Genetic Susceptibility and Molecular Characterization of Glioma

Karim Labreche

▶ **To cite this version:**

Karim Labreche. Genetic Susceptibility and Molecular Characterization of Glioma. Quantitative Methods [q-bio.QM]. Université Paris Saclay (COmUE), 2018. English. NNT : 2018SACLS161 . tel-02170512

HAL Id: tel-02170512
https://theses.hal.science/tel-02170512

Submitted on 2 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université PARIS-SACLAY

NNT : 2018SACLS161

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

Genetic susceptibility and molecular characterization of glioma

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École doctorale n°582 Cancérologie, biologie, médecine, santé (CBMS)
Discipline : Aspects moléculaires et cellulaires de la biologie

Thèse présentée et soutenue à Paris, le 27 juin 2018, par

Karim LABRECHE

Composition du Jury :

Pr François Ducray PU-PH, Université Claude Bernard Lyon 1 Cancer Research Centre of Lyon, INSERM U1052, CNRS UMR5286	Président
Pr Jean Mosser PU-PH, Université de Rennes 1, CHU de Rennes, Institut de génétique et développement, CNRS–UR1, IGRD UMR 6290	Rapporteur
Dr Franck Bourdeaut Praticien CLCC, Institut Curie, PSL Research University INSERM U830	Rapporteur
Dr Jacques Grill PH, Université Paris-sud, Institut Gustave Roussy Département de Cancérologie de l'Enfant et de l'Adolescent et UMR8203	Examinateur
Dr Alex Duval DR, UPMC, Hôpital Saint-Antoine UMR-S UPMC/Inserm 938, centre de recherches Saint-Antoine	Examinateur
Pr Marc Sanson PU-PH, Sorbonne Universités, UPMC. Hôpital Pitié-Salpêtrière Institut du Cerveau et de la Moelle épinière, ICM, Inserm U 1127, CNRS UMR 7225, UMR S 1127	Directeur de thèse
Pr Richard Houlston PU-PH, University of London. Institute of Cancer Research, ICR Molecular and Population Genetics Team	Co-Directeur de thèse
Dr Emmanuelle Huillard CR, Sorbonne Universités, UPMC. Hôpital Pitié-Salpêtrière Institut du Cerveau et de la Moelle épinière, ICM, Inserm U 1127, CNRS UMR 7225, UMR S 1127	Invitée

Thèse de doctorat

« *Autant que savoir, douter me plaît* »

Dante

Acknowledgements

Most importantly, I would like to thank my supervisors Marc Sanson and Richard Houlston. Marc has provided me with rich set of opportunities and a freedom to learn, thank you for guidance and patience. Richard has imparted his knowledge of scientific rigour and provided continuous guidance. I have learned so much during the past four years and this will serve me well to my future career.

Thank you also to all members of the 'Molecular and Population Genetics' team in London and the 'Neuro-oncologie experimentale' team in Paris past and present for your tremendous support and friendship. I spent unforgettable moments with you, at work and outside (Phbar). I have meet people who have helped me to be a better person. I will never forget you.

Thanks to The Institute of Cancer Research in London, ICR and the Brain and spine Institute, ICM in Paris for hosting me and to the 'Association pour la Recherche sur les Tumeurs Cérébrales Malignes' - A.R.T.C for funding me during all my thesis.

Thanks to our collaborators on the gliomas projects across Europe. With special mention to Iva Simeonova, Emmanuelle Huillard, Karim Mokhtari and the members of Ligue Nationale Contre Le Cancer for their collaborative support on the TCF12 project. I also thank Melissa Bondy, Beatrice Melin for their kind collaboration using the Glioma International Case Control (GICC) data. I greatly thank Yannick Marie, members of the OncoNeuroteck and the clinical staff of Mazarin for their work in patient recruitment. I Hope we can continue our collaboration in glioma research in order to make real changes for patients.

To my friends, Youcef, Islem, Rima, Guillaume, Raphael , Vivien, Matthieu, Samir, Romain, Aziz, Alex and Jim, thank you for your sincere and strong friendship during all this years.

Finally, to my Dad, Mum, my sister and my brother, Sahim, Dalal, Ines, Badis, tonton Mourad and Khalou, I am hugely thankful for your love and your support.

Table of contents

List of abbreviations

%	Percent
1958-BC	1958 British birth cohort
95% CI	95 percent confidence interval
A	Adenine
AD	Alternate depth
ALSPAC	The Avon longitudinal study of parents and children
APC	Adenomatous polyposis coli
APL	Acute promyelocytic leukaemia
BAM	Binary SAM
BLAST	Basic local alignment search tool
bp	base pairs
BQSR	Base quality score recalibration
BRCA1	Breast cancer associated 1
BRCA2	Breast cancer 2, early onset
BWA	Burrows-Wheeler alignment
C	Cytosine
CADD	Combined annotated dependent depletion
CDKN2A	Cyclin Dependent Kinase Inhibitor 2A
CEU	Utah residents with northern and western european ancestry
CGEMS	Cancer Genetic Markers of Susceptibility
CHB	Han chinese in Beijing, China
CHEK2	Checkpoint kinase 2
Chr	Chromosome
ChromHMM	Chromatin hidden Markov model
CI	Confidence interval
CNV	Copy number variants
Condel	Consensus deleteriousness score
CRC	Colorectal cancer
D'	D-prime
dbSNP	Database of short genetic variations
df	Degrees of freedom
DNA	Deoxyribonucleic acid

DNase	Deoxyribonuclease
dNTP	Deoxynucleotide triphosphate
EGA	European genome-phenome archive
EGFR	Epidermal growth factor receptor
ENCODE	Encyclopedia of DNA elements
eQTL	Expression quantitative trait locus
EVS	Exome variant server
ExAC	Exome aggregation consortium
FDR	False discovery rate
FPR	False positive rate
FWER	Family-wise error rate
G	Guanine
GATK	Genome analysis toolkit
GBM	Glioblastoma multiforme
G-CIMP	Glioma CpG island methylator phenotype
GCTA	Genome-wide complex trait analysis
GEO	Gene expression omnibus
GERP	Genomic evolutionary rate profiling
GO	Gene ontology
GSEA	Gene-set enrichment analysis
GTEx	Genotype-tissue expression
GWAS	Genome-wide association study
h^2	Heritability
HapMap	The human haplotype project
HMM	Hidden markov model
HWE	Hardy-Weinberg equilibrium
I^2	I-squared statistic
IBD	Identity by descent
IBS	Identity by similarity
IDH1	Isocitrate dehydrogenase 1
IDH2	Isocitrate dehydrogenase 2
IDH	IDH1 or IDH2 mutation
i-GSEA4GWAS	Improved gene-set enrichment analysis for genome-wide association study

Indel	Insertion/deletion
JPT	Japanese in Tokyo, Japan
kb	kilo bases
KORA	Co-operative Health Research in the region of Augsburg
LB	Luria broth
LD	Linkage disequilibrium
LGG	Low grade glioma
lincRNA	Long intergenic non-coding RNA
LOH	Loss of heterozygosity
LTL	Leukocyte telomere length
MAF	Minor allele frequency
Mb	Megabase
miRNA	Micro ribonucleic acid
MLH1	MutL homolog 1 (E.coli)
MMR	Mismatch repair
mRNA	Messenger ribonucleic acid
MSH2	MutS homolog 2 (E. coli)
MSH6	MutS homolog 6 (E. coli)
MsigDB	Molecular signatures database
Myc	Avian myelocytomatosis viral oncogene homolog
NBT	National brain tumour
NCBI	The national centre for biotechnology information
NCRN	National cancer research network
NF1	Neurofibromatosis type 1
NF2	Neurofibromatosis type 2
NOS	Not otherwise specified
OD	Optical density
OR	Odds ratio
OS	Overall Survival
p	Short arm of chromosome
P	*P*-value
PCA	Principal components analysis
PCR	Polymerase chain reaction
P^{het}	*P*-value of heterozygosity

PHLDA1	Pleckstrin homology-like domain, family a, MEMBER 1
PHLDB1	Pleckstrin homology-like domain family B member 1
Phylo-HMM	Phylogenetic hidden markov model
PMS2	Postmeiotic segregation increased 2
POLR3B	Polymerase III, RNA, subunit b
PolyPhen-2	Polymorphism Phenotyping 2
POPGEN	Population Genetic Cohort
PTV	Protein truncating variant
q	Long arm of chromosome
Q	Cochran's Q statistic
QC	Quality control
Q-Q	Quantile-quantile
r^2	Correlation co-efficient
RAF	Risk allele frequency
RB	Retinoblastoma
RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid-sequence
RR	Relative risk
RTEL1	Regulator of telomere elongation helicase 1
RTK	Receptor tyrosine kinase
s.d.	Standard deviation
SAM	Sequence alignment/map
SE	Standard error
SEMA3A	Semaphorin 3A
SHAPEIT	Segmented haplotype estimation and imputation tool
SIFT	Sorting intolerant from tolerant
SIR	Standardised incidence ratio
SNAP	SNAP annotation and proxy search
SNP	Single nucleotide polymorphism
SU.VI.MAX	SUpplementation en VItamines et MineralesAntioXydants
SVM	Support vector machine
T	Thymine
TAD	Topological associated domain
TagSNP	Tagging SNPs

TCGA	The Cancer Genome Atlas
TCF12	Transcription Factor 12
TERC	Telomerase RNA component
TERT	Telomerase reverse transcriptase
TF	Transcription factor
TP53	Tumour protein 53
TPR	True positive rate
UCSC	The University of California, Santa Cruz
UK	United Kingdom
UTR	Untranslated region
VCF	Variant call format
VEP	Variant effect predictor
visPIG	Visual plotting interface for genetics
VT	Variable threshold
VTI1A	Vesicle transport through interaction with t-SNAREs 1A
WHO	World Health Organisation
WTCCC2	Wellcome Trust case control consortium 2
YRI	Yoruba in Ibadan, Nigeria
ZBTB16	Zinc finger and BTB domain-containing protein 16
Λ	Lambda
Mg	Micrograms
µl	Microlitres
χ^2	chi-squared

List of figures

List of tables

CHAPTER 1

Introduction

1.1. Overview of central nervous system (CNS) tumours

Malignant tumours of the central nervous system (CNS) are characterised by high morbidity and mortality [1]. Gliomas and meningiomas are the most common types of primary CNS tumour [2][3] (Figure 1.1). Gliomas account for almost 30% of all primary CNS tumours, and 80% of all malignant ones.

Approximately half of newly diagnosed gliomas are classified as glioblastoma, which is the most malignant type of CNS tumour — with median patient survival of approximately 14–17 months in contemporary clinical trials [4][5][6] and approximately 12 months in population-based studies [2][7].

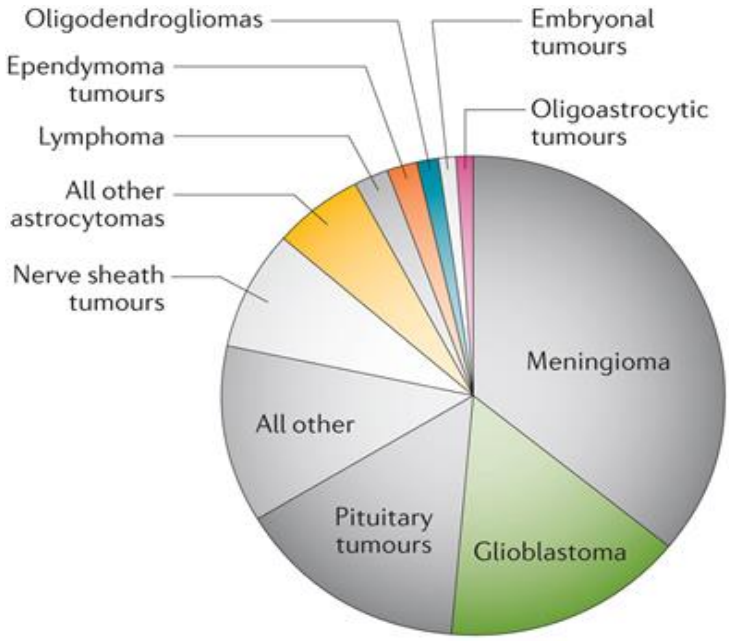


Figure 1.1 Relative frequency of primary brain and central nervous system tumours.Taken from [8]. The figure shows the Central Brain Tumour Registry of the United States (CBTRUS) statistical report, which classified central nervous system tumours by histological groupings (n = 343,175).

During my thesis, the classification of tumours of the CNS by the World Health Organization (WHO) have changed. The 2016 CNS WHO stepped forward over the 2007 WHO classification. In this introduction some parts reference the earlier classification and some the newer.

1.1.1 Histological classification of glioma

Gliomas are tumours that arise from glial or precursor cells (Figure 1.2). On the basis of their histological appearance, they have been traditionally classified as astrocytic, oligodendroglial or ependymal tumours and assigned WHO grades I–IV, which indicate different degrees of malignancy.

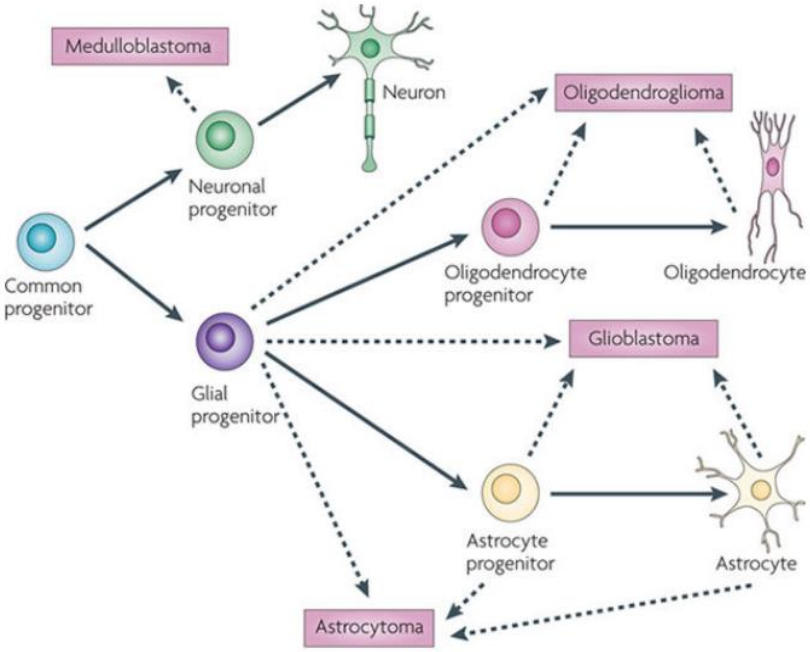


Figure 1.2 Brain cells and brain tumours. Taken from [9]. Self-renewing, common progenitors are thought to produce committed neuronal and glial progenitors that eventually differentiate into mature neurons, astrocytes and oligodendrocytes. Although the precise cells of origin for diffuse glioma variants and medulloblastoma remain largely unknown, a selection of likely candidates for each (dashed arrows) is indicated.

Grade I gliomas pilocytic astrocytomas are benign tumours that occur primarily in children. Astrocytomas, oligodendrogliomas and oligoastrocytomas correspond to low-grade (II) or high-grade (III and IV), which are invasive tumours and can progress to glioblastoma. Grade IV gliomas are glioblastomas including primary and secondary GBM [10][11] (Table 1.3).

	Histologic types (grades)	Age at diagnosis (years)	Survival time (years)
Astrocytic tumours	Pilocytic astrocytoma (I)	children	>20
	Diffuse astrocytoma (II)	young adults	4-10
	Anaplastic astrocytoma (III)	~41	2-5
	Glioblastoma (IV)	45-75	1-2
Oligodendroglial tumours	Oligodendroglioma (II)	50-60	8-20
	Anaplastic Oligodendroglioma (III)	50-60	2-10
Mixed gliomas	Oligoastrocytoma (II)	35-45	5-12
	Anaplastic oligoastrocytoma (III)	~45	2-8

Table 1.1 Histological classification of gliomas based on WHO (2007) guidelines. Based on [10].

In term of frequency the distribution of the histological subtypes of glioma vary, while GBMs account for 56% of all glioma tumours, oligodendrogliomas represent only 5% of cases (Figure 1.3)

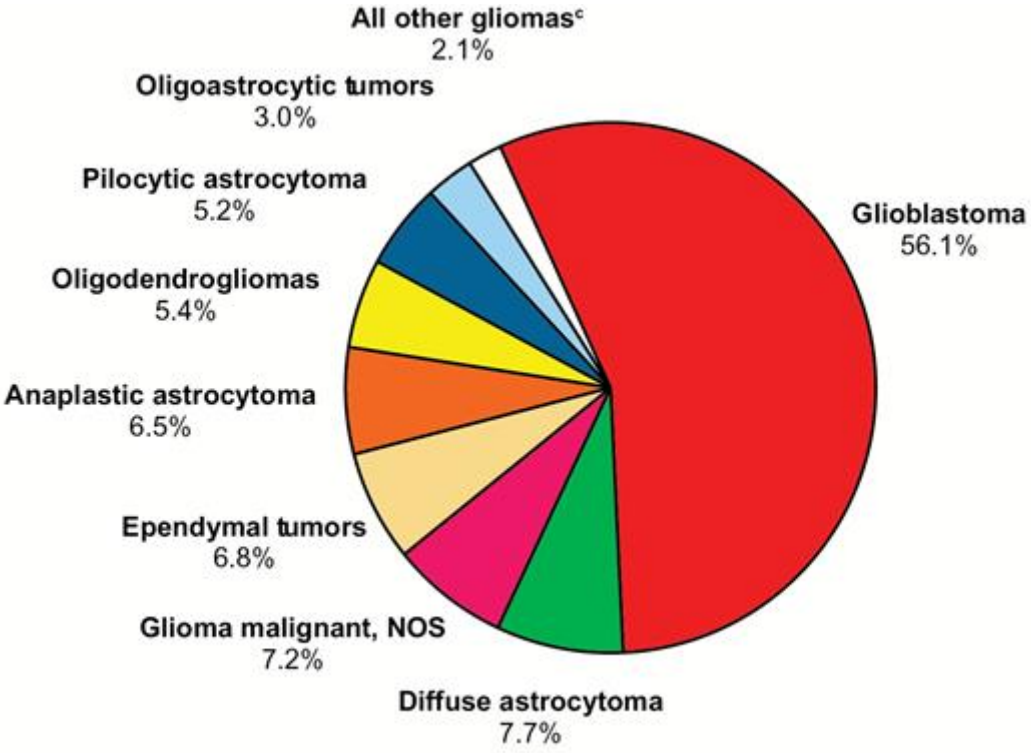


Figure 1.3 Distribution of primary brain and other CNS gliomas by histology subtypes (N=100,619). Taken from CBTRUS Statistical Report: Primary brain and other central nervous system tumours diagnosed in the United States in 2010–2014 [2].

1.1.2 Epidemiology of glioma

Gliomas are diagnosed most commonly in middle-age, with a median age at diagnosis of 56 years in the European population [2][12]. Glioma are rare, with the overall incidence rate for all gliomas being 5.5 per 100,000 in Europeans of which about half were glioblastomas [2]. These tumours are more common in men, with an incidence rates per 100,000 for glioblastomas ranging from 3.95 to 4.03 in men compared with 2.49 to 2.56 in women (Table 1.2). Glioma shows a regional variation, with an incidence rate of gliomas in Japan being less than half of that in Northern Europe [10][11]. The reasons of this regional difference are presently not known.

