5142-01A-01D-1467-05	Glioma IDH mutant	Astrocytoma	III	0.99	0.96	M
C7	TCGA-CS-5394-01A-01D-1467-05	Glioma IDH mutant	Subclass Astro/1q/19p codeleted oligodendroglioma	III	0.99	0.99	M
C7	TCGA-DU-7296-01A-11D-2025-05	Glioma IDH mutant	Subclass Astro/1q/19p codeleted oligodendroglioma	III	0.99	0.84/0.13	F
C7	TCGA-E1-5305-01A-01D-1894-05	Glioma IDH mutant	Subclass Astro/1q/19p codeleted oligodendroglioma	III	0.99	0.85/0.1	M
C7	TCGA-HT-7606-01A-11D-2087-05	Glioma IDH mutant	Subclass HG Astrocytoma	II	0.99	0.99	F
C7	TCGA-P5--A5EV-01A-11D-A27L-05	Glioma IDH mutant	Astrocytoma	II	0.99	0.98	M
C7	TCGA-S9-A6U5-01A-12D-A33U-05	Glioma IDH mutant	Subclass Astro/1q/19p codeleted oligodendroglioma	II	0.99	0.98	M

Supplementary table 2.Molecules associated with altered regions in more than three clusters and their associated pathways.

Ingenuity Canonical Pathways	-log(p-value)	Ratio	Molecules associated with altered region and pathway
Axonal Guidance Signaling	5,79	0,224	*ABLIM1,ABLIM3,ADAM19,ADAM20,ADAM21,ADAM23,ADAMTS15,ADAMTS16,ADAMTS19,ADAMTS2,ADAMTS8, ARPC2,BDNF,BMP10,CFL1,CRKL,CXCR4,DPYSL5,ECE2,EFNA5,EFNB2,EPHA3,EPHA4,EPHA6,EPHB3,FARP2,FZD4, FZD5,FZD7,GDF7,GLI2,GNB1L,GNB3,GNB4,GNG2,ITGA4,KALRN,LNPEP,MAPK1,MME,MMP10,MMP12,MMP13, MMP20,MMP21,MMP26,MMP27,MMP3,MMP7,MYL1,MYL3,NCK1,NGEF,NTF3,PAK2,PDGFB,PFN3,PIK3C2A, PIK3CA,PIK3CB,PIK3R1,PIK3R4,PLCD1,PLCD4,PLCH1,PLCL2,PLXNA1,PLXNB1,PLXNB2,PLXND1,PPP3R1,PRKAR2A, PRKCE,PRKCH,PRKCI,PRKD1,PRKD3,PSMD14,RAC2,RALB,RAP2A,RAP2B,RASD2,RHOA,RHOD,ROBO1,ROBO2,ROBO3, ROCK2,RRAS2,RTN4R,SEMA3F,SEMA3G,SEMA4C,SEMA5A,SHANK2,SOS1,TUBA3C/TUBA3D,TUBA3E,UNC5A,VEGFC, WNT10A,WNT11,WNT5B,WNT6,WNT7A,WNT7B*
Cell Cycle: G1/S Checkpoint Regulation	0,754	0,0455	*FOXO1,RB1,TFDP1,TGFB3, CCNE2, MYC*
Chondroitin Sulfate Biosynthesis	1,56	0,118	*HS3ST5,SULT1B1,SULT1E1,SULT2A1,SULT2B1,UST*
Chondroitin Sulfate Biosynthesis (Late Stages)	1,9	0,14	*HS3ST5,SULT1B1,SULT1E1,SULT2A1,SULT2B1,UST*

Coagulation System	3,14	0,206	*F11,F13B,FGA,FGG,PLG,PROS1,TFPI,BDKRB1,PLAUR,PLG,SERPINA5*
Dermatan Sulfate Biosynthesis	1,97	0,13	*DSE,HS3ST5,SULT1B1,SULT1E1,SULT2A1,SULT2B1,UST*
Dermatan Sulfate Biosynthesis (Late Stages)	2,64	0,171	*DSE,HS3ST5,SULT1B1,SULT1E1,SULT2A1,SULT2B1,UST*
ERK/MAPK Signaling	1,08	0,182	*CRKL,DUSP1,DUSP2,ELF1,ETS1,HSPB3,ITGA4,MAPK1,MYCN,PAK2,PIK3C2A,PIK3CA,PIK3CB,PIK3R1,PIK3R4,PLA2G3,* *PPARG,PPM1L,PPP1CA,PPP1CB,PPP1R14B,PPP1R7,PPP2CA,PPP2R5B,PRKAR2A,PRKCE,PRKCI,* *RAC2,RALB,RAP2A,* *RAP2B,RASD2,RRAS2,SOS1,YWHAQ*
ERK5 Signaling	1,01	0,0857	*FOXO3,NGF,NRAS,RAP1A,RRAS,SGK1*
Glioblastoma Multiforme Signaling	3,32	0,244	*APC,CCND1,CDKN1B,CTNNB1,E2F8,FOXO1,FZD4,FZD5,FZD7,MAPK1,PDGFB,PIK3C2A,PIK3CA,PIK3CB,* *PIK3R1,PIK3R4,PLCD1,PLCD4,PLCH1,PLCL2,RAC2,RALB,RAP2A,RAP2B,RASD2,RB1,RHOA,RHOB,RHOD,* *RHOG,RHOJ,RHOQ,RRAS2,SOS1,WNT10A,WNT11,WNT5B,WNT6,WNT7A,WNT7B*
Glioma Invasiveness Signaling	2,55	0,274	*F2R,ITGB5,MAPK1,PIK3C2A,PIK3CA,PIK3CB,PIK3R1,PIK3R4,RAC2,RALB,RAP2A,RAP2B,RASD2,RHOA,* *RHOB,RHOD,RHOG,RHOJ,RHOQ,RRAS2*
Glioma Signaling	0,363	0,0273	*RAP2A,RB1,TFDP1,CALM1,IGF2R,PRKCH,RB1,RRAS*
Heparan Sulfate Biosynthesis	1,15	0,0938	*HS3ST5,SULT1B1,SULT1E1,SULT2A1,SULT2B1,UST*
Heparan Sulfate Biosynthesis (Late Stages)	1,36	0,105	*HS3ST5,SULT1B1,SULT1E1,SULT2A1,SULT2B1,UST*

Hepatic-Fibrosis/Hepatic-Stellate cell Activation	1,29	0,0432	*COL4A1,COL4A2,CXCL3,CXCL8,CXCL9,FGF2,FLT1,VEGFC*
IL-12 Signaling and Production in Macrophages	1,88	0,0985	*APOE,IFNA1/IFNA13,IFNA2,IFNA21,IFNA4,IFNA5,IFNA6,IFNA8,LPA,NFKBIB,PRKCH,TGFB1,TGFB3*
IL-15 Production	1,19	0,0847	*FLT1,FLT3LG,FRK,FYN,IFNA1/IFNA13,IFNA21,IFNA5,IFNB1,JAK2,LMTK3*
IL-17 Signaling	1,01	0,05	*CXCL5,CXCL8,MAPK10,RAP2A*
IL-17A Signaling in Airway Cells	1,29	0,0625	*CXCL3,CXCL5,CXCL6,MAPK10*
IL-3 Signaling	3,66	0,304	*CRKL,CSF2RB,FOXO1,GAB2,IL3,JAK2,MAPK1,PIK3C2A,PIK3CA,PIK3CB,PIK3R1,PIK3R4,PPP3R1, PRKCE,PRKCH,PRKCI, PRKD1,PRKD3,RALB,RAP2A,RAP2B,RASD2,RRAS2,SOS1*
NF-κB Signaling	0,455	0,0281	*BMPR1B,FLT1,RAP2A,TLR3,TNFSF11*
p53 Signaling	0,967	0,194	*ATR,CCND1,CCND2,CCNG1,CHEK1,CTNNB1,EP300,JMY,KAT2B,PIK3C2A,PIK3CA,PIK3CB,PIK3R 1 PIK3R4,RB1,RPRM,TIGAR,TOPBP1,TP53AIP1*
PI3K Signaling in B Lymphocytes	0	0,0455	*CALM1 (includes others),FOXO3,FYN,NRAS,RAP1A,RRAS*
PI3K Signaling in B Lymphocytes	0	0,0145	*IRS2,RAP2A*
PI3K/AKT Signaling	0,491	0,0606	*FOXO1,FOXO3,GSK3A,HSP90AA1,JAK2,NFKBIB,NFKBID,RRAS*
Protein Kinase A Signaling	0	0,0483	*ACP4,AKAP11,AKAP12,AKAP7,CALM1,GSK3A,KDELR1,NFKBIB,NFKBID,PDE7B,PPP1R14C,PRKC H,PTP4A1,PTPN21,TGFB1,TGFB3,TNNI3,TULP2,VASP*
PTEN Signaling	0,304	0,0484	*FOXO3,FOXO1,IGF2R,ITGA4,NRAS,RAP1A,RAP2A,FLT1,RRAS,GSK3A,IGF2R*
Retinoate Biosynthesis I	0,77	0,0588	*ADH1C,ADH7,RDH12,RDH13*
Retinoate Biosynthesis II	1,97	0,75	*RBP1,RBP2,RBP5*
STAT3 Pathway	0,893	0,037	*BMPR1B,FGF2,JAK2,MAPK10,MYC,FLT1,IGF2R,JAK2,MAP3K10,RRAS,TGFB1*