Mortality rates in glioma differ significantly by histology and age. For example, patients with glioblastoma multiform (GBM) have a 5-year survival rate of 2.7% (Table 1.3), whereas patients with lower grade gliomas, such as pilocytic astrocytoma, oligodendroglioma, and ependymoma, have 5-year survival rates of >70% and patients with diffuse astrocytoma and anaplastic astrocytoma have 5-year survival rates <40%. Overall, and for most histologies, the 5 year survival rate decreases with age.

Glioma histology	Median age	Incidence for 100,000 (95% confidence interval)		
		Overall	Male	Female
Diffuse astrocytoma	48	0.48(0.47-0.49)	0.55(0.54-0.57)	0.42(0.41-0.43)
Anaplastic astrocytoma	53.0	0.40 (0.39-0.41)	0.46 (0.44-0.47)	0.35 (0.33-0.36)
GBM	64.0	3.20 (3.17-3.23)	3.99 (3.95-4.03)	2.52 (2.49-2.56)
Oligodendroglioma	43.0	0.24 (0.23-0.25)	0.28 (0.26-0.39)	0.21 (0.20-0.22)
Anaplastic oligodendroglioma	50.0	0.11 (0.10-0.11)	0.12 (0.11-0.13)	0.09 (0.09-0.10)
Oligoastrocytoma	41.0	0.19 (0.18-0.20)	0.22 (0.21-0.23)	0.16 (0.15-0.17)

Table 1.2 Age-adjusted incidence rates per 100,000 by histology and sex. Based on Central Brain Tumor Registry of the United States (CBTRUS) Statistical Report: Primary brain and other central nervous system tumours diagnosed in the United States in 2010–2014 [2].

Glioma histology	5-year relative survival (95% confidence interval)
Other glioma	38.5 (35.4-41.7)
Astrocytoma unspecified	38.5 (35.9-41.1)
Oligodendroglioma	67.2 (62.5-71.6)
Anaplastic astrocytoma	15.8 (13.6-18.2)
Anaplastic oligodendroglioma	31.5 (25.0-38.3)
Glioblastoma Multiform	2.7 (2.3-3.2)

Table 1.3 Five-year relative survival across all Europe. Data from [15].

1.2. Molecular classification of glioma

In recent years, there has been substantial progress in our understanding of the molecular pathogenesis of glioma allowing generation of a molecular classification of these tumours. The distinction of glioma entities based on their IDH mutation and the status of the co-deletion of chromosome arms 1p and 19q was the fundamental improvement in the 2016 WHO classification [16] (Table 1.4) comparing to previous 2007 classification.

The work in this thesis is focussed on the genetic study of the diffuse astrocytic and oligodendroglia tumours.

The diffuse astrocytic and oligodendroglial tumour category of brain cancers comprises IDH-mutant astrocytic gliomas of WHO grades II–IV, IDH-mutant and 1p/19q co-deleted oligodendroglial tumours of WHO grades II–III, IDH-wild-type glioblastomas of WHO grade IV, and a newly introduced class of histone H3-K27M (H3-K27M)-mutant diffuse midline gliomas of WHO grade IV (Table 1.4).

Tumour classification	WHO grade
Diffuse astrocytic and oligodendroglial tumours	
Diffuse astrocytoma, IDH-mutant	II
• Gemistocytic astrocytoma, IDH-mutant	
Diffuse astrocytoma, IDH-wild-type*	II
Diffuse astrocytoma, NOS	II
Anaplastic astrocytoma, IDH-mutant	III
Anaplastic astrocytoma, IDH-wild-type*	III
Anaplastic astrocytoma, NOS	III
Glioblastoma, IDH-wild-type	IV
• Giant-cell glioblastoma	
• Gliosarcoma	
• Epithelioid glioblastoma*	
Glioblastoma, IDH-mutant	IV
Glioblastoma, NOS	IV
Diffuse midline glioma, H3-K27M-mutant	IV
Oligodendroglioma, IDH-mutant and 1p/19q co-deleted	II
Oligodendroglioma, NOS	II
Anaplastic oligodendroglioma, IDH-mutant and 1p/19q co-deleted	III
Anaplastic oligodendroglioma, NOS	III
Oligoastrocytoma, NOS‡	II
Anaplastic oligoastrocytoma, NOS‡	III

Table 1.4 2016 WHO classification of diffuse astrocytic and oligodendroglial tumours. Based on [16]. NOS categories are reserved for the rare instances that a tumour cannot be molecularly tested or that test results remain inconclusive. H3-K27M, K27M-mutated histone H3; NOS, not otherwise specified. *Provisional tumour entities or variants. ‡The diagnosis of 'oligoastrocytoma, NOS' or 'anaplastic oligoastrocytoma, NOS' is discouraged in the 2016 WHO classification of gliomas[16]: oligoastrocytic (mixed) gliomas should be assigned either to an astrocytic or an oligodendroglial tumour entity via appropriate molecular testing for *IDH1/2* mutation and 1p/19q codeletion. §The pilomyxoid astrocytoma variant is not assigned to a definite WHO grade.

IDH mutation (*IDH1* or *IDH2* mutation) is more common in WHO grade II and III gliomas (60-80%) than in WHO IV glioblastoma (5-10%) [17][18]. IDH-mutation is among the earliest genetic aberrations that occur during the development of glioma [19]. Theses mutation have been identified as driver genes in low grade gliomas and secondary GBMs, but not primary GBM [19][20][21]. Findings in mice indicate that IDH mutation alone is not sufficient for tumourigenesis [22]. The exact mechanism by which IDH mutations contribute to glioma progression remains to be established, but could result from metabolic changes [23]. The association of IDH mutated gliomas with a glioma CpG island methylator phenotype (G-CIMP)

[24], suggest that progression in glioma is driven by large scale epigenetic changes (Figure 1.4)

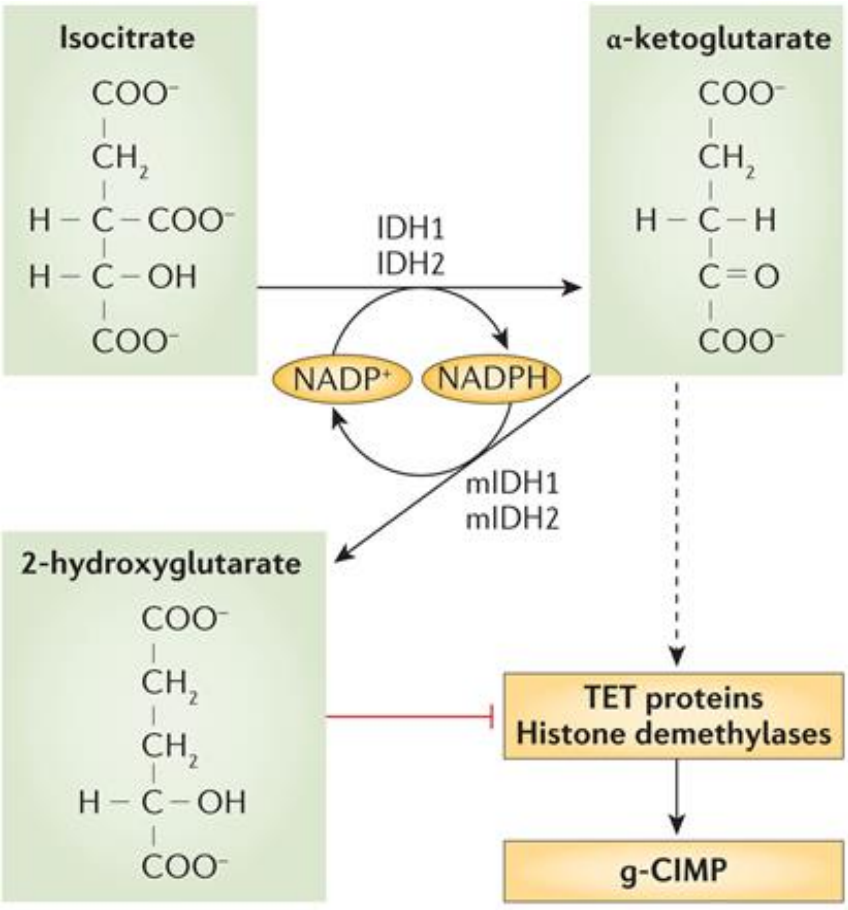


Figure 1.4 Biochemical consequences of glioma-associated isocitrate dehydrogenase mutations. Taken from [8].

The co-deletion of chromosomes arms 1p and 19q is caused by an unbalanced (1;19)(q10;p10) translocation [25]. IDH-mutation associated with 1p/19q co-deletion is the genetically signature of oligodendroglioma tumours.

TERT promoter mutation is associated with the majority of glioma with IDH mutation and 1p/19q co-deletion [20][26][27]. In addition oligodendroglial tumours have been shown to contain *FUBP1* and *CIC* mutation in more than one and two thirds of patients, respectively [28].

Astrocytomas as well as secondary GBMs (which progress from astrocytomas) commonly contain mutations in *TP53* and *ATRX*. These mutation are mutually exclusive with 1p/19q co-deletion suggesting that following IDH-mutation, acquisition of either 1p/19q co-deletion or *TP53*/*ATRX* mutation determines differentiation along the oligodendroglial or astrocytic lineages respectively [20]. The remaining subset (20%) of low-grade gliomas that do not contain IDH mutations are typically grade III and are genetically and clinically similar to primary GBMs [20].

EGFR amplification is detectable in about 40% of IDH-wild-type glioblastomas, with half of these tumours also harbouring a genetic rearrangement that results in deletion of EGFR exons 2–7 [29] referred to as *EGFRvIII* [30]. Additionally, PI(3)K pathway components are often mutated, and *CDKN2A* and *NF1* tumour suppressor genes are commonly deleted [20][31].

1.2.1 Molecular model of glioma development

In 2015, Eckel-Passow *et al* [32] developed a classification that stratified gliomas into five subtypes based on combinations of IDH mutation, *TERT* promoter mutation and 1p/19q co-deletion [32]. The different groups were found to be associated with distinct tumour alterations, age of diagnosis distributions and survival (Table 1.5 and Figure 1.5). In both low grade and GBM tumours patients the *TERT* promoter mutation only group had the poorest overall survival (OS) while in low grade glioma triple-positive tumours and gliomas with *TERT* and IDH mutations had a better survival than patients with triple-negative gliomas [32].

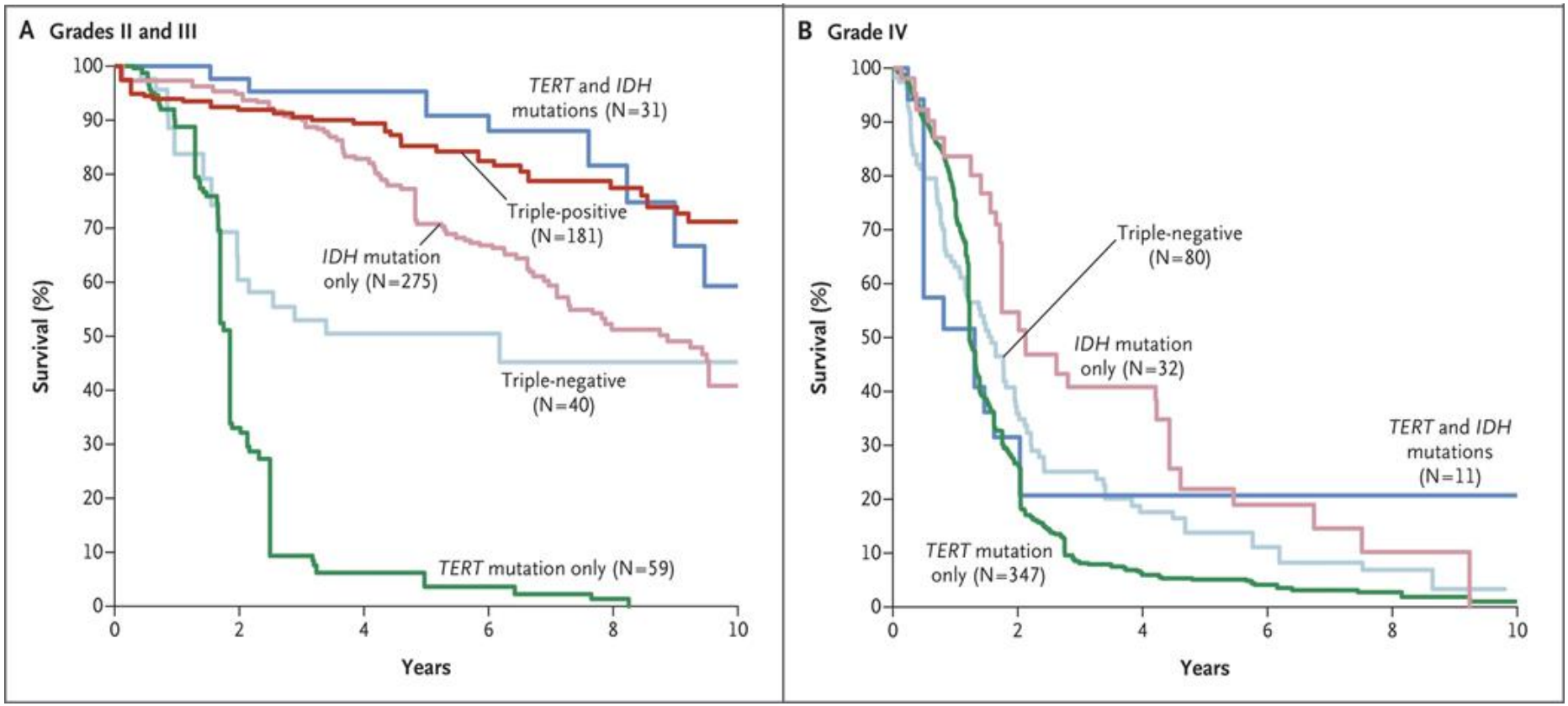


Figure 1.5 Adjusted estimates of overall survival in the glioma molecular groups. Taken from [32]. Overall Kaplan-Meier survival estimates were adjusted for sex and age at diagnosis (on the basis of the 2010 US white population) with the use of the reweighted (direct adjustment) method. Because there was only one triple-positive case among patients with grade IV gliomas, this group was not included in Panel B.

	Molecular classification groupings				
Feature	**Triple positive**	***TERT* and IDH**	**IDH**	**Triple negative**	***TERT***
Grouping alterations	IDH, 1p/19q, and *TERT*	*TERT* and IDH	IDH	-	*TERT*
Histology	Oligodendroglioma/ oligoastroctyoma	Astroctyomas/ oligoastrocytoma	GBM (67%)	GBM (85%) + astrocytoma	No specific association
Mean age at diagnosis	44 years	46 years	37 years	50 years	59 years
Grade II/III (615)	29%	5%	45%	7%	10%
Grade IV (472)	<1%	2%	7%	17%	74%
Common acquired mutations	*CIC*, *FUBP1*, *NOTCH1*, either *PIK3CA* or *PIK3R1*	*TP53*	*TP53* and *ATRX*	*EGFR*, *PTEN* and *NF1*	*EGFR*, *EGFRvIII, PTEN, NF1, RB1,* either *PIK3CA* or *PIK3R1*
Common acquired copy-number alterations	Chr 4 loss, hemizygous *CDKN2A/B* loss "genomically quiet"	Chr 7 gain, 8q24 (*MYC*) duplication, homozygous *CDKN2A/B* loss, *PTEN* deletion	Chr 7q duplication, 8q24 (*MYC*) duplication, hemizygous *CDKN2A/B* loss, 19q deletion	Similar to *TERT* mutation only (at lower prevalence)	Chr 4 loss, chr 7 gain, chr 19 gain, *EGFR* amplification, *CDKN2A/B* homozygous loss, *PTEN* deletion

Table 1.5 Summary of somatic alterations in adult gliomas. Adapted from [32]. IDH indicates mutation in either *IDH1* or *IDH2*. *TERT* indicates *TERT* promoter mutation. Chr, chromosome.

1.3. Clinical and biological aspects of glioma

1.3.1 Glioma origins

The cell of origin for glioma has been an issue for discussion, with evidence pointing to neural stem cells (NSCs), or NSC-derived astrocytes or oligodendrocyte precursor cells (OPCs). Consideration of cell of origin suggests that glioma formation may result from acquisition of mutations in a variety of neural and glial cell backgrounds.

For example, GBMs have further been sub-classified based on gene expression signatures into classical, mesenchymal, proneural and neural subtypes [33]. Moreover, further subclasses on the basis of microRNA expression resemble radial glia, oligoneuronal precursors, neuronal precursors, neuroepithelial/neural crest precursors or astrocyte precursors [34].

1.3.2 Prognosis of glioma

Prognosis is dependent on both grade and molecular profile: diffuse gliomas are divided into three prognostic molecular subgroups: the IDH wild type have the poorest outcome (median OS is 2 years for grade 3), the IDH mutation and 1p/19q co-deleted gliomas have the best survival (median OS >14 years for grade 3) and the IDH mutated non co-deleted (median OS 5-7 years for grade 3). Outcome is also dependent on age, and performance status.

1.3.3 Treatments of glioma

Gliomas grade III and IV are typically treated by surgical resection (if possible) followed by radiotherapy and chemotherapy. Alkylating agents, notably nitrosourea and temozolomide, have shown benefits on patient survival particularly in tumours with IDH mutation and/or with *MGMT* promoter methylation [35][36]. Grade IV (ie glioblastomas) are treated with radiotherapy and concomitant and adjuvant temozolomide [37]. In grade III gliomas, the modality and type of chemotherapy is dependent on genomic profile: IDH wild type grade III are assimilated to GBM (see above), IDH mutated co deleted are treated with radiotherapy and adjuvant nitrosourea based chemotherapy (PCV), IDH mutated non co-deleted are treated with radiotherapy and adjuvant chemotherapy (PCV or TMZ) [38][39][40][41]. Management of grade II gliomas is based on surgical resection which may be iterative, with wait and see periods, chemotherapy, and radiotherapy associated with adjuvant chemotherapy in case of “high risk” grade II glioma [42].

There is no standard of treatment at recurrence. Targeted therapies, antiangiogenic therapies [4][43][44], and immunotherapies have been disappointing so far. While, targeting EGFR initially appeared to be an attractive therapeutic strategy in GBM tumours, clinical effectiveness has so far been limited by both upfront and acquired drug resistance [45]. A vaccine targeting the most common IDH1 alteration (p.Arg132His) has recently been demonstrated to introduce anti-tumour immunity and has been proposed as a viable future therapy for tumours with this mutation [46].

1.4. Genetic architecture of susceptibility to cancer

1.4.1 Overview

Genetic susceptibility, also called genetic predisposition or genetic risk, refers to the increased risk of developing a particular disease based on a person's germline DNA. The two- to three-fold familial risks associated with glioma and other cancers are compatible with a range of effect sizes and frequencies of predisposition alleles observed in the population. The composition of risk alleles for a given disease is typically described as the genomic architecture of disease susceptibility (Figure 1.6). More than 40 years ago, Anderson [47] stated that the magnitude of these familial risks seen for almost all cancers was not indicative of strong genetic effects but instead suggested a mechanism involving many genes with smaller effect acting in concert with environmental or non-genetic factors with larger and more important effects [47].