Spermidine Biosynthesis	1,05	0,5	*AMD1*
VEGF Signaling	0,445	0,0606	*EIF2B2,EIF2S1,FLT1,FOXO1,FOXO3,RRAS,RAP2A,VEGFC,MAPK1,PIK3C2A,PIK3CA,PIK3CB,PIK3R1,PIK3R4*
Wnt/β-catenin Signaling	0,593	0,0289	*CDH12,DKK2,FZD6,MYC,SFRP2*
Xenobiotic Metabolism Signaling	1,07	0,0654	*CITED2,FTL,GSTM2,GSTM5,HS3ST5,NRAS,PPM1J,RAP1A,RRAS,SULT1B1,SULT1E1,SULT2A1,SULT2B1,UGT2B11,UGT2B15, UGT2B4,UST*

Bidii Stephen Ngalah-Curriculum vitae

Date of birth: 02/06/1987 **Nationality:** Kenyan   **Gender** Male   (+49) 1632266075 |

stephenbidii@gmail.com   Im Neuenheimer Feld 135, 69120, Heidelberg, Germany

WORK EXPERIENCE

01/06/2017 – CURRENT – Heidelberg, Germany

POSTGRADUATE UNDER GERMAN ACADEMIC EXCHANGE SERVICE (DAAD) – UNIVERSITAT HEIDELBERG

- Analysis of copy number variation profiles in brain tumors in the context of a methylation based classifier.
- Using genome wide sequencing approach to identify germline predisposing variants in familial cancers.
- Molecular prevention and genotoxicity research Heidelberg, Germany

15/06/2020 – CURRENT – Freiburg, Germany

GUEST PHD RESEARCH SCIENTIST – UNIVERSITATSKLINIKUM FREIBURG-MOLECULAR PREVENTIVE MEDICINE

- *In vitro* and *in vivo* genotoxicity assessments of chromosomal anomalies using human cell lines and predictive cancer biomarkers e.g Micronucleus and nuclear buds.

01/01/2018 – 01/01/2020 – Heidelberg, Germany

DOCTORAL RESEARCHER – GERMAN CANCER RESEARCH CENTER-DKFZ

1. Helmholtz international graduate school for cancer research PhD student training
2. Germline mutation screening in familial cancers with focus to neuroendocrine tumours and multiple myeloma
3. Brain tumour screening/methylome diagnosis using illumina epic 450k and 850K arrays.
4. Identification of copy number alteration profiles and their roles in grading of brain tumour

06/04/2012 – CURRENT

PART-TIME LECTURER-DEPARTMENT OF PURE AND APPLIED SCIENCES – TECHNICAL

UNIVERSITY OF MOMBASA AND PWANI UNIVERSITY

- lecturing medical microbiology, molecular biology, epidemiology and introduction to bioinformatics units in both medical and applied sciences departments.
- I participate in review of curriculum and set exams. Mombasa, Kenya

06/01/2013 – 01/01/2015

POSTGRADUATE TRAINEE – DEPARTMENT OF EMERGING INFECTIOUS DISEASES,GLOBAL

EMERGING INFECTIOUS SURVEILLANCE(DEID-GEIS) AND R

I completed a clinical trial-based project entitled " major Genome region underlying artemisinin resistance in malaria parasite isolates from Kisumu county, western Kenya". General objective: To determine Plasmodium falciparum genetic polymorphism from uncomplicated malaria subjects with distinct clearance after artemisinin administration and parasite before introduction of artemisinin.