1.4.2 Multi-locus/multi-allele hypothesis

In terms of evidence to validate these models, a number of rare high penetrance cancer susceptibility genes were successfully identified by linkage studies of highly selected families across 1980s-2000s, hence validating the “multi-locus/multi-allele” model. Examples of these include most of the currently known high-penetrance susceptibility genes, for example *BRCA1* and *BRCA2* in breast cancer, *MLH1* in colorectal cancer and *CDKN2A* in melanoma [48][49][50][51] (Figure 1.6). In recent years the search for additional rare high penetrance mutations has continued, using High-Throughput Sequencing (HTS) techniques, which offer greater resolution than genetic linkage. In fact the increasing cost effectiveness, quality, throughput and bioinformatics resources supporting HTS are enabling comprehensive studies of the entire exome or genome in large patient cohorts.

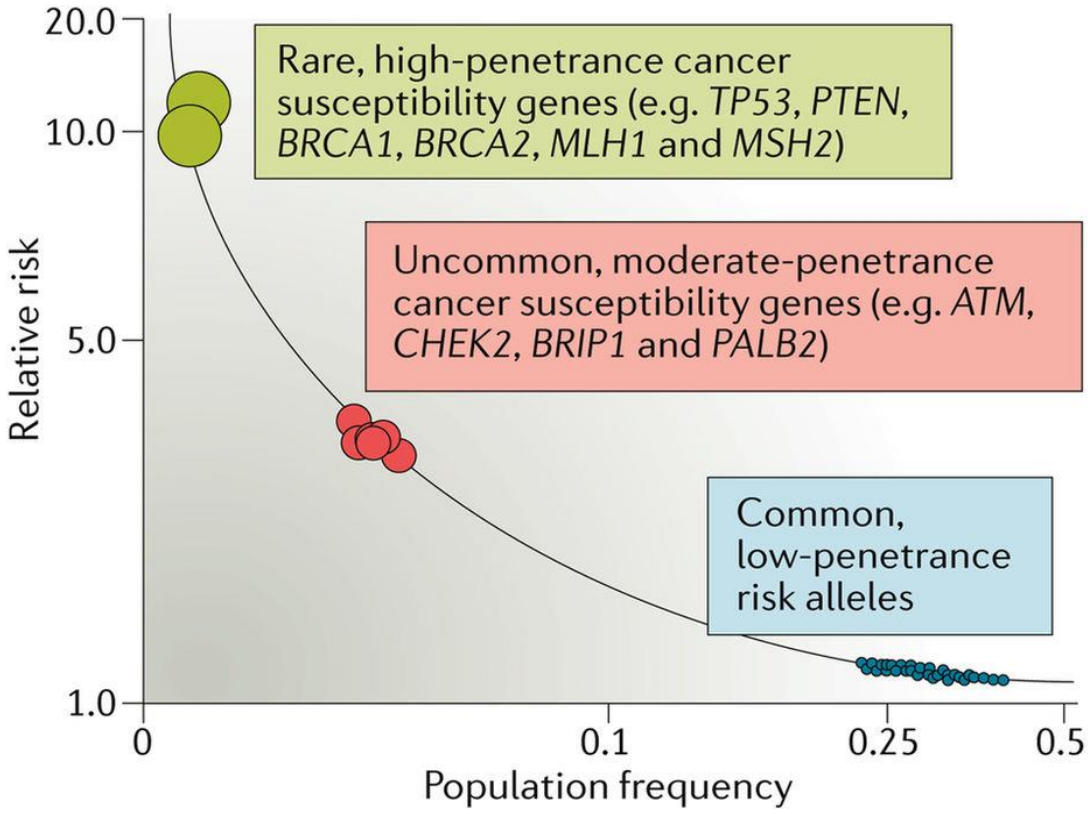


Figure 1.6 Genetic architecture of cancer risk. Taken from [52].This graph depicts the low relative risks (RRs) associated with common, low-penetrance genetic variants (such as single nucleotide polymorphisms (SNPs) identified in genome-wide association studies (GWAS)); moderate RRs associated with uncommon, moderate-penetrance genetic variants (such as ataxia telangiectasia mutated (*ATM*) and checkpoint kinase 2 (*CHEK2*)); and higher RRs associated with rare, high-penetrance genetic variants (such as pathogenic mutations in *BRCA1* and *BRCA2* associated with hereditary breast and ovarian cancer). *BRIP1*, *BRCA1* interacting protein C-terminal helicase 1; *MLH1*, mutL homologue 1; *MSH2*, mutS homologue 2; *PALB2*, partner and localizer of *BRCA2*.

Increased risk of glioma is now recognised to be associated with a number of these Mendelian cancer predisposition syndromes, notable neurofibromatosis (NF1 and NF2), Li-Fraumeni and Turcot's [53][54][55][56][57][58][59][60][61][62]. Additionally, germline mutation of *CDKN2A* has been reported to be a cause of the astrocytoma-melanoma syndrome [63][64]. A number of these cancer syndromes are now recognised to be associated with an increased risk of glioma (Table 6).

Syndrome	Inheritance	Gene	Location	Tumours	Reference
Li-Fraumeni	Dominant	*TP53*	17p13.1	Sarcoma, breast, brain, leukaemia, adrenocortical carcinoma	[53][54][55] [57][58][59]
Turcot's type 1 (hereditary nonpolyposis cancer syndrome)	Dominant/ Recessive	*MLH1*, *MSH2*, *MSH6*, *PMS2*	3p22.2, 2p16.3, 2p21, 7p22.1	Colorectal carcinoma, glioma	[62][65]
Turcot's type 2	Dominant	*APC*	5q22.2	Colorectal carcinoma, glioma	[65]
Neurofibromatosis type 1 (NF1)	Dominant	*NF1*	17q11.2	Glioma, neurofibroma, pheochromocytoma, meningioma, schwannoma	[61]
Neurofibromatosis type 2 (NF2)	Dominant	*NF2*	22q12.2	Bilateral acoustic schwannoma, meningioma, glioma, neurofibroma, ependymoma	[60]
Melanoma-astrocytoma	Dominant	*CDKN2A*	9p21.3	Melanoma, astrocytoma	[63][64]
BRCA	Dominant	*BRCA1*, *BRCA2*	17q21.31, 13q13.1	Breast, ovarian, prostatic, pancreatic, glioma	[66]

Table 1.6 Inherited cancer syndromes associated with high risk of glioma.

Inherited mutations in these genes are typically very rare at a population level and are consistent with Knudson's "two-hit" hypothesis of cancer development [67]. Collectively however these syndromes are rare and account for little of the two-fold of familial risk of glioma in the population [68].

1.4.3 More recent models of genetic susceptibility to glioma

The identification of susceptibility genes to glioma through linkage analysis has been limited. In a segregation study of four Finnish families with two or more gliomas non-significant linkage was attained at 15q23-q26.3 [69]. In 2011, linkage analysis by Shete *et al* using high-density SNP arrays of 46 US families provided suggestive linkage at 17q12-q21.32 [70]. however replication genotyping of an independent series of 29 families has failed to provide evidence for causal basis of the linkage signal [71][72].

Linkage studies are not powered to detect moderate and low-penetrance alleles conferring more modest risk of disease, which are unlikely to cause multiple cases in families [73]. Statistical modelling of glioma has suggested that much of the heritable risk is polygenic and enshrined in common risk variants, involving the co-inheritance of multiple genetic factors (Figure1.7).

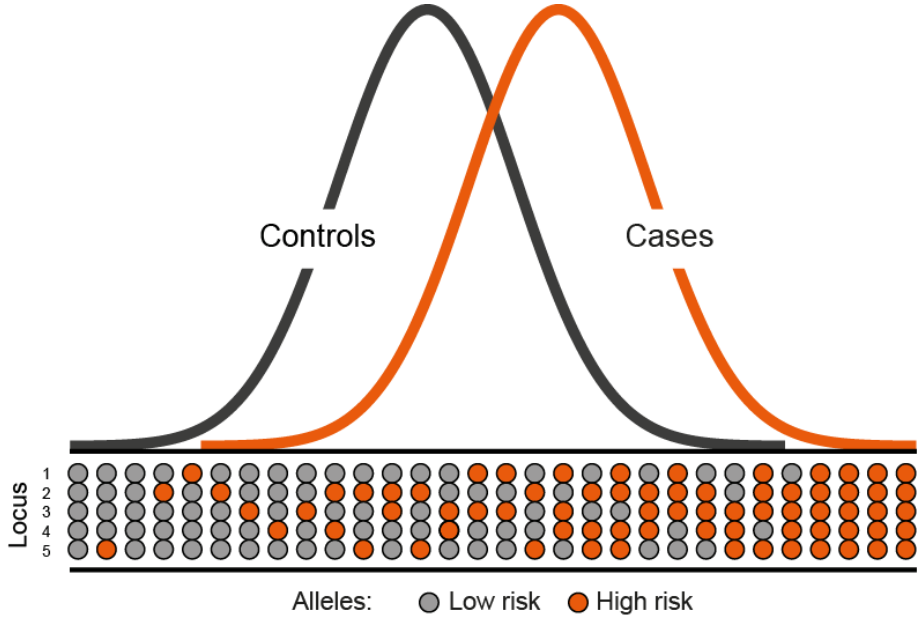


Figure 1.7 Polygenic model of disease susceptibility. The distribution of risk alleles in both cases and controls follows a normal distribution. However, cases have a shift towards a higher number of risk alleles.

1.4.3.1 Rare, moderately-penetrant disease-causing variants

The "rare variant" hypothesis suggests that a proportion of the remaining heritability of glioma could be due to the combined effect of rare, moderately-penetrant risk alleles [74]. This hypothesis suggests that such variants act independently and confer modest but detectable increases in risk. Studies of rare variants through sequencing of candidate genes in glioma cases and controls have failed to identify genes associated with glioma. A recent study of 1,662 cases and 1,301 controls failed to replicate 52 variants previously identified by candidate gene studies [75].

Thus in summary, both models of genetic susceptibility have proven to be correct and across all tumour types wide continuums of differing genomic architectures have been observed. For example prostate cancer has a genetic susceptibility predominantly based on common low risk alleles, whereas in ovarian cancer a very substantial proportion is accounted for by rare high penetrance mutations, with the majority of other cancers somewhere in between.

1.5 Identification of common low-penetrance allele

The “common disease, common variant” hypothesis posits that a substantial proportion of the genetic risk of common diseases can be accounted for by the action of multiple low-penetrance alleles that have a relatively high population frequency [76]. While each variant may individually cause very modest increases in risk, collectively they could underscore a substantial proportion of disease genetic risk. These alleles are highly unlikely to cause multiple cases in families and therefore would have eluded prior detection through linkage studies [73].

1.5.1 Genome-wide association studies

Genome-wide association studies (GWAS) emerged in 2005 as a powerful tool for the identification of common genetic markers associated with disease risk. A marker allele is associated with disease if one allele is found significantly more frequently in cases than in disease-free controls. Single nucleotide polymorphisms (SNPs), the marker variants generally used for association studies, are common in the human genome and account for over 90% of all sequence variation [77]. Adjacent SNPs in the genome are not randomly inherited; they are strongly correlated and likely to co-segregate together in a haplotype. The strong correlation of genetically nearby SNPs is termed linkage disequilibrium (LD); the strength of which decreases rapidly with increasing genomic distance [76]. The nature of this haplotype structure allows certain SNPs across the genome to be selected as “tagging SNPs”, which are expected to capture the majority of sequence variation across a given region (Figure 1.7).

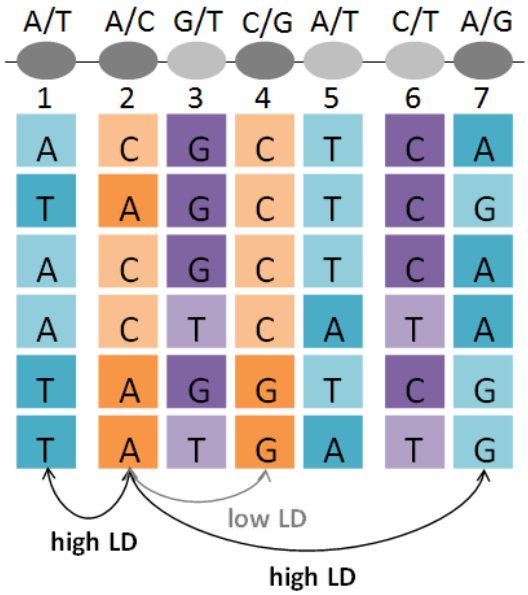


Figure 1.8 Tagging SNPs. It is possible to identify genetic variation without genotyping every SNP in a chromosomal region. For example through genotyping SNP 2 it is possible to infer the genotypes of SNP 1, SNP 4 and SNP 7

GWAS arrays typically directly genotype 300,000-1,000,000 tagging SNPs (tag SNPs) across the genome simultaneously. They allow identification of regions associated with a disease or trait (termed “risk loci”) without prior knowledge of genomic location or function. The power of an association study is the likelihood of detecting a true genetic association. The sample size required to yield sufficient power is dependent on the frequency of the disease allele under study, the effect size of the variant on the trait of interest and the significance threshold required to declare a true association. The main advantage of the association design over linkage studies is that single cases are much more readily available than large extended pedigrees. This allows for much larger sample sizes and therefore greater power to detect variants with small effects. Additionally, multiple studies can be combined in a meta-analysis resulting in further increases in power. An alternative approach is to select cases that are genetically enriched for disease, such as those with a family history or early age of disease onset [78]. Since 2005 GWAS have been successfully applied across a broad range of disease types, and the NHGRI-EBI catalogue of published GWAS [79] currently lists over 13,000 published disease associating SNPs. GWAS have also been extensively applied to cancer, with disease-associated SNPs identified for the majority of tumour types.

1.5.2 Imputation

Risk SNPs identified through GWAS represent proxies for the association signal but are not themselves necessarily the functional or causative variant at the risk locus. The causative SNP in the association is likely to be correlated with the sentinel tag SNP at the GWAS association peak while not being directly genotyped on a GWAS array. These SNPs can be recovered and the disease risk locus fine-mapped through imputation, which is a computational method that aims to predict the likely genotypes at un-genotyped loci across the genome. This method makes use of the information provided by haplotypes in a reference panel of sequenced samples such as the 1000 Genomes project [80] and UK10K project [81] (Figure 1.10). Additionally, a genome-wide approach to imputation can be used to identify new regions of association at variants that are incompletely tagged by GWAS tag SNPs or at insertion/deletions (indels) that are not fully captured by GWAS arrays. This genome-wide imputation approach has been successfully implemented in a recent study which identified rare variants in *BRCA2* and *CHEK2* with a large effect on lung cancer risk (OR>2.4) [82]. Imputation is limited by the choice of reference panel, the quality and size of which can impact on imputation fidelity. Therefore robust methodological practices are required to avoid erroneous associations, however when conducted correctly imputation can be a valuable tool in risk loci discovery [83].

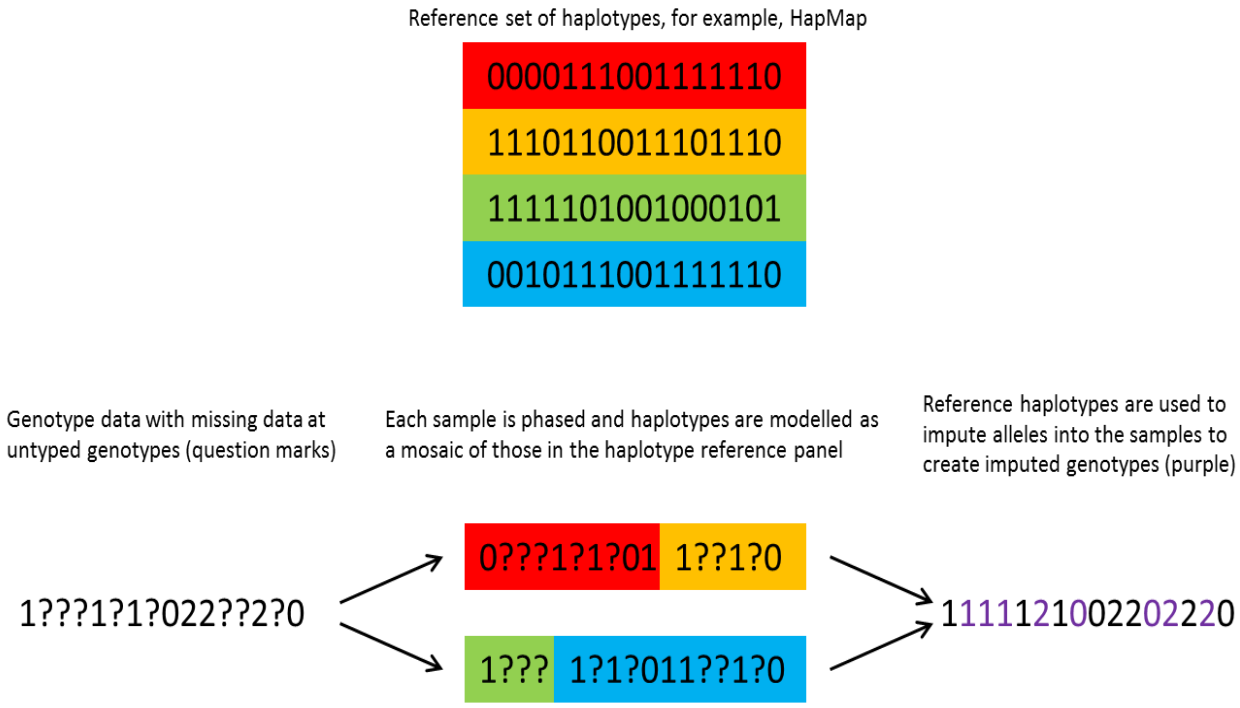


Figure 1.9 Overview of Imputation. Adapted from [84]**.**

1.6. Genetic susceptibility to glioma

1.6.1 Association studies in glioma

Outside of the work detailed in this thesis, fourteen glioma susceptibility loci have been identified in European populations (Table 1.7
Table) [83][85][86][87][88][89][90][91].In 2009 Shete *et al* carried out the first glioma GWAS [85] that comprised a discovery case-control series of UK and European-American individuals (totalling 1,878 cases and 3,670 controls) and replication series of French, German and Swedish individuals (totalling 2,545 cases and 2,953 controls). This study identified five susceptibility loci at 5p15.33, 8q24.21, 9p21.3, 11q23.3 and 20q13.33 [85]. The loci at 9p21.3 and 20q13.33 were independently confirmed by Wrensch *et al* [89] in a contemporaneous study of European-American individuals comprising a discovery phase of 692 high-grade glioma cases and 3,992 controls as well as a replication phase of 176 high-grade glioma cases and 174 controls [89]. In 2011, a GWAS carried out by Sanson *et al* [87], making use of data from the UK and European-American studies previously reported by Shete *et al* [80] as well as two additional case-control series from France and Germany (totalling 4,147 cases and 7,435 controls). This study identified 7p11.2 as a susceptibility locus for glioma, which contained two statistically independent SNP associations with glioma risk [83]. In 2014 a GWAS was carried out by Walsh *et al* [90] comprising a UK and European-American discovery series of 1,013 high-grade glioma cases and 6,595 controls (in part overlapping with the study of Wrensch *et al* [89]), as well as a European-American replication series of 631 GBM cases and 1,141 controls. This study reported a novel glioma risk locus at 3q26.2 (near *TERC*) [90].

Most recently Kinnersley *et al* [92] performed a meta-analysis of GWAS data previously generated on four non-overlapping case–control series of Northern European ancestry, totalling 4,147 cases and 7,435 controls (comprising the previous data; the UK-GWAS [93], the French-GWAS [87], the German-GWAS [87] and the US-GWAS [85]). The study led to the identification of additional susceptibility loci at 12q23.33, 10q25.2, 11q23.2, 12q21.2 and 15q24.2 and taking the total count of risk loci to 12 [92]. Intriguingly across all of the four GWAS data sets the authors did not replicate the association between rs1920116 (near *TERC*) at 3q26.2 and risk of high-grade glioma recently reported by Walsh *et al*[91].

In addition to this, a sequence-based association study in the Icelandic population led to the discovery of 17p13.1 (*TP53*) as a risk locus for several cancers including glioma. The association with glioma was confirmed in an independent European study [83]. To refine the association signal at 8q24.21 in glioma, the region was fine-mapped by sequencing as well as statistical imputation of pre-existing GWAS datasets. This led to the identification of rs55705857 as being responsible for the 8q24.21 glioma association, with the SNP exhibiting a much larger effect size than the initial GWAS tagSNPs and being highly restricted to low-grade IDH mutated glioma [86][91].

Table 1.7 Glioma risk loci identified outside of the work detailed in this thesis. Odds ratios derived with respect to the risk allele, highlighted in bold. Risk allele frequencies are according to the European population in 1000 Genomes Project. RAF, risk allele frequency. *Associations are statistically independent.

Locus	SNP	Alleles	RAF	*P*-value	Odd ratio	Reported subtype	Reference
3q26.2	rs1920116	A/**G**	0.710	$8.3x10^{-9}$	1.30	GBM	[90]
5p15.33	rs2736100	C/**T**	0.499	$1.4x10^{-15}$	1.39	GBM	[85]
7p11.2	rs2252586	T/**G**	0.281	$2.09x10^{-8}$	1.18	GBM	[87]
7p11.2	rs11979158	**A**/G	0.83	$7.03x10^{-8}$	1.23	GBM	[87]
8q24.21	rs55705857	A/**G**	0.057	$2.3x10^{-94}$	4.3	Non-GBM	[85]
9p21.3	rs4977756	T/**G**	0.40	$1.41x10^{-12}$	1.22	GBM	[85][89]
10q25.2	rs11196067	A/**T**	0.41	$4.32x10^{-8}$	1.09	Non-GBM	[92]
11q23.2	rs648044	**A**/G	0.38	$6.26x10^{-11}$	1.25	Non-GBM	[92]
11q23.3	rs498872	**G**/C	0.307	$1.07x10^{-8}$	1.18	Non-GBM	[85]
12q21.2	rs12230172	**G**/A	0.45	$7.35x10^{-11}$	1.00	Non-GBM	[92]
12q23.3	rs3851634	T/**C**	0.27	$3.02x10^{-9}$	1.00	GBM	[92]
15q24.2	rs1801591	G/**A**	0.10	$5.71x10^{-9}$	1.36	Non-GBM	[92]
17p13.1	rs78378222	T/**G**	0.01	$6.86x10^{-24}$	3.74	All	[88]
20q13.33	rs6010620	T/**C**	0.80	$4.7x10^{-19}$	1.56	GBM	[85][89]

1.6.2 Perspectives from glioma GWAS

The glioma GWAS risk loci so far provide support for a polygenic model of disease susceptibility. Aside from the fine-mapped associations at 8q24.21 and 17p13.1, the glioma GWAS SNPs identified so far are relatively common (European MAF>0.2) and have modest effect sizes (1.18<OR<1.56). The loci implicate genes known to be important in glioma and cancer biology, for example *EGFR* at 7p11.2, *CDKN2A/B* at 9p21.3, *MYC* at 8q24.21, *TP53* at 17p13.1. Additionally, through identification of risk loci at *TERC* (3q26.2), *TERT* (5p15.33) and *RTEL1* (20q13.33) GWAS associations reveal telomere maintenance as an important feature in glioma progression.

Recent methods allow the estimation of SNP-based heritability from GWAS datasets and have been applied to a variety of complex traits including cancer [94][95]. Analysis performed by Kinnersley *et al* [96] shows that substantial proportion (approximately 25%) of the heritability of developing glioma can be ascribed to common genetic variation. These results suggest that most of the heritable risk attributable to common genetic variants remains to be identified and that further GWAS efforts will lead to the identification of additional risk loci.

1.7 .Strategies to identify novel glioma susceptibility alleles

Collectively, the architecture of glioma predisposition encompasses a small proportion of high-penetrance single gene mutations as well as the combined effect of multiple common low-penetrance polymorphisms (Figure 1.11).

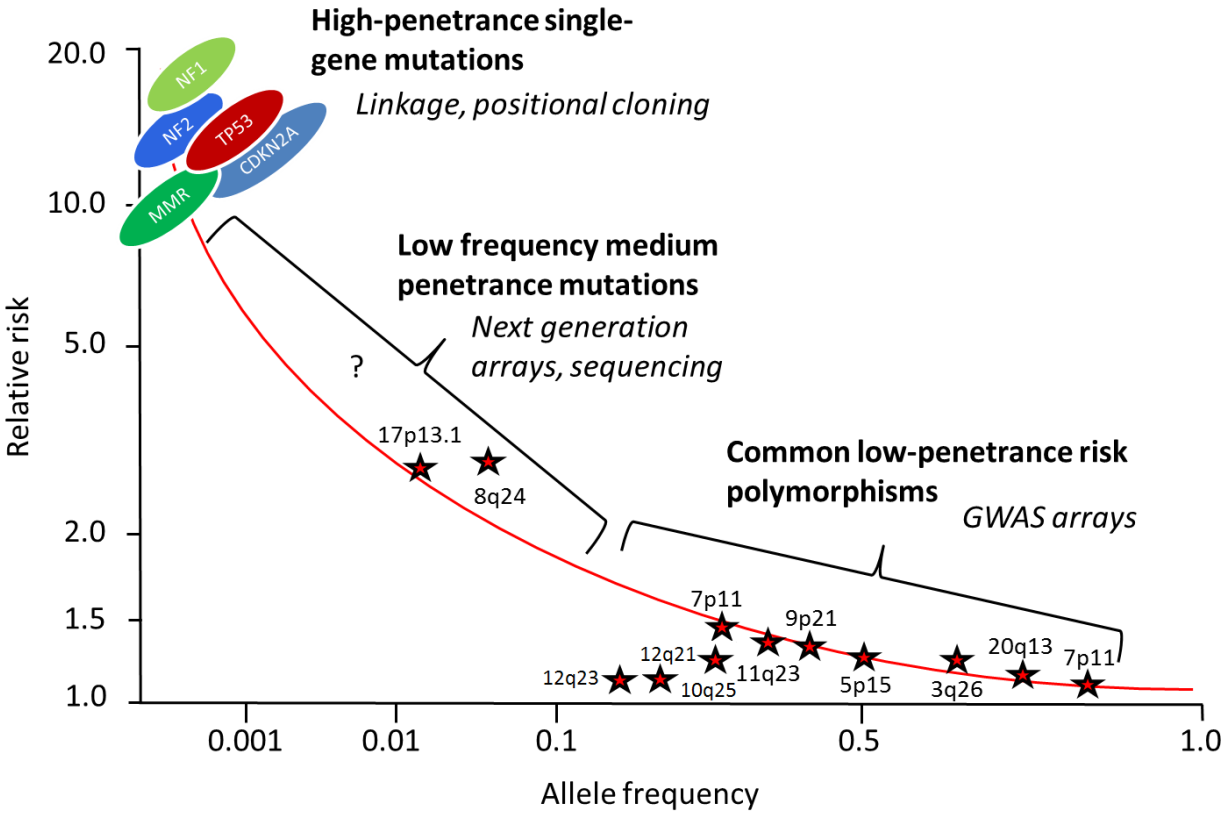


Figure 1.10 Architecture of glioma predisposition. Graph of allele frequency against relative risk for glioma risk variants. Highlighted are the three major classes of risk allele, and the methods used to identify them. GWAS, genome-wide association study; MMR, mismatch repair.

1.7.1 GWAS, Imputation and meta-analysis

Given that many GWAS exhibit long tails of associations with small effect sizes, much of the underlying genetic architecture of cancer susceptibility may be due to a large number of common susceptibility alleles, which individually account for a small proportion of the inherited risk [97]. New susceptibility loci are likely to be identified through imputation using larger reference panels and generation of larger GWAS [96], involving large-scale meta-analysis and replication. Additionally, given that many of the currently identified glioma GWAS risk SNP show a degree of specificity to glioma histological subtypes it is therefore likely

that further studies combining pre-existing and additional GWAS datasets with subtype data will identify further glioma risk loci.

1.7.2 Next-generation arrays

Low-frequency risk variants (MAF ~1%) are hypothesised to contribute significantly to the risk of glioma. While current GWAS arrays are designed to capture common risk variants, they do not adequately capture variation at MAF < 5% [98][99]. Using pools of reference haplotypes such as that provided by the 1000 genomes project and UK10K project, whole-genome imputation may extend the frequency range for which associations can be detected from existing datasets [80][81].

Recently there has been development of new disease specific arrays such as the Illumina OncoArray which contains approximately 533,000 markers with nearly 50% of the markers selected as a GWAS backbone (Illumina HumanCore). These markers were selected to tag the large majority of known common variants, via imputation. The remaining markers were selected from the disease consortia representing the main cancer sites [100]. These arrays enable the identification of new susceptibility loci, performing fine mapping of new or known loci associated with either single or multiple cancers, assessing the degree of overlap in cancer causation and pleiotropic effects of loci that have been identified for disease-specific risk, and jointly model genetic, environmental and lifestyle related exposures.

1.7.3 Functional annotation of risk SNPs

Many functional classes of genetic variation have been implicated as the basis of risk loci identified from GWAS (Figure 1.12). A small number of the loci identified from cancer GWAS directly impact on the amino acid sequence of the expressed protein. The mechanistic interpretation of such variants is presumed to be relatively simple, owing to the implied direct relationship between genotype and function [82][101]. Similarly, a direct relationship can be inferred for those variants affecting RNA processing [88] and those affecting splice sites such as the inhibitory splice isoform [102]. However, it is possible that coding variants could have more subtle effects that do not necessarily involve disrupting protein function but instead involve tagging functional non-coding variants.

The majority of risk loci map to non-coding regions of the genome (for example, to gene introns or promoters and intergenic regions). Risk loci identified from GWAS have been

demonstrated to map to genomic regions of cell-type-specific active chromatin and show an over-representation of expression quantitative trait loci, methylation quantitative trait loci [103][104] and transcription factor (TF) binding [105]. Chromatin conformation studies have helped link regulatory regions, which SNPs identified by GWAS localise to, with their respective target genes [106][107].

To date, relatively few risk loci have been comprehensively studied. However, insights into the genetic and biological basis of cancer susceptibility mediated through common variation are emerging.

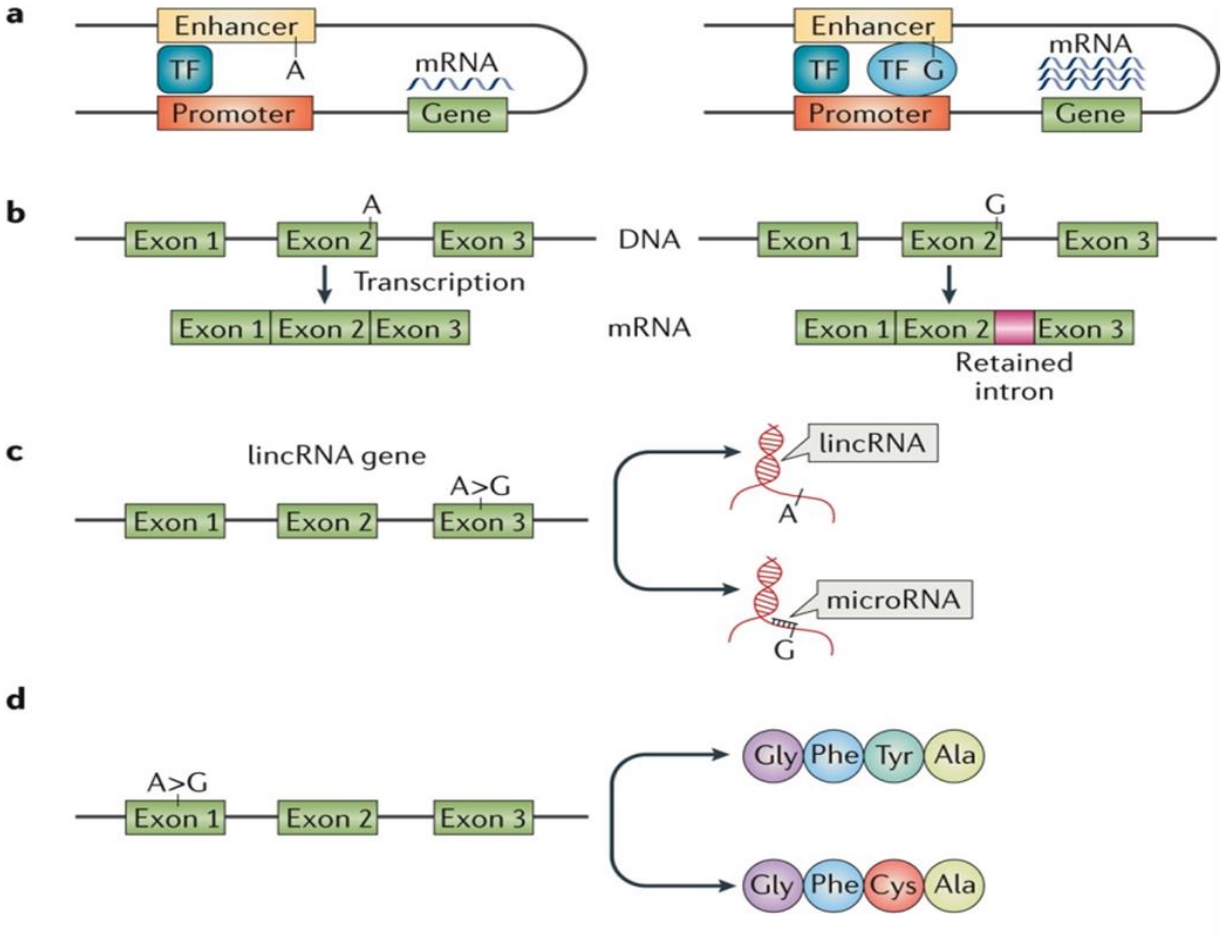


Figure 1.11 Potential molecular mechanisms by which risk polymorphisms mediate cancer susceptibility. Taken from [52] The A>G polymorphism is affecting gene transcription by altering transcription factor (TF) binding through a looping promoter–enhancer-complex interaction (part a); the A>G polymorphism occurs at an intron splice site and results in intron retention, thereby affecting mRNA processing (for example, by modulating splicing and poly-adenylation) (part b); the A>G polymorphism leads to the generation of a novel microRNA binding site on the large intergenic non-coding RNA (lincRNA) (part c); and the A>G polymorphism affects the protein sequence by causing an amino acid substitution of tyrosine to cytosine (part d). GWAS, genome-wide association studies; SNPs, single nucleotide polymorphisms.

1.8. Study aims and scope of enquiry

The identification of the discussed risk loci to glioma is consistent with architecture of inherited predisposition to glioma involving both high-penetrance mutations in single genes and multiple low-penetrance risk SNPs. However, the estimation of SNP-based heritability (approximately 25%), suggest that most of the heritable risk attributable to common genetic variants remains to be identified. In addition, the discovery of a new recurrent somatically mutated gene help to a better classification of glioma entities and will offer the potential to support drug development and advance precision medicine for these tumours.

The work detailed in this thesis was therefore aimed at gaining further insight into these questions, studying both the inherited genetic basis and somatic mutational features of glioma, making use of currently available technologies and analytical methods. It is anticipated that this research will lwad to increased insight into the biological and genetic basis of glioma development, with potential to support the development of improved treatment strategies and predictive biomarkers of therapeutic outcome.

Specifically:

- **Chapter 3** reports on the identification of novel common germline risk loci for glioma
- **Chapter 4** reports on the investigation of the relationship between risk SNPs and glioma molecular subtype.
- **Chapter 5** reports the results of somatic whole-exome sequencing of a series of glioma subgroup (anaplastic oligodendroglioma)

CHAPTER 2

Materials and methods

2.1. Subjects and samples

The case/control samples and datasets used in this thesis can be described as follows:

2.1.2 Germline gliomas cases controls samples

Here I describe the GWAS data from seven studies used in Chapter 3. Cases and controls samples used in Chapter 4 are described in 4.2.1.

GICC GWAS

Studies participating in GICC comprised 5,189 glioma cases and 3,827 controls that were ascertained through centers in the USA, Denmark, Sweden and the UK. Cases had newly diagnosed glioma, and controls had no personal history of central nervous system tumour at the time of ascertainment. Table 2.1 describe the summary characteristics of the GICC sub-studies. Detailed information regarding recruitment protocol is given in Amirian *et al* [108].

UK GWAS

The previously published UK GWAS [85][87][92] was based on 636 cases (401 males; mean age 46 years) of Northern European ancestry who were ascertained through the INTERPHONE study [93]. Individuals from the 1958 Birth Cohort (n = 2,930) served as a source of controls [109].

German GWAS

The German GWAS published in Kinnersley *et al* [92], comprised 880 patients of Northern European ancestry who had undergone surgery for a glioma at the Department of Neurosurgery, University of Bonn Medical Center, between 1996 and 2008. Control subjects were taken from three population studies: KORA (Co-operative Health Research in the Region of Augsburg; n = 488) [110]; POPGEN (Population Genetic Cohort; n = 678) [111] and the Heinz Nixdorf Recall study (n = 380) [112].

MDA GWAS

The MDA GWAS [85] was based on 1,281 cases of Northern European ancestry (786 males; mean age 47 years) who were ascertained through the MD Anderson Cancer Center, Texas, between 1990 and 2008. Individuals from the Cancer Genetic Markers of Susceptibility (CGEMS, n = 2,245) studies served as controls [113][114].

UCSF adult glioma case-control study (SFAGS–GWAS)

The SFAGS-GWAS included participants of the San Francisco Bay Area Adult Glioma Study (AGS). Details of subject recruitment for AGS have been reported previously [32][89][91][115][116]. Briefly cases were adults (>18 years of age) with newly diagnosed, histologically confirmed glioma. Population-based cases who were diagnosed between 1991 and 2009 (series 1–4) and who were residing in the six San Francisco Bay area counties were ascertained using the Cancer Prevention Institute of California's early-case ascertainment system. Clinic-based cases who were diagnosed between 2002 and 2012 (series 3–5) were recruited from the UCSF Neuro-oncology Clinic, regardless of the place of residence. From 1991 to 2010, population-based controls from the same residential area as the population-based cases were identified using random digit-dialing and were frequency matched to population-based cases for age, gender and ethnicity. Between 2010 and 2012, all controls were selected from the UCSF general medicine phlebotomy clinic. Clinic-based controls were matched to clinic-based glioma cases for age, gender and ethnicity. Consenting participants provided blood, buccal and/or saliva specimens, and information, during in-person or telephone interviews. A total of 677 cases and 3,940 controls were used in the current analysis.

GliomaScan GWAS

The previously published GliomaScan GWAS [117] comprise In total 1,653 cases and 2,725 controls were used in the current study.

French GWAS

The French GWAS is detailed in 4.2.1.

	All Glioma						GBM			Non-GBM		
	Pre-QC			**Post QC**			**Post QC**			**Post QC**		
	Total	**Cases**	**Controls**	**Total**	**Cases**	**Controls**	**Total**	**Cases**	**Controls**	**Total**	**Cases**	**Controls**
Baylor College of Medicine	40	40	0	11	11	0	6	6	0	5	5	0
Brigham and Women's Hospital	247	225	22	215	193	22	123	101	22	98	76	22
Columbia	215	64	151	166	40	126	150	24	126	141	15	126
Case Western Reserve University	74	60	14	67	56	11	44	33	11	34	23	11
Denmark	1,054	522	532	1,008	496	512	811	299	512	706	194	512
Duke	876	622	254	782	578	204	627	423	204	338	134	204
Mayo	833	376	457	803	358	445	639	194	445	604	159	445
MD Anderson	1,783	1,505	278	1,140	921	219	571	352	219	774	555	219
Memorial Sloan Kettering	652	283	369	531	239	292	416	124	292	396	104	292
North Shore	306	133	173	264	123	141	217	76	141	187	46	141
Sweden	1,400	476	924	1,356	465	891	1,162	270	891	1,079	188	891
University of California, San Francisco	673	333	340	506	277	229	381	152	229	350	121	229
UK	914	798	116	874	766	108	491	383	108	366	258	108
University of Southern California	297	98	199	135	49	86	115	29	86	105	19	86
GICC	**9,364**	**5,535**	**3,829**	**7,858**	**4,572**	**3,286**	**5,754**	**2,466**	**3,286**	**5,183**	**1,897**	**3,286**

* Israeli samples and whole genome amplified samples were excluded at the initial QC regarding DNA quality control.

Table 2.1 Summary characteristics of the GICC sub-studies.

2.1.3 Anaplastic oligodendroglioma matched tumour/normal samples

Samples were obtained with informed and written consent and the after approval of the institutional review boards (IRBs) study was approved by Comité de Protection des Personnes Ile de France-VI (October 2008) of respective hospitals participating in the Prise en charge des oligodendrogliomes anaplasiques (POLA) network. All patients were aged 18 years or older at diagnosis and tumour histology was centrally reviewed and validated according to World Health Organization (WHO) guidelines [118].

In addition to the datasets generated through the work reported in this thesis, in Chapter 5 I made use of The Cancer Genome Atlas (TCGA) study of low grade glioma as described in 5.2.1

2.2 Molecular methods

2.2.1 Illumina whole-exome sequencing

Whole-exome sequencing was used to generate data analysed in the course of this thesis. Here is a brief description of the sequencing technology

2.2.1.1 Sample and library preparation

DNA was quantified using the Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies). Libraries were generated robotically using the SureSelectXT Automated Human All Exon Target Enrichment for Illumina Paired-End Multiplexed Sequencing (Agilent) as per the manufacturer's recommendations. Libraries were quantified using the Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies) and the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (D-Mark). Average size fragment was determined using a LaChip GX (PerkinElmer) instrument.

2.2.1.2 Target capture

Regions of interest are selected for by a 24hour hybridisation step with biotinylated RNA library baits followed by a cleanup step using magnetic streptavidin beads. The baits can be custom designed using Agilent's SureDesign software. PCR is then used to amplify these regions which are then ready for sequencing. (Figure 1.2)

2.2.1.3 High-throughput sequencing

Finally samples then underwent paired end sequencing using the Ilumina HiSeq2000 platform with a 100-bp read length. The Illumina HiSeq 2000 platform carries out sequencing by synthesis whereby millions of DNA fragment clusters are sequenced in parallel. Briefly, as each deoxynucleotide

triphosphate (dNTP) is added, an attached fluorescently labelled reversible terminator is imaged before being cleaved to allow incorporation of the following base. Incorporation bias is minimised by natural competition generated through the presence of all four possible terminator-bound dNTPs throughout the reaction. Base calls are made directly from the signal intensity during each incorporation cycle.

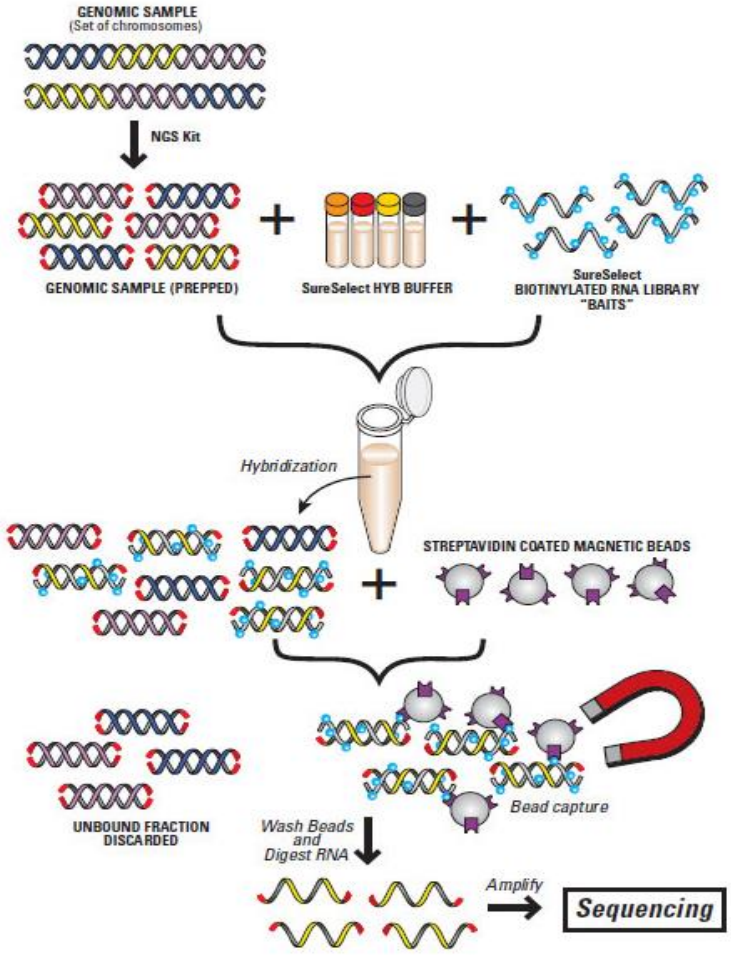


Figure 2.1 SureSelect Target Enrichment System Capture Process. Taken from, agilent exome enrichment kit datasheet**.**

2.2.2 Illumina transcriptome sequencing (RNA-seq)

Illumina transcriptome sequencing was performed on 39 tumours on Chapter 5. Here is a brief description of the used sequencing technology.

2.2.2.1 Sample and library preparation

RNA-seq library construction protocols include similar basic steps, which require elimination of ribosomal RNA (rRNA), reverse transcription of the desired RNA species, fragmentation, adapter ligation, and enrichment (Figure2.2) Extracted RNA from tumours was cleaned using the RNeasy MinElute Cleanup Kit (Qiagen) and the RNA integrity assessed using an Agilent 2100 Bioanalyzer and quantified using a Nanodrop 1000. Libraries for stranded total RNA-sequencing were prepared with Illumina Stranded Total RNA protocol (RS-122-2301). Libraries were assessed by Agilent 2100 Bioanalyzer.

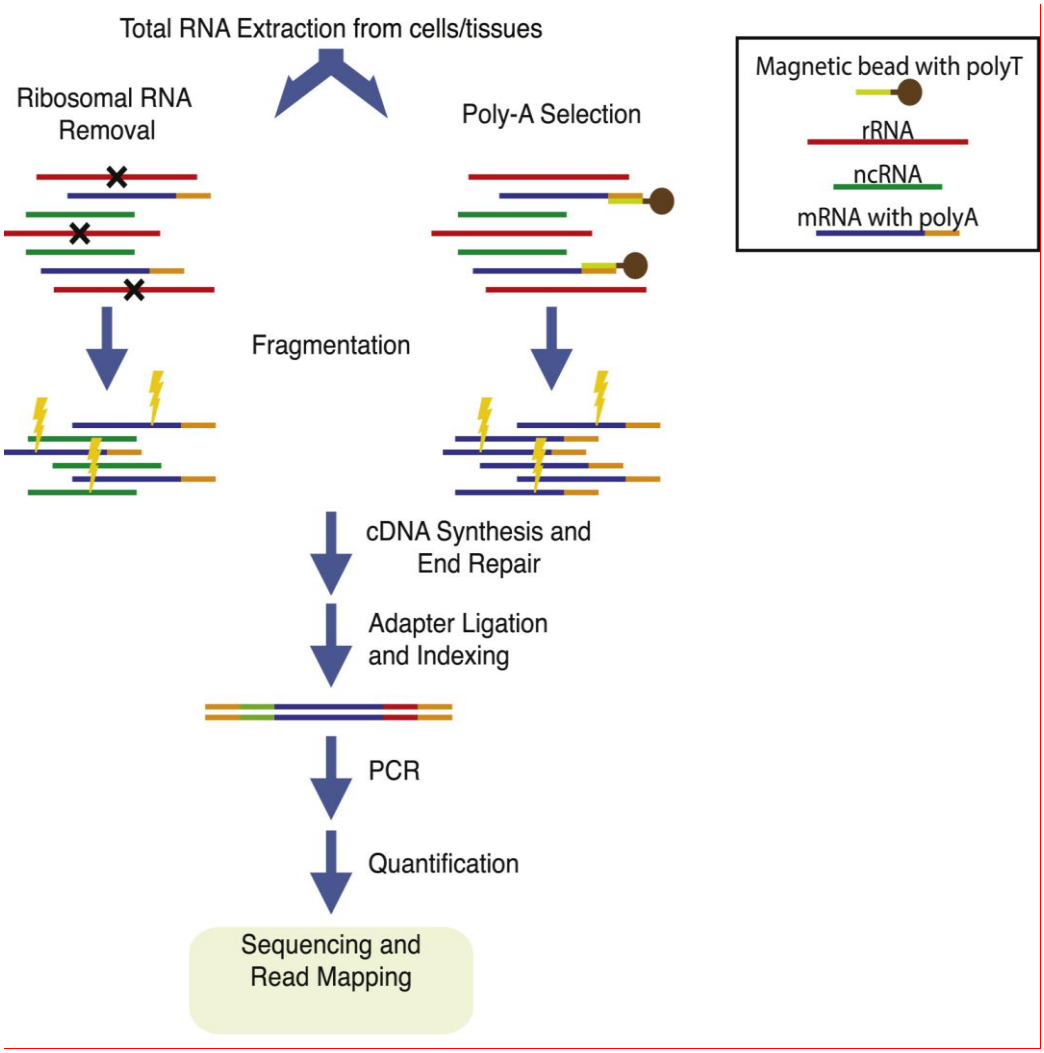


Figure 2.2 RNA sequencing workflow and analysis. Taken from [119].

2.2.2.2 High-throughput sequencing

Sequencing was performed by pooling 4 libraries per lane at a 9pM dilution on an Illumina HiSeq 2000 instrument for 2x 100 cycles using the recommended manufacturer's conditions. PhiX control was added at 1% on each lane.

2.2.3 Genotyping

2.2.3.1 Genome wide array genotyping

Most of the GWAS datasets presented in Chapter 3 and 4 were genotyped on Illumina BeadChip SNP arrays. The GICC GWAS dataset presented in Chapter 3 were genotyped using a custom Infinium OncoArray-500K BeadChip (Oncoarray) from Illumina (Illumina, San Diego, CA, USA), comprising a 250K SNP genome-wide backbone and 250K SNP custom content selected across multiple consortia within COGS (Collaborative Oncological Gene-environment Study). Oncoarray genotyping was conducted in accordance with the manufacturer's (Illumina Inc.).

The principles of BeadChip arrays can be illustrated by the Illumina infinium II assay.

Prior to genotyping DNA samples were quantified by Picogreen, normalised and 50ng/µl aliquots plated in 96 deep-well plates. The Illumina infinium II assay is a genome-wide genotyping assay carried out in a single tube using high-density BeadArray technology. Briefly, genomic DNA (~750ng) is isothermally amplified before fragmentation. After alcohol precipitation and DNA resuspension, samples are hybridised onto BeadChip arrays containing locus-specific 50-mer oligonucleotides. Allele detection through a two-step process provides high call rates and accuracy. An oligonucleotide primer hybridises to a complementary region, forming a duplex, with the primer's terminal 3' end directly adjacent to the nucleotide base to be identified (Figure 2.3). The primer is enzymatically extended a single base by a labelled nucleotide terminator complementary to the nucleotide being identified. The intensities of the beads' fluorescence are detected by the Illumina BeadArray Reader and analysed using Illumina's software for automated genotype calling (Figure 2.3; http://www.illumina.com/technology/beadarray-technology/infinium-hd-assay.ilmn).

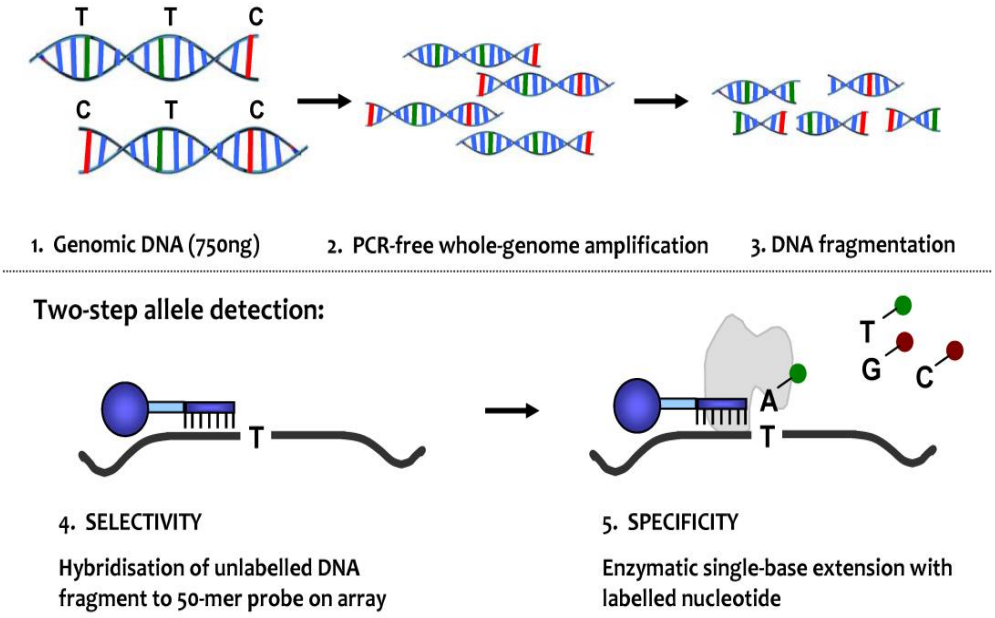


Figure 2.3 The Illumina infinium II genotyping assay.

2.3 Statistical and bioinformatics methods

2.3.1 General statistical methods

2.3.1.1 Software

Statistical analyses were carried out using the following statistical software programs: R v3.01 (http://www.r-project.org/) [120], PLINK v1.07 (http://pngu.mgh.harvard.edu/~purcell/plink/) [121] and custom perl/python scripts.

2.3.1.2 Assessing statistical significance

When assessing statistical significance, the *P*-value is defined as the probability of obtaining a value that is at least as extreme as that of the actual sample by chance. If the *P*-value is smaller than a pre-set threshold then the null hypothesis of no association is rejected and the result is considered significant. For a single test $P<0.05$ is deemed significant in order to control the family wise error rate (FWER; the probability of making even one type I error) at 0.05. To minimise type I error and keep the FWER at 0.05, a Bonferroni correction of the *P*-value can be applied. The corrected *P*-value is given by the equation $P = \alpha/n$, where α equates to the initially accepted level of significance (0.05) and n to the number of independent tests performed. For GWAS, previous simulations generating an infinitely dense set of polymorphisms identified a *P*-value cut off of $5x10^{-8}$ as appropriate in genome-wide studies [122][123][124]. Additional analyses were explicitly corrected according to the number of tests carried out unless stated otherwise. Continuous variables were analysed using Student's t tests. Study power is defined as the probability of rejecting the null hypothesis (H_0) of no association when the alternative hypothesis (H_1) is true [125].

2.3.2 General Bioinformatics techniques

2.3.2.1 Databases and publically available data resources

The following public databases were utilised in this thesis:

University of California Santa Cruz genome browser

The University of California, Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu/) is a virtual map of the human genome, annotated with known genes, transcripts, polymorphic variation, repeated sequences, conservation, structural variation and experimental data from external databases such as ENCODE (see below). These features are mapped against their physical positions in the genome. Various bioinformatics tools are contained within the website and were utilised as follows:

- *Genome Browser* tool was used to query specific regions of DNA and visualise genes, introns, regulatory elements and other features of the genomic location.
- *BLAT* tool was used to assess the binding accuracy of primers designed for PCR by finding possible spurious binding sites with >95% similarity to the sequence of interest.
- *LiftOver* tool was used to convert genome coordinates between different genome assemblies. Specifically, early GWAS SNPs may be mapped to NCBI Build 36 (hg18) whereas sequencing reads are mapped to the more recent Build 37 (hg19).
- *Table Browser* tool was used to download data associated with specific tracks in the genome browser. For example this tool was used to download genomic coordinates of genes, histone modifications and predicted transcription factor binding sites across specific regions and genome-wide.

National Centre for Biotechnology Information

The National centre for biotechnology information (NCBI) web server (http://www.ncbi.nlm.nih.gov/) hosts a multitude of databases and bioinformatics tools [126]. Specific tools used in this work are:

- *PubMed* for literature searches and citations..
- *RefSeq* to obtain reference sequences of chromosomes, genomic contigs, mRNAs and proteins. These data can also be queried in UCSC.
- *dbSNP* database of short genetic variations to query specific SNPs for position, allele and frequency information.
- *ClinVar* to query genetic variant pathogenicity

The Encyclopedia of DNA Elements

The encyclopedia of DNA elements (ENCODE) [105] was established in order to build a comprehensive list of functional elements in the human genome, including elements that act at the protein and RNA level, as well as DNA regulatory elements. The ENCODE project integrates genome-wide experimental data for over 100 different cell types. Data includes: chromatin structure (*e.g.* Hi-C), open chromatic prediction (*e.g.* DNase hypersensitivity), histone modifications and transcription factor binding prediction (ChIP-seq) and RNA transcription (RNAseq). All data is publicly available for download and can be viewed in the UCSC genome browser. From this data the

functionality of specific genomic regions can be inferred which is critical in fine-mapping studies and prioritisation of sequence variants.

1000 Genomes project

The 1000 Genomes Project (http://www.1000genomes.org/) was established to provide a comprehensive catalogue of human genetic variation with frequencies >1% through sequencing large numbers of individuals at 4x coverage [127]. Combining data from all individuals will then allow for accurate imputation of variants not directly covered in this low coverage sequencing. Data from the pilot phase, phase one and phase three of the project have been made publicly available. It is currently the largest publicly available resource for genome-wide variant frequency data across different populations worldwide.

Variant data from 1000 Genomes project were used for the following purposes:

- Haplotype data, as part of a reference panel for imputation.
- Variant frequency data, as part of a rare variant screening pipeline.

UK10K project

The UK10K project (http://www.uk10k.org/) aims to sequence 10,000 phenotyped people at 6x coverage in order to better understand the link between low-frequency and rare genetic changes and human disease [128]. The 10,000 individuals are split into three cohorts; the Twins UK and ALSPAC cohorts comprise 1,854 and 1,927 whole-genome sequenced individuals respectively and a further 6,000 individuals with extreme health problems (neurodevelopment, obesity and rare diseases) are to be exome sequenced. It is currently the largest publicly available resource for variant frequency data in the UK population.

Ensembl genome browser

The Ensembl genome browser (http://www.ensembl.org) is a genome annotation database supported by the European bioinformatics institute. Along with the ensembl biomart (http://www.ensembl.org/biomart/) it is of particular use for retrieval of gene information including genomic organisation of exons, introns and known regulatory domains, known transcripts, proteins, homologues and recorded variation within the gene sequence and also hosts the Variant Effect Predictor (VEP) for annotation of variant effects (See 0) [129].

Exome Aggregation Consortium (ExAC) Browser

The Exome Aggregation Consortium (ExAC) Browser (http://exac.broadinstitute.org/) contains variant frequencies from 60,706 unrelated individuals (of which 33,370 are non-Finnish European) sequenced as part of various disease-specific and population genetic studies. ExAC is currently the largest publicly available resource for coding variant sequence data worldwide.

The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) project has generated comprehensive, multi-dimensional maps of the key molecular changes in 33 types of cancer. The dataset encompasses DNA/RNA sequencing, methylation and SNP array platforms, together with clinical notes. Data has been generated on matched tumour/normal tissue for more than 15,000 patients and is publically available, with wide usage across the cancer research community.

The Genotype-Tissue Expression (GTEx) project

The Genotype-Tissue Expression (GTEx) (http://www.gtexportal.org/home/) project is a resource aiming to study human gene expression and regulation, and its relationship to genetic variation, in multiple tissue types. Expression Data, from Affymetrix Expression Array or Illumina TrueSeq RNA sequencing, is collected from tissue samples along with germline genotypes, from Illumina OMNI 5M SNP Array. GTEx contains integrated data from <7,000 samples, across >40 different tissue-types. By analysing global RNA expression within individual tissues and treating the expression levels of genes as quantitative traits, variations in gene expression that are highly correlated with genetic variation can be identified as expression quantitative trait loci (eQTLs).

2.3.2.2 Gene-set enrichment analysis

Gene set enrichment analysis (GSEA; http://www.broadinstitute.org/gsea/index.jsp) is a well-established, widely used and publicly available computational method that determines whether an *a priori* defined set of genes show statistically significant differences between two biological states (e.g. phenotypes) [130]. GSEA was used in Chapter 5.

2.3.2.3 High-throughtput-sequencing (HTS) pipeline

In Chapters 5 HTS methods were used to conduct whole exome sequencing, and the following data formats were utilised:

FASTQ format

The FASTQ format is a text-based format for storing nucleotide next-generation sequence reads and their corresponding per-base quality scores [131]. Additional information relating to whether reads are single-end or paired-end is also stored. Base quality scores (Q) are Phred-based and related to the probability (p) of a base call being false by the equation:

$Q = -10\ log_{10}\ p$

For example, a Q score of 10 corresponds to a 1 in 10 chance of an incorrect base call, whereas a Q score of 30 corresponds to a 1 in 1,000 chance.

Sequence alignment/map (SAM) format

The sequence alignment/map (SAM) format is the most widely used file format for storing read alignments against reference sequences [132]. Details of aligned and unaligned reads are stored along with associated mapping qualities. SAM files are typically stored in the binary form as BAM files.

Variant call format (VCF)

The variant call format (VCF) is a widely used specification for storing genetic sequence variations relative to a specified reference genome [133]. These files are typically generated by variant calling algorithms [134]. A variant in this format is defined as containing an allele (called the alternate allele) that is not the reference allele at that position. For a given genetic variant, the likely genotype is given along with a Phred-based genotype quality score, information about read depths for the reference and alternate alleles, genotype likelihoods as well as any additional meta-information.

These data types were generated using the following tools:

bcl2fastq (FASTQ extraction)

Illumina sequencing instruments generate per-cycle BCL basecall files as primary sequencing output, but many downstream analysis applications use per- read FASTQ files as input. bcl2fastq (https://support.illumina.com/tools.html) combines these per-cycle BCL files from a run and translates them into FASTQ files. At the same time as converting, bcl2fastq also separates multiplexed samples (demultiplexing). Multiplexed sequencing allows you to run multiple individual samples in one lane.

Stampy/BWA (sequence alignment)

Stampy (http://www.well.ox.ac.uk/project-stampy) [135] is a package designed for sensitive and fast single-end and paired-end mapping of short reads produced by Illumina-based sequencing. In its recommended hybrid mode, the Burrows-Wheeler aligner BWA (http://bio-bwa.sourceforge.net/) [136] is first used to map the majority of reads which are closely representative of the reference sequence. The remaining reads that could not initially be aligned are then mapped using the Stampy algorithm, which features a more detailed statistical model to aid sensitivity. In the exome sequence analysis pipeline, alignment to human build 37 reference genome was carried out in BWA (v. 0.5.10) and Stampy (v.1.0.23)

Picard tools (removing PCR duplicates)

Picard (http://broadinstitute.github.io/picard/) is a set of command line tools for working with next generation sequencing data in a reliable and efficient manner. In the exome sequence analysis pipeline, Picard (v.1.48) was used to filter duplicate reads and generate coverage metrics.

Genome Analysis Toolkit (local indel realignment and base score recalibration)

The Genome Analysis Toolkit (GATK; https://www.broadinstitute.org/gatk/) is a widely used software package developed for use in analysis of high-throughput sequencing data [137][138]. It was chosen (v. 3.1-1) for its ability to perform a wide range of analyses including local realignment, base score calibration and coverage estimation. Target realignment locally realigns target (*e.g.* exome capture) regions which may have been incorrectly mapped due to the presence of indels. The base quality score recalibration (BQSR) package attempts to recalibrate base quality scores of sequence reads in a BAM file. The aim is for these quality scores to more truly reflect the probability of mismatching the reference genome through correcting for variation in quality with machine cycle and sequence context. Coverage was estimated using the GATK DepthOfCoverage tool.

MuTect (Somatic variant calling)

MuTect (v. 1.1.4) was used for somatic variant detection (Chapter 5). MuTect is a widely used tool for accurate identification of point mutations found somatically in tumour tissue, and was chosen due to its low false positive rate. MuTect starts by preprocessing aligned reads in tumour and normal sequencing data, ignoring reads with low quality scores. Two Bayesian classifiers are then used to identify candidate somatic mutations, the first aims to detect whether the tumour is non-reference at a given site and then when this is found, the second classifier makes sure the normal does not carry the variant allele. Finally post-processing of candidate somatic mutations is completed, to eliminate artifacts of next-generation sequencing, short read alignment and hybrid capture.

IndelGenotyper

Somatic indels in Chapter 5 were called using IndelGenotyper. This GATK (https://software.broadinstitute.org/gatk/) tool uses a Bayesian genotype likelihood model to estimate simultaneously the most likely genotypes and allele frequency in a population of N samples, emitting a genotype for each sample.

2.3.2.4 In-silico prediction of variant effect

These programs and methods were used to predict variant functional effect:

Polyphen-2

Polymorphism phenotyping v2 (PolyPhen-2) (http://genetics.bwh.harvard.edu/pph2/) is an automatic web-based tool for prediction of possible functional impact of amino acid substitutions

on human proteins [139]. Sequence- and structure-based features of the substitution site are fed into a probabilistic classifier trained using a supervised machine-learning approach (Naive Bayes classifier). PolyPhen-2 calculates the posterior probability that a mutation is damaging and reports estimates of false positive rate (FPR) and true positive rate (TPR) in addition to a qualitative assessment that the mutation is benign, possibly damaging or probably damaging based on FPR thresholds.

SIFT

The sorting intolerant from tolerant (SIFT) algorithm (http://sift.jcvi.org/) predicts whether an amino acid substitution is likely to affect protein function [140]. SIFT assumes important positions within the protein sequence will be conserved through evolution and therefore mutations at these positions may affect protein function. By assessing this sequence conservation, SIFT predicts effects of all possible substitutions in a protein sequence. A score is output which ranges from 0 to 1. The amino acid substitution is predicted to be damaging if the score is ≤0.05 and tolerated if score is >0.05. While there exist a number of *in-silico* prediction algorithms, both the SIFT and Polyphen-2 methods are well-established and widely used, facilitating easier interpretation of their output among the scientific community.

CONDEL

The consensus deleteriousness score (CONDEL) method (http://bg.upf.edu/fannsdb/) integrates the output of up to five computational tools (SIFT, MutationAssessor, PolyPhen-2, LogRE and FATHMM) by computing a weighted average of the scores output from these tools [141]. Relative weights are calculated using the probability that a predicted deleterious mutation is not a false positive and the probability that a predicted neutral mutation is not a false negative. CONDEL predictions have been demonstrated to be more reliable than using the individual tools contributing to the algorithm alone [141].

CADD

The combined annotation dependent depletion (CADD) tool (http://cadd.gs.washington.edu/) scores the deleteriousness of SNVs and insertions/deletions in the human genome. C-scores are calculated by contrasting naturally occurring variants that have survived natural selection with simulated variants [142]. This information is integrated with 63 annotations of conservation, functional genomics and protein-level scores (SIFT, PolyPhen-2) derived from the Ensembl Variant

Effect predictor, ENCODE and the UCSC Genome Browser and used to train a support vector machine (SVM). Phred-like scaled C-scores ranging from 1 to 99 are calculated based on the rank of each variant relative to all possible 8.6 billion substitutions in the human genome. CADD scores were made use of as they allow genome-wide *in-silico* prediction of variant effect, as opposed to algorithms such as SIFT and PolyPhen-2 which are restricted to missense variants in coding regions.

Variant Effect Predictor

The Ensembl variant effect predictor (http://www.ensembl.org/info/docs/tools/vep/index.html) annotates the likely effect of genomic variants on genes, transcripts and protein sequence as well as regulatory, non-coding regions [129]. Along with the location (*e.g.* upstream of a transcript, in non-coding RNA, regulatory) and consequence of the variant (*e.g.* stop gained, missense), allele frequencies and predicted impacts from SIFT and PolyPhen-2 are returned, where available.

2.3.3 Methods for genome-wide association studies

Genome wide association study (GWAS) analyses in Chapter 4 were conducted using PLINK v1.07, a whole genome association analysis toolset which is designed to perform a range of basic, large-scale analyses [121]. PLINK provides a computationally efficient platform to store GWAS genotype data and to perform a number of quality control steps and association analyses in a typical GWAS analysis pipeline.

2.3.3.1 SNP quality control filtering

The GWAS SNP data (Chapter 3 and Chapter 4) was filtered as follows: all SNPs were excluded with minor allele frequency <1%, a call rate of <95% in cases or controls or with a minor allele frequency of 1–5% and a call rate of <99%. In addition SNPs deviating from Hardy-Weinberg equilibrium ($P < 10^{-12}$ in controls and $P < 10^{-5}$ in cases) were also removed. The Hardy-Weinberg principle states that the allele and genotype frequencies in a population will remain constant from generation to generation in the absence of evolutionary influences [143]. At a single locus with two alleles denoted A and a with frequencies $f(A)=p$ and $f(a)=q$, respectively, expected genotype frequencies are $f(AA)=p^2$, $f(aa)=q^2$ and $f(Aa) = 2pq$ for the AA homozygote, aa homozygote and Aa heterozygote respectively. As the sum of all genotype frequencies must equal 1: $p^2 + 2pq + q^2 = 1$. If a genetic locus satisfies this equation it is said to be in Hardy-Weinberg equilibrium (HWE), with deviation from HWE assessed using the χ^2-test [143].

2.3.3.2 Association analysis and meta-analysis

In Chapter 4 association between imputed SNPs and glioma was performed using logistic regression under an additive genetic model in SNPTESTv2.5 [144]. Overall significance was assessed using a fixed-effects meta-analysis in PLINK v1.07. In Chapter 3 and Chapter 4, for the new primary GWAS tests of association between imputed SNPs and glioma were performed under a probabilistic dosage model in in SNPTESTv2.5 [144], adjusting for principal components. Meta-analyses were performed using the fixed-effects inverse-variance method based on the β estimates and standard errors from each study using META v1.6 [145]. In Chapter 3 and 4 Cochran's Q-statistic to test for heterogeneity and the I^2 statistic to quantify the proportion of the total variation due to heterogeneity were calculated [146]. Throughout all GWAS studies a threshold of $P<5.0\text{x}10^{-8}$ was used to denote genome-wide significance. For each new locus discovered evidence of departure from a log-additive (multiplicative) model was examined for, to assess any genotype specific effect. Individual genotype data ORs were calculated for heterozygote (OR_{het}) and homozygote (OR_{hom}) genotypes, which were compared to the per allele ORs. A difference in these 1d.f. and 2d.f. logistic regression models was tested for, to assess for evidence of deviation ($P<0.05$) from a log-additive model. Subtype analyses were conducted to test for an association between SNP genotype and glioma risk for each individual histological subtype using logistic regression.

2.3.3.3 Assessment of inflation

Quantile-quantile (Q-Q) plots were used to assess the adequacy of case-control matching and the possibility of differential genotyping of cases and controls by comparing the distribution of observed test statistics from that of a null distribution. A Q-Q plot is a probability plot comparing two probability distributions by plotting their quantiles against each other. The highest observed value is plotted against the highest expected value. If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line y = x. The comparatively few variants with much higher observed than expected values are assumed to represent true associations. The inflation factor λ was calculated by dividing the median of the test statistics by the median expected values from a χ^2 distribution with 1 degree of freedom for the 90% least-significant SNPs [147]. However, it is recognised that the degree of inflation will increase with experiment size; thus standardisation is required to correct for experiment size. Therefore, an estimate of lambda corrected to an equivalent statistic as if the study were of 1,000 cases and 1,000 controls (λ_{1000}), was obtained using the formula:

$$\lambda_{1,000} = 1 + 500\ (1 / N_{cases} + 1 / N_{controls}) * (\lambda - 1)$$

where N_{cases} and $N_{controls}$ are the number of cases and controls, respectively. Q-Q plots were generated and inflation factors estimated using R.

2.3.3.4 Estimating linkage disequilibrium

SNPs adjacent in the genome are not randomly inherited; they are strongly correlated and likely to co-segregate together in a haplotype. This non-random association of alleles is termed linkage disequilibrium (LD). The most common measures of LD are *D'* and the correlation coefficient (r^2). *D'* is determined by dividing the disequilibrium co-efficient (*D*) by its maximum possible value (D_{max}), given the allele frequencies at the two loci. *D'* varies between 0 and 1 with a value of 1 corresponding to complete LD. Values less than one indicate disrupted LD and have no clear statistical interpretation particularly as *D'* is strongly inflated in small sample sizes and only measures recombinational history. Therefore, intermediate values should not be used to measure the extent of LD. The more stable r^2 is the preferred measure of the extent of LD as it summarises both the recombinational and the mutational history of the markers [148][149]. The r^2 statistic is equal to *D'* divided by the product of the allele frequencies at the two loci. Perfect LD is indicated by r^2=1 while high values of LD are generally defined as r^2>0.34 [150]. LD has been exploited by GWAS to maximise the coverage of the SNP genotyping platforms employed. Tagging SNPs are representative SNPs in a region of the genome in LD termed LD blocks or haplotypes.

Haploview v4.2 (http://www.broadinstitute.org/scientific-community/science/ programs/medical-and-population-genetics/haploview/Haploview) and SNP Annotation and Proxy search (SNAP) (http://www.broadinstitute.org/mpg/snap/) tools were used to calculate LD scores. LD blocks were defined using the HapMap recombination rates (cM/Mb) and defined using the Oxford recombination hotspots [151].

2.3.3.5 Imputation

Genome-wide imputation was performed on the MDA GWAS, SFAGS–GWAS, GICC GWAS Oncoarray and TCGA Affymetrix datasets. The 1000 genomes phase 3 data (2014 release) was used as a reference panel, with haplotypes pre-phased using SHAPEIT2 [152]. Imputation was performed using IMPUTE2 software [153] and association between imputed genotype and Glioma was tested

using SNPTEST [84], under a frequentist model of association. QC was performed on the imputed SNPs; excluding those with INFO score < 0.8 and MAF < 0.01.

2.3.4 Methods for functional analysis of genomic data

2.3.4.1 Measures of sequence conservation

These well-established methods were used to functionally annotate SNPs in Chapters 3 and 4. As with CADD (2.3.2.4), they allow genome-wide assessment of variant effect.

2.3.4.2 GERP

Genomic evolutionary rate profiling (GERP) [154] identifies sequence conservation by searching for substitution deficits in multiple sequence alignments. These substitutions would be expected to occur if the site were neutral and not under purifying selection. GERP scores vary from -12.3 to 6.17 with a score >2 taken as evidence of evolutionary constraint.

2.3.4.3 PhastCons

PhastCons is a statistical program which identifies evolutionarily conserved elements in multiple species alignments given a phylogenetic tree using a phylogenetic hidden Markov model (phylo-HMM) [155]. PhastCons produces base-by-base conservation scores and predictions of discrete conserved elements both of which can be visualised and downloaded from the UCSC genome browser. Predictions can be based on 100 vertebrate genomes, 46 primate genomes or just placental mammals. Conservation scores range from 0 to 1 with a score of >0.3 taken as evidence of sequence conservation.

2.3.5 Annotation of regulatory elements

Used in combination these tools can be used to derive increased insight into the potential function of a query risk SNP.

2.3.5.1 ChromHMM

ChromHMM (chromatin hidden markov model) is a software package for learning and characterising chromatin states. Multiple genomic datasets (e.g. ChIP-seq, histone marks) are integrated into a hidden Markov model that models the presence or absence of each chromatin mark to demarcate the genome into a defined number of states corresponding to different biological functions (e.g.

active promoter, strong enhancer or repetitive) [156]. This inference of regulatory elements aids in interpretation of SNP effect.

2.3.5.2 HaploReg

The Broad Institute's HaploReg database [157] is a tool for exploring non-coding variant genomic annotations. This web-based tool allows for visualisation of all linked variants, predicted chromatic state, sequence conservation and effects on regulatory motifs. This tool was used to functionally annotate SNPs in Chapter 3.

2.3.5.3 RegulomeDB

RegulomeDB [158] (http://www.regulomedb.org/) is a database that annotates non-coding SNPs with known and predicted regulatory elements from the gene expression omnibus (GEO) and ENCODE projects as well as published literature. The web-based interface can be queried for specific variants or genomic regions. Variants are scored from 1 to 6 corresponding to the overlapping regulatory elements identified. This tool was used to functionally annotate SNPs in Chapter 3.

2.3.5.4 Super-enhancer regions

Hnisz *et al* [159] propose the existence super-enhancers, which are large clusters of transcriptional enhancers that play key roles in human cell identity in health and in disease. They provide a catalogue of super-enhancers in 86 human cell and tissue types [159], allowing interrogation of DNA sequence of interest and potentially enabling increased insight into the functional effect of query risk SNPs. In Chapter 3 SNPs were annotated for overlap with super-enhancers in U87 GBM cells, astrocyte cells and brain tissue.

2.3.5.5 Roadmap epigenomics project

The Roadmap epigenomics project (http://www.roadmapepigenomics.org/) aims to investigate the hypothesis that the origins of health and susceptibility to disease are partly due to epigenetic regulation [160]. The goal of the project is to produce a public resource of human epigenomic data, for example DNA methylation, histone modifications, chromatin accessibility in stem cells and primary tissues, expanding the more limited range available from the ENCODE project. Chapter 3 made use of 15-state chromHMM data (se 2.3.5.1) from H1 derived neuronal progenitor cells available from the Epigenome roadmap project.

2.3.5.6 Expression quantitative trait locus (eQTL) analysis

For the Geuvadis the relationship between SNP and expression of genes located within 1 Mb was analysed using the Matrix eQTL package under a linear model. In all the datasets, SNPs in LD ($r^2 > 0.8$) with the potential pleiotropic associations were explored, and were included where FDR adjusted *P*-value < 0.05.

2.3.5.6.1 Summary-data-based Mendelian Randomisation

To examine the relationship between SNP genotype and gene expression, Summary-data-based Mendelian Randomization (SMR) analysis (http://cnsgenomics.com/software/smr/) was carried out as per Zhu *et al* [104]. Briefly, if b_{xy} is the effect size of *x* (gene expression) on *y* (slope of y regressed on the genetic value of *x*), b_{zx} is the effect of *z* on *x*, and b_{zy} be is the effect of *z* on *y*. Therefore b_{xy} (b_{zy}/b_{zx}) is the effect of *x* on *y*. To distinguish pleiotropy from linkage where the top associated *cis*-eQTL is in LD with two causal variants, one affecting gene expression the other affecting trait, heterogeneity was tested for in dependent instruments, using multiple SNPs in each *cis*-eQTL region. Under the hypothesis of pleiotropy b_{xy} values for SNPs in LD with the causal variant will be identical. Thus testing against the null hypothesis that there is a single causal variant is equivalent to testing heterogeneity in the b_{xy} values estimated for the SNPs in the *cis*-eQTL region. For each probe that passed significance threshold for the SMR test, heterogeneity in the b_{xy} values estimated for multiple SNPs in the *cis*-eQTL region using the HEIDI method [104].

2.3.5.7 Transcription factor binding motif analysis

To examine enrichment in specific TF binding across risk loci a variant set enrichment method was used. Briefly, for each risk locus, a region of strong LD (defined as $r^2 > 0.8$ and $D' > 0.8$) was determined, and these SNPs were termed the associated variant set (AVS). Transcription factor ChIP-seq uniform peak data were obtained from ENCODE for the GM12878 cell line, and included data for 82 TF. For each of these marks the overlap of the SNPs in the AVS and the binding sites was determined to produce a mapping tally. SNPs with the same LD structure as the risk associated SNP were randomly selected to calculate a null mapping tally. A null distribution was produced by repeating this process 10,000 times, and approximate *P*-values were calculated as the proportion of permutations where the null mapping tally was greater or equal to the AVS mapping tally. An enrichment score was calculated by normalising the tallies to the median of the null distribution.

Thus the enrichment score is the number of standard deviations of the AVS mapping tally from the mean of the null distribution tallies.

2.3.5.8 Hi-C analysis

To investigate the significant contacts between glioma risk SNPs and nearby genes in chapter 4, I made use of the HUGIn browser [161], which is based on analysis by Schmitt *et al*, 2016 [162]. I restricted analysis to Hi-C data generated on H1 Embryonic Stem Cell and Neuronal Progenitor cell lines, as originally described in Dixon *et al*, 2015 [163]. Plotted topologically associating domain (TAD) boundaries were obtained from the insulating score method [164] at 40-kb bin resolution. We searched for significant interactions between bins overlapping the glioma risk SNP and all other bins within 1Mb at each locus (*i.e.* "virtual 4C").

2.3.6 Methods for somatic genomic analysis

Tumour/normal somatic sequencing analysis was conducted as follows:

2.3.6.1 Somatic variant calling and driver gene analysis

The core HTS processing pipeline was followed, as described in 2.3.2.3, with final BAM files generated as normal. Single nucleotide variations (SNVs) were then called using MuTect (v. 1.1.4). Data was quality filtered using FoxoG software, based on methods as described in Costello et al 2013 [165], including removal of potential artefactual variants introduced through DNA oxidation. FoxoG ensured variants were supported by minimum of 1 alternative read in each strand direction, a mean Phred base quality score of > 26, mean mapping quality ≥50 and an alignability site score of 1.0. Small-scale insertion/deletions (Indels) were called using GATK IndelGenotyper. MutSigCV (v.1.4) was used to identify genes somatically mutated more often than would be expected by chance [166]. MutSigCV was run using the standard genomic covariates of (i) global gene expression data, (ii) DNA replication time and (iii) Hi-C statistic of open vs. closed chromatin states. Oncodrive-fm [167] was used as implemented within the IntOGen-mutations platform [168] for pathway analysis, using data mutation data from multiple tumour studies.

2.3.6.2 Somatic copy number alteration analysis

SNP array analysis

Genomic profiles were divided into homogeneous segments by applying the circular binary segmentation algorithm to both log R ratio and BAF values. We then used the Genome Alteration Print method to determine the ploidy of each sample, the level of contamination with normal cells

and the allele-specific copy number of each segment. Chromosome aberrations were defined using empirically determined thresholds as follows: gain, copy number ≥ploidy+1; loss, copy number ≤ploidy −1; high-level amplification, copy number >ploidy+2; homozygous deletion, copy number=0. Finally, we considered a segment to have undergone LOH when the copy number of the minor allele was equal to 0. Lists of homozygous deletions and focal amplifications, defined by at least five consecutive probes, were generated and verified manually to remove doubtful events. Significantly recurrent copy number changes were identified using the GISTIC2.0 algorithm [169].

2.3.6.3 Methods for RNA-seq analysis

Paired-end reads from RNA-seq were aligned to the following database files using BWA 0.5.5: (i) the human GRCh37-lite reference sequence, (ii) RefSeq, (iii) a sequence file representing all possible combinations of non-sequential pairs in RefSeq exons and (iv) the AceView database flat file downloaded from UCSC representing transcripts constructed from human ESTs. The mapping results from databases (ii)-(iv) were aligned to human reference genome coordinates. The final BAM file was constructed by selecting the best alignment.

2.3.6.4 Fusion detection

To identify fusion transcripts we analysed RNAseq data using Chimerascan software [170] (version 0.4.5). As advocated algorithmic output was analyzed for high-confidence fusion transcripts imposing filters (i) spanning reads > 2 (ii) total supported reads ≥10 [171]. In absence of corresponding paired normal tissue samples, we made use of data from the human body map project data to identify fusions seen in normal tissue.

2.3.7 Plotting tools

2.3.7.1 VisPIG

Visual plotting interface for genetics (visPIG; http://vispig.icr.ac.uk/) [234] is a web application for producing multi-region, multi-track, multi-scale plots of genetic data. Making use of code from SNAP [172] association plots can be generated. Additional tracks can be plotted to aid interpretation, for example chromHMM (see 2.3.5.1). At the time of writing, no other publicly available web-based tool exists to produce high-quality plots in this way. In this thesis therefore all association plots were generated using visPIG (*e.g.* in Chapter 3).

2.3.7.2 ggplot2

The ggplot2 package [173], created by Hadley Wickham, offers a powerful graphics language for creating elegant and complex plots. ggplot2 allows you to create graphs that represent both univariate and multivariate numerical and categorical data in a straightforward manner. Grouping can be represented by colour, symbol, size, and transparency. The creation of trellis plots (i.e., conditioning) is relatively simple.ggplot2 were used to generate plots in Chapter 4

2.3.8 Survival analysis

In Chapter 4, survival plots were generated using the *survfit* package in R which computes an estimate of a survival curve for censored data using the Kaplan–Meier method. Log-rank tests were used to compare curves between groups and power to demonstrate a relationship between different groups and overall survival was estimated using sample size formulae for comparative binomial trials. The Cox proportional-hazards regression model was used to investigate the association between survival and age, grade, molecular group and number of risk alleles. Individuals were excluded if they died within a month of surgery. Date of surgery was used as a proxy for the date of diagnosis.

CHAPTER 3

Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumours

In this Chapter, I had contributed to the bioinformatics and statistical analysis related to this project. I performed SNP quality control filtering, haplotypes phasing and the SNP imputation of the GICC GWAS, MDA GWAS and SFAGS–GWAS datasets described in 2.1. In addition, I had contributed to the meta-analysis.

The published version of this chapter with the figures is attached on appendix 2.

Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma.

Beatrice S Melin*#[1], Jill S Barnholtz-Sloan#[2], Margaret R Wrensch#[3,4] , Christoffer Johansen#[5], Dora Il'yasova#[6-8,] Ben Kinnersley#[9], Quinn T Ostrom[2], Karim Labreche[9,12,] Yanwen Chen[2], Georgina Armstrong[10], Yanhong Liu[10], Jeanette E Eckel-Passow[11], Paul A Decker[11], Marianne Labussière[12,] Ahmed Idbaih[12,13,] Khe Hoang-Xuan[12,13], Anna-Luisa Di Stefano[12,13], Karima Mokhtari[12,14,] Jean-Yves Delattre[12,13,] Peter Broderick[9], Pilar Galan[15] , Konstantinos Gousias[16], Johannes Schramm[16], Minouk J. Schoemaker[17], Sarah J Fleming[18], Stefan Herms[18,] Stefanie Heilmann[19], Markus M Nöthen[19], Heinz-Erich Wichmann[20-22], Stefan Schreiber[23], Anthony Swerdlow[17,24], Mark Lathrop[25], Matthias Simon[16,] Marc Sanson[12,13,] Ulrika Andersson[1], Preetha Rajaraman[26], Stephen Chanock[26], Martha Linet[26], Zhaoming Wang[26], Meredith Yeager[26], GliomaScan consortium[27], John K Wiencke[3,4], Helen Hansen[3], Lucie McCoy[3], Terri Rice[3], Matthew L Kosel[11], Hugues Sicotte[11], Christopher I Amos[28], Jonine L Bernstein[29], Faith Davis[30], Dan Lachance[31], Ching Lau[32], Ryan T Merrell[33] , Joellen Shildkraut[7,8], Francis Ali-Osman[7,34], Siegal Sadetzki[35,36,] Michael Scheurer[32], Sanjay Shete[37], Rose K Lai+[38], Elizabeth B Claus+[39,40], Sara H Olson+[29], Robert B Jenkins+[41], Richard S Houlston+*[9,42], Melissa L Bondy+*[10].

* Corresponding authors

shared first authors

+ shared last authors

** members presented in acknowledgement

Author affiliations:

1. Department of Radiation Sciences, Umeå University, Umeå, Sweden

2. Case Comprehensive Cancer Center, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA

3. Department of Neurological Surgery, School of Medicine, University of California, San Francisco, San Francisco, California, USA

4. Institute of Human Genetics, University of California, San Franciso, California, USA

5. Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen, Denmark, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark

6. Department of Epidemiology and Biostatistics, School of Public Health, Georgia State University, Atlanta, Georgia, USA

7. Duke Cancer Institute, Duke University Medical Center, Durham, North Carolina, USA

8. Cancer Control and Prevention Program, Department of Community and Family Medicine, Duke University Medical Center, Durham, North Carolina, USA

9. Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey, SM2 5NG, UK

10. Department of Medicine, Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, USA

11. Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, Minnesota, USA

12. Sorbonne Universités UPMC Univ Paris 06, INSERM CNRS, U1127, UMR 7225, ICM, F-75013 Paris, France.

13. AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de neurologie 2-Mazarin, Paris, France

14. AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de neurologie 2-Mazarin, Paris, France

15. Université Paris 13 Sorbonne Paris Cité, Inserm (U557), Inra (U1125), Cnam, France

16. Department of Neurosurgery, University of Bonn Medical Center, Sigmund-Freud-Str. 25, 53105 Bonn, Germany

17. Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK

18. Centre for Epidemiology and Biostatistics, Faculty of Medicine and Health, University of Leeds, Leeds LS2 9JT, UK

19. Institute of Human Genetics, University of Bonn, Germany

20. Helmholtz Center Munich, Institute of Epidemiology I, Germany

21. Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany

22. Institute of Medical Statistics and Epidemiology, Technical University Munich, Germany

23. 1st Medical Department, University Clinic Schleswig-Holstein, Campus Kiel, Germany

24. Division of Breast Cancer Research, The Institute of Cancer Research, London, UK

25. Génome Québec, Department of Human Genetics, McGill University, Montreal, Quebec, H3A 0G1, Canada

26. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, USA

27. Gliomascan consortium, the members of this consortium and their affiliations are listed in the Acknowledgements

28. Department of Biomedical Data Science Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA

29. Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, USA

30. School of Public Health, University of Alberta, Edmonton, Alberta, Canada

31. Department of Neurology, Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, Minnesota, USA

32. Department of Pediatrics Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, USA

33. Department of Neurology, NorthShore University HealthSystem, Evanston, Illinois, USA

34. Department of Surgery, Duke University Medical Center, Durham, North Carolina, USA

35. Cancer and Radiation Epidemiology Unit, Gertner Institute, Chaim Sheba Medical Center, Tel Hashomer, Israel

36. Department of Epidemiology and Preventive Medicine, School of Public Health, Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel

37. Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, Texas, USA

38. Departments of Neurology and Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA

39. School of Public Health, Yale University, New Haven, Connecticut

40. Department of Neurosurgery, Brigham and Women's Hospital, Boston, Massachusetts, USA

41. Department of Laboratory Medicine and Pathology, Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, Minnesota, USA

42. Division of Molecular Pathology, The Institute of Cancer Research, Sutton, Surrey, SM2 5NG, UK

Genome-wide association studies (GWAS) have transformed our understanding of glioma susceptibility, but individual studies have had limited power to identify risk loci. We performed a meta-analysis of existing GWAS and two new GWAS, totalling 12,496 cases and 18,190 controls. We identified five new loci for glioblastoma (GBM) at 1p31.3 (rs12752552; $P=2.04\times10^{-9}$, odds ratio (OR)=1.22), 11q14.1 (rs11233250; $P=9.95\times10^{-10}$, OR=1.24), 16p13.3 (rs2562152; $P=1.93\times10^{-8}$, OR=1.21), 16q12.1 (rs10852606; $P=1.29\times10^{-11}$, OR=1.18), 22q13.1 (rs2235573; $P=1.76\times10^{-10}$, OR=1.15) and eight for non-GBM at 1q32.1 (rs4252707; $P=3.34\times10^{-9}$, OR=1.19), 1q44 (rs12076373; $P=2.63\times10^{-10}$, OR=1.23), 2q33.3 (rs7572263; $P=2.18\times10^{-10}$, OR=1.20), 3p14.1 (rs11706832; $P=7.66\times10^{-9}$, OR=1.15), 10q24.33 (rs11598018; $P=3.39\times10^{-8}$, OR=1.14), 11q21 (rs7107785; $P=3.87\times10^{-10}$, OR=1.16), 14q12 (rs10131032; $P=5.07\times10^{-11}$, OR=1.33) and 16p13.3 (rs3751667; $P=2.61\times10^{-9}$, OR=1.18). These data substantiate genetic susceptibility to GBM and non-GBM being highly distinct, likely reflecting different etiology.

Glioma accounts for around 27% of all primary brain tumors and is responsible for approximately 13,000 cancer-related deaths in the US each year[1,2]. Gliomas can be broadly classified into glioblastoma (GBM) and lower-grade non-GBM[3]. Gliomas typically have a poor prognosis irrespective of medical care, with the most common form, GBM, having a five-year survival rate of only 5%[4].

So far, no environmental exposures have been robustly linked to risk of developing glioma except for moderate to high doses of ionizing radiation, which accounts for a small proportion of cases[5]. Evidence for inherited predisposition to glioma is provided by a number of rare inherited cancer syndromes, such as Turcot's and Li–Fraumeni syndromes, and neurofibromatosis. Even collectively, however these account for little of the two-fold familial risk of glioma[6]. Our understanding of the heritability of glioma has been transformed by recent genome-wide association studies (GWAS), which have identified single nucleotide polymorphisms (SNPs) at 13 loci influencing risk[7-14].

Previous individual studies have had limited statistical power for additional discovery of novel glioma risk loci[15]. Therefore, to gain a more comprehensive insight to glioma etiology, we performed a meta-analysis of previously published GWAS and two new GWAS, allowing us to identify 13 new risk loci for glioma.

We analysed GWAS SNP data passing quality control for 12,496 cases (6,191 classified as GBM, 5,819 as non-GBM) and 18,190 controls from eight studies of European ancestry; a new GWAS of 4,572 cases and 3,286 controls performed by the Glioma International Case Control Consortium (GICC) **(Supplementary Table 1)**, a new GWAS of 1,591 cases and 804 controls from University of California, San Francisco (UCSF)-Mayo and six previously reported GWAS[9,10,13] totalling 6,405 cases and 14,100 controls (**Supplementary Table 2**). To increase genomic resolution, we imputed >10 million SNPs. Quantile-Quantile (Q-Q) plots for SNPs with minor allele frequency (MAF) >1% post imputation did not show evidence of substantive over-dispersion (λ=1.02–1.10, λ_{90}=1.02–1.05; **Supplementary Fig. 1).** We derived joint odds ratios (ORs) and 95% confidence intervals (CIs) under a fixed-effects model for each SNP with MAF >1% and associated per allele principal component (PCA) corrected *P*-values for all glioma, GBM and non-GBM cases versus controls (**Fig. 1**).

In the combined meta-analysis, among previously published glioma risk SNPs, those for all glioma at 17p13.1 (*TP53*), GBM at 5p15.33 (*TERT*), 7p11.2 (*EGFR*), 9p21.3 (*CDKN2B-AS1*) and 20q13.33 (*RTEL1*) and for non-GBM at 8q24.21 (*CCDC26*), 11q23.2, 11q23.3 (*PHLDB1*) and 15q24.2 (*ETFA*) showed even greater evidence for association (**Supplementary Table 3, Supplementary Fig. 2**). SNPs at 10q25.2 and 12q12.1 for non-GBM tumors retained genome-wide significance (*i.e.* $P<5.0\times10^{-8}$). Associations at the previously reported 3q26.2 (near *TERC*)[11] and 12q23.33 (*POLR3B*)[10] loci for GBM did not retain statistical significance (respective *P*-values for the most associated SNPs = 2.68×10^{-5} and 1.60×10^{-5}; **Supplementary Table 3**).

In addition to previously reported loci, we identified genome-wide significant associations marking novel loci (**Table 1, Supplementary Fig. 3, Supplementary Data 1**) for GBM at 1p31.3 (rs12752552; $P=2.04\times10^{-9}$), 11q14.1 (rs11233250; $P=9.95\times10^{-10}$), 16p13.3 (rs2562152; $P=1.93\times10^{-8}$), 16q12.1 (rs10852606; $P=1.29\times10^{-11}$), 22q13.1 (rs2235573; $P=1.76\times10^{-10}$) and for non-GBM at 1q32.1 (rs4252707; $P=3.34\times10^{-9}$), 1q44 (rs12076373; $P=2.63\times10^{-10}$), 2q33.3 (rs7572263; $P=2.18\times10^{-10}$), 3p14.1 (rs11706832; $P=7.66\times10^{-9}$), 10q24.33 (rs11598018; $P=3.39\times10^{-8}$), 11q21 (rs7107785; $P=3.87\times10^{-10}$), 14q12 (rs10131032; $P=5.07\times10^{-11}$) and 16p13.3 (rs3751667; $P=2.61\times10^{-9}$). Conditional analysis confirmed the existence of two independent association signals at 7p11.2 (*EGFR*) as previously reported[7] but did not provide evidence for additional signals at any of the other established identified risk loci or the 13 newly identified loci. Case-only analyses confirmed the specificity of 11q14.1, 16p13.3 and 22q13.1 associations for GBM and 1q44, 2q33.3, 3p14.1, 11q21 and 14q12 for non-GBM tumors (**Supplementary Table 4**, **Fig. 2**). Collectively our findings provide strong evidence for subtype associations for glioma consistent with their distinctive molecular profiles presumably resulting from different etiological pathways.

Across the new and known risk loci, we found a significant enrichment of overlap with enhancers in H9 derived neuronal progenitor cells (P=8.2x10^{-5}; **Supplementary Data 2**). These observations support the assertion that the GWAS loci influence glioma risk through effects on neural cis-regulatory networks, and are strongly involved in transcriptional initiation and enhancement. To gain further insight into the biological basis for associations at the 13 new risk loci we performed an expression quantitative trait loci (eQTL) analysis using RNA-Seq data on 10 regions of normal human brain from up to 103 individuals from GTEx[16] and blood eQTL data on 5,311 individuals from Westra *et al*.[17] We used Summary level Mendelian Randomization (SMR)[18] analysis to test for an concordance between GWAS signal and cis-eQTL for genes within 1Mb of the sentinel and correlated SNPs (r^2>0.8) at each locus (**Supplementary Data 3**), deriving b_{XY} statistics which estimate the effect of gene expression on glioma risk. Additionally for each of the risk SNPs at the 13 new loci (as well as correlated variants) we examined published data[19,20] and made use of the online resources, HaploRegv4, RegulomeDB, and SeattleSeq for evidence of functional effect (**Supplementary Table 5**).

At 16q12.1 the GBM association signal was significantly associated with *HEATR3* expression in nine of ten regions of the brain (P_{SMR}=3.38x10^{-6}-6.55x10^{-10}, b_{XY}=0.14-0.24; **Supplementary Data 3**, **Supplementary Fig. 4**). The C-risk allele of rs10852606 being associated with reduced *HEATR3* expression is consistent with differential expression of *HEATR3* being the functional basis of the 16q12.1 association. The observation that variation at 16q12.1 is associated with risk of testicular[21] (rs8046148) and esophageal[22] (rs4785204) cancer (pairwise r^2 and D' with rs10852606, 0.67, 1.0 and 0.16, 1.0 respectively) suggests the locus has pleiotropic effects on tumor risk, compatible with generic effects as shown by the observation of a *HEATR3* eQTL signal in blood (P_{SMR}=5.84x10^{-11}, b_{XY}=0.30).

Similarly, significant associations between gene expression and glioma risk were observed at the GBM loci 1p31.3 (*JAK1*, brain cortex and cerebellar hemisphere), 16p13.3 (*POLR3K*, whole blood) and 22q13.1 (*CTA-228A9.3*, brain cerebellum; *PICK1*, brain hippocampus) (**Supplementary Data 3, Supplementary Fig. 4**). The non-GBM 1q32.1 association marked by rs4252707 (**Supplementary Fig. 3**) maps to intron eight of the gene encoding *MDM4* (mouse double minute 4 homolog) a p53-binding protein. SNP rs4252707 is in strong LD with rs12031912 and rs12028476 (r^2=0.92), which both map to the *MDM4* promoter. While no significant eQTL was shown in any brain tissue an association with *MDM4* was seen in blood (P_{SMR}=4.74x10^{-6}, b_{XY}=0.31; **Supplementary Data 3, Supplementary Fig. 4**). Over-expression of *MDM4* is a feature in *TP53*-wildtype and *MDM2*-amplification negative glioma, consistent with *MDM4* amplification being a mechanism by which the p53-dependent growth control is inactivated[23].

The 1q44 association marked by rs12076373 maps to the eighth intron of *AKT3* (v-akt murine thymoma viral oncogene homolog 3) one of the major downstream effectors of phosphatidylinositol 3-kinase which is highly expressed during active neurogenesis, with haploinsufficiency causing postnatal microcephaly and agenesis of the corpus callosum[24]. Importantly *AKT3* is hyper-expressed in glioma playing an important role in tumor viability by activating DNA repair[25]. While rs12076373 does not map to a regulatory element, correlated SNPs rs12124113 and rs59953491 (r^2=0.94 and 0.90 respectively), locate within an enhancer element in brain cells/tissues including H9 derived neuronal progenitor cultured cells, cortex derived primary cultured neurospheres and NH-A astrocytes.

The 3p14.1 association marked by rs11706832 localizes to intron 2 of *LRIG1* (leucine-rich repeats- and immunoglobulin-like domains-containing protein 1). Although we did not identify an eQTL *LRIG1* is highly expressed in the brain and is a pan-negative regulator of the *EGFR* signaling pathway which inhibits hypoxia-induced vasculogenic mimicry via *EGFR*/*PI3K*/*AKT* pathway suppression and epithelial-to-mesenchymal transition[26]. Reduced *LRIG1* expression is linked tumor aggressiveness, temozolomide-resistance and radio-resistance[27,28]. We have previously shown an association for glioma at *EGFR* (7p11.2)[7], which is well established to be pivotal in both initiation of primary GBM and progression of lower-grade glioma to grade IV. Although speculative our new findings now suggest a more extensive pathway involving variation at *LRIG1* and *AKT3*.

Of particular interest is rs7572263 mapping to 2q33.3 which localizes within intron three of *C2orf80* and is 50 kb telomeric to *IDH1* (isocitrate dehydrogenase 1). Mutation of *IDH1* is a driver for gliomagenesis[29,30] and is responsible for the CpG island methylator (G-CIMP) phenotype[31,32]. *IDH* mutation predominates in non-GBM glioma[33,34] therefore the association at 2q33.3 is plausible as a basis for susceptibility to non-GBM glioma. In the absence of convincing eQTL or other functional support, this does not preclude *C2orf80* or another gene mapping to the region of LD being the functional basis for the 2q33.3 association.

Maintenance of telomeres is central to cell immortalization and plays a central role in gliomagenesis[35]. We have previously shown the risk of GBM is strongly linked to genetic variation in the telomere-related genes *TERT* (5p15.33) and *RTEL1* (20q13.33), and possibly also *TERC* (3q26.2)[8,9,11]. The 10q24.33 association marked by rs11598018 lies intronic to *OBFC1* (oligonucleotide/oligosaccharide-binding fold-containing protein 1), which functions in a telomere-associated complex protecting telomeres independently of POT1[36]. The CST

complex encoded by *OBFC1*, *CTC1*, and *TEN1* competes with shelterin for telomeric DNA inhibiting telomerase-based telomere extension[37]. The significant association between risk of non-GBM and *OBFC1* variation is particularly intriguing in light of our recent exome sequencing report demonstrating that rare germline loss-of-function mutations in shelterin-complex genes are a cause of familial oligodendroglioma[38]. The glioma risk alleles at *TERT*, *TERC* and *OBFC1* are associated with increased leukocyte telomere length thereby supporting a relationship between genotype and biology (**Supplementary Table 6**)[35,39,40]. However the *RTEL1* locus is not consistent with such a postulate and recent data which has not shown a relationship between the *TERT* promoter mutation and telomere length in glioma[41] raises the possibility of a role for extra-telomeric effects.

Deregulation of pathways involved in telomere length and *EGFR* signalling are thus consistent with glioma risk being governed by pathways important in the longevity of glial cells and substantiate early observations that genetic susceptibility to GBM and non-GBM is highly distinct, presumably reflecting different aetiologies between GBM and non-GBM tumors (**Fig. 2**).

The other associations we identified mark genes with varying degrees of plausibility for having a role in glioma oncogenesis. The GBM association at 16p13.33 marked by rs2562152 localizes 3 kb telomeric to *MPG* which encodes a N-methylpurine DNA glycosylase whose expression is linked to temozolomide resistance in glioma[42]. Although attractive as a candidate, the only genes for which there was found to be a significant association between expression and glioma risk were *POLR3K* and *C16ORF33* in blood (**Supplementary Data 3, Supplementary Fig. 4)**. At 1p31.3 only *JAK1* provided convincing evidence for a significant eQTL with glioma risk SNPs in brain. The strongest association was shown in the cortex (P_{SMR}=1.61x10^{-6}, b_{XY}=0.22; **Supplementary Data 3, Supplementary Fig. 4**) with the T-risk allele of rs12752552 increasing *JAK1* expression. The cis-eQTL signal for *JAK1* in the cortex maps to 65.3Mb-65.35Mb and shows a consistent direction of effect with the glioma associated SNPs. J*AK1-STAT6* signaling is increasingly being recognized to be relevant to glioma progression[43]. Hence, while *JAK1* remains an attractive candidate mechanistic basis for the glioma association at 1p31.3 we cannot exclude the possibility of the cluster of SNPs between 65.3Mb and 65.35Mb containing the true causal variant. In the absence of functional data potential target genes for associations at 11q14.1 (GBM), 16p13.3 (non-GBM), 11q21 (non-GBM) and 14q12 (non-GBM) remain to be elucidated.

In conclusion, we have performed the largest glioma GWAS to date identifying 13 new glioma risk loci, thereby providing further evidence for a polygenic basis of genetic susceptibility to glioma. Histological

classification of glioma is in part being superseded by molecular profile[34,44]; hence, it is important to understand the biology behind these risk variants in the context of molecularly defined glioma subtypes. Currently identified risk SNPs for glioma account for at best around 27% and 37% of the familial risk of GBM and non-GBM tumors respectively (**Supplementary Table 7**). Therefore further GWAS-based studies in concert with functional analyses should lead to additional insights into the biology and etiological basis of the different glioma histologies. Importantly, such information can inform gene discovery initiatives and thus have a measurable impact on the successful development of new therapeutic agents.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGEMENTS

The GICC was supported by grants from the National Institutes of Health, Bethesda, Maryland

(R01CA139020, R01CA52689, P50097257, P30CA125123). Additional support was provided by the McNair Medical Institute and the Population Sciences Biorepository at Baylor College of Medicine.

In Sweden work was additionally supported by Acta Oncologica through the Royal Swedish Academy of Science (BM salary) and The Swedish Research council and Swedish Cancer foundation. We are grateful to the National clinical brain tumor group, all clinicians and research nurses throughout Sweden who identified all cases.

In the UK, funding was provided by Cancer Research UK (C1298/A8362 supported by the Bobby Moore Fund), the Wellcome Trust and the DJ Fielding Medical Research Trust. The National Brain Tumour Study is supported by the National Cancer Research Network and we acknowledge the contribution of all clinicians and heath care professionals to this initiative. The UK INTERPHONE study was supported by the European Union Fifth Framework Program 'Quality of life and Management of Living Resources' (QLK4-CT-1999-01563)

and the International Union against Cancer (UICC). The UICC received funds from the Mobile Manufacturers' Forum and GSM Association. Provision of funds via the UICC was governed by agreements that guaranteed INTERPHONE's scientific independence (http://www.iarc.fr/ENG/Units/RCAd.html) and the views expressed in the paper are not necessarily those of the funders. The UK centres were also supported by the Mobile Telecommunications and Health Research (MTHR) Programme and the Northern UK Centre was supported by the Health and Safety Executive, Department of Health and Safety Executive and the UK Network Operators.

In France, funding was provided by the Ligue Nationale contre le Cancer, the fondation ARC, the Institut National du Cancer (INCa; PL046), the French Ministry of Higher Education and Research and the program "Investissements d'avenir" ANR-10-IAIHU-06. This study was additionally supported by a grant from Génome Québec, le Ministère de l'Enseignement supérieur, de la Recherche, de la Science et de la Technologie (MESRST) Québec and McGill University.

In Germany, funding was provided to M.S. and J.S. by the Deutsche Forschungsgemeinschaft (Si552, Schr285), the Deutsche Krebshilfe (70-2385-Wi2, 70-3163-Wi3, 10-6262) and BONFOR. Funding for the WTCCC was provided by the Wellcome Trust (076113&085475). The KORA Ausburg studies are supported by grants from the German Federal Ministry of Education and Research (BMBF) and were mainly financed by the Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg. This work was financed by the German National Genome Research Network (NGFN) and supported within the Munich Center of Health Sciences (MC Health) as part of LMUinnovativ. Generation of the German control data was partially supported by a grant of the German Federal Ministry of Education and Research (BMBF) through the Integrated Network IntegraMent (Integrated Understanding of Causes and Mechanisms in Mental Disorders), under the auspices of the e:Med research and funding concept (01ZX1314A). Markus M. Nöthen received support from the Alfried Krupp von Bohlen und Halbach-Stiftung and is a member of the DFG-funded Excellence Cluster ImmunoSensation.

For the UK-GWAS, we acknowledge the funders, organizations and individuals who contributed to the blood sample and data collection as listed in Hepworth et al. (BMJ 2006, 332, 883). MD Anderson acknowledges the work on the MDA-GWAS of Phyllis Adatto, Fabian Morice, Hui Zhang, Victor Levin, Alfred W.K. Yung, Mark Gilbert, Raymond Sawaya, Vinay Puduvalli, Charles Conrad, Fredrick Lang and Jeffrey Weinberg from the Brain and Spine Center. For the French study, we are indebted to A. Rahimian (Onconeurotek), A.M. Lekieffre and M Brandel for help in collecting data, and Y Marie for database support. For the German study, we are

indebted to B. Harzheim (Bonn), S. Ott and Dr A. Müller-Erkwoh (Bonn) for help with the acquisition of clinical data and R. Mahlberg (Bonn) who provided technical support. The UK study made use of control genotyping data generated by the Wellcome Trust Case–Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. The MDA-GWAS made use of control genotypes from the CGEMS prostate and breast cancer studies. A full list of the investigators who contributed to the generation of the data is available from http://cgems.cancer.gov/. French controls were taken from the SU.VI.MAX study. The German GWA study made use of genotyping data from three population control sources: KORA-gen39, The Heinz-Nixdorf RECALL studyand POPGEN. The HNR cohort was established with the support of the Heinz Nixdorf Foundation. Franziska Degenhardt received support from the BONFOR Programme of the University of Bonn, Germany.

The UCSF Adult Glioma Study was supported by the National Institutes of Health (grant numbers R01CA52689, P50CA097257, R01CA126831, and R01CA139020), the Loglio Collective, the National Brain Tumor Foundation, the Stanley D. Lewis and Virginia S. Lewis Endowed Chair in Brain Tumor Research, the Robert Magnin Newman Endowed Chair in Neuro-oncology, and by donations from families and friends of John Berardi, Helen Glaser, Elvera Olsen, Raymond E. Cooper, and William Martinusen. This project also was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through UCSF-CTSI Grant Number UL1 RR024131. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The collection of cancer incidence data used in this study was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201000140C awarded to the Cancer Prevention Institute of California, contract HHSN261201000035C awarded to the University of Southern California, and contract HHSN261201000034C awarded to the Public Health Institute; and the Centers for Disease Control and Prevention's National Program of Cancer Registries, under agreement # U58DP003862-01 awarded to the California Department of Public Health. The ideas and opinions expressed herein are those of the author(s) and endorsement by the State of California Department of Public Health, the National Cancer Institute, and the Centers for Disease Control and Prevention or their Contractors and Subcontractors is not intended nor should be inferred. Other significant contributors for the UCSF Adult Glioma Study include: M Berger, P Bracci, S Chang, J Clarke, A Molinaro, A Perry, M Pezmecki, M Prados, I Smirnov, T Tihan, K Walsh, J Wiemels, S Zheng.

At Mayo the authors wish to acknowledge study participants, the clinicians and research staff at the participating medical centers, the Mayo Clinic Biobank and Biospecimens Accessioning and Processing Shared

Resource (in particular its manager Mine Cicek). Work at the Mayo Clinic beyond the GICC was also supported by the US National Institutes of Health (NIH; grants P50CA108961 and P30CA15083), the National Institute of Neurological Disorders and Stroke (grant RC1NS068222Z), the Bernie and Edith Waterman Foundation and the Ting Tsung and Wei Fong Chao Family Foundation.

The GliomaScan group comprised: Laura E. Beane Freeman, Stella Koutros, Demetrius Albanes, Kala Visvanathan, Victoria L. Stevens , Roger Henriksson, Dominique S. Michaud , Maria Feychting, Anders Ahlbom , Graham G. Giles Roger Milne , Roberta McKean-Cowdin, Loic Le Marchand, Meir Stampfer, Avima M. Ruder, Tania Carreon, Goran Hallmans, Anne Zeleniuch-Jacquotte, J. Michael Gaziano, Howard D. Sesso, Mark P. Purdue, Emily White , Ulrike Peters, Howard D. Sesso, Julie Buring.

UK10K data generation and access was organised by the UK10K consortium and funded by the Wellcome Trust.

We are grateful to all the patients and individuals for their participation and we would also like to thank the clinicians and other hospital staff, cancer registries and study staff in respective centers who contributed to the blood sample and data collection.

AUTHOR CONTRIBUTIONS

M.B., B.M., R.S.H. and J.B-S performed project management; R.S.H., M.B., B.M., J.B-S., R.B.J., Q.T.O., B.K. and M.W. drafted the manuscript; Q.T.O., K.L., B.K., J.E.E-P. and P.D performed statistical analyses; Y.C, K.L., Y.L. and B.K. performed bioinformatic analyses; B.M., J.B-S, M.R.W., J.K.W., C.J., D.Il'y, R.L., G.A., P.A.D., U.A., T.R., H.M.H., L.M., M.L.K., T.M., H.S., J.B., F.D., D.L, C.I.A, C.L., R.T.M., J.S., F.A.O., S.S., M.S, S. Shete, E.B.C., S.H.O., R.B.J., R.S.H., M.L.B. developed the GICC protocol and performed sample acquisition; P.R., S.C., M.L, Z.W. and M.Y. provided NCI data; in the UK, P.B, A.S., M.J.S., S.J.F. and R.S.H. developed patient recruitment, sample acquisition and performed sample collection of cases, P.B. oversaw DNA isolation and storage and performed case and control ascertainment and supervision of DNA extractions; in Germany, M.S., M.M.N., H.-E.W., S.S. and J.S. developed patient recruitment and blood sample collection, M.S. oversaw DNA isolation and storage and performed case and control ascertainment and supervision of DNA extractions, S.

Herms, S. Heilmann and K.G. performed experimental work; in France, M. Sanson and J.-Y.D. developed patient recruitment, M.L., A.-L.D.S, P.G, K.M., A.I K.H-X performed patient ascertainment; M. Lathrop performed laboratory management and oversaw genotyping of the French samples; all authors contributed to the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Bondy, M.L. *et al.* Brain tumor epidemiology: consensus from the Brain Tumor Epidemiology Consortium. *Cancer* **113**, 1953-68 (2008).

2. Ostrom, Q.T. *et al.* CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2008-2