10/02/2011 – 20/12/2011

LABORATORY IN-INCHARGE COAST PROVINCE- MICRONUTRIENT NATIONAL SURVEY –

KENYA MEDICAL RESEARCH INSTITUTE(KEMRI)- CENTER OF PUBLIC HEALTH RESEARCH

Responsible for a field and laboratory-based epidemiology study aimed in identification of micro-nutrient deficiencies burden in Kenya05/05/2010 – 05/05/2011

INTERNSHIP – UNIVERSITY OF WASHINGTON PARTNER IN PREVENTION CLINICAL TRIALS

AT KENYATTA NATIONAL HOSPITAL

- ◦ I worked under a clinical trial project to determine efficacy of Truvada and Tenofovir HIV prophylactic drugs in discordant couples.

Nairobi, Kenya

06/06/2006 – 14/06/2007

EPIDEMIOLOGY DEPARTMENT, RTS,S VACCINE CLINICAL TRIAL, FIELD OFFICER – KEMRI-WELLCOME TRUST KILIFI

- ◦ Conducting pre and post vaccination follow ups
- ◦ Rapid test for malaria in children below 5 years
- ◦ Collection of blood sample for the lab
- ◦ Reporting cases to study clinician for incidence and prevalence estimation
- ◦ Conducting both active and passive case detections Mombasa, Kenya

EDUCATION AND TRAINING

- ●

01/04/2012 – 08/06/2015 – Nairobi, Kenya

MSC. MOLECULAR MEDICINE – Jomokenyatta University, College of health sciences-Institute of Tropical

Medicine and infectious diseases

Thesis: Major genome region underlying artemisinin resistance in malaria parasite isolates from Kisumu County, Western Kenya.

05/05/2007 – 20/12/2010 – Nairobi, Kenya

BSC. MEDICAL MICROBIOLOGY- FIRST CLASS HONORS – Jomokenyatta university of agriculture and Technology.

Fourth year research project : Isolation of *Streptococcus pneumoniae* serotypes in patients attending Kilifi district hospital.

LANGUAGE SKILLS

Mother tongue(s): ENGLISH AND SWAHILI

Other language(s):

	UNDERSTANDING		SPEAKING		WRITING
	Listening	Reading	Spoken production	Spoken interaction	
GERMAN	B1	B1	B1	B1	B1

Levels: A1 and A2: Basic user; B1 and B2: Independent user; C1 and C2: Proficient user

ORGANISATIONAL SKILLS

Organisational skills

Good organisation skills gained during project management in my career.

COMMUNICATION AND INTERPERSONAL SKILLS

Communication and interpersonal skills

Good communication skills gained in my scientific and lecturing career

JOB-RELATED SKILLS

Job-related skills

- Molecular biology techniques e.g PCR, qPCR , sanger sequencing and western-blot
- Sequenome MassArray for genotyping single nucleotide polymorphism
- Whole genome sequence analysis, mutation, polymorphisms and gene alteration screening .
- Copy number alteration profiles identification and survival analysis
- Ingenuity pathway analysis to identify the pathways affected by altered genes.
- Culture technique and cloning of genes
- Site directed Mutagenesis and beginner in trans-infections experiments.
- Good background in clinical pathogen diagnosis
- Conducting epidemiology studies on infectious diseases
- Intermediate R-user
- Intermediate Bioinformatic skills
- Advanced Biostatistics analysis skills

- **Referees**

1. Dr.Helke Hillebrand
   Heidelberg Graduate academy
   helke.hillebrand@uni-heidelberg.de
   +49 (0) 6221-54 -19760

2. PD.Dr. Obul Reddy Bandapalli
   German cancer research center
   Hopp Children's Cancer Center Heidelberg (KiTZ)
   Im Neuenheimer Feld 580
   D-69120 Heidelberg
   +49 (0) 6221-42-1809

3. Dr. rer. nat. Stefan Pusch
   German cancer research s.pusch@dkfz-heidelberg.de
   +49 (0) 6221-42-1473

4. Colonel, Dr. EDWIN KAMAU
   USAMRU-K/KEMRI- GEIS PROGRAM Kisumu.
   Edwin.Kamau@usamru-k.org
   +1 (301) 273-5658

Acknowledgements

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema

   Analysis of copy number variation profiles in brain tumors in the context of a methylation based classifier

handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keener unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In-oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.*

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum   Heidelberg, 10th November 2021   Unterschrift

Wenn dies nicht zutrifft, machen Sie folgende Angaben:

3. *Die Arbeit oder Teile davon habe ich wie folgt an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt:*

   Titel der Arbeit:

   Hochschule und Jahr:

   Art der Prüfungs- oder Qualifikationsleistung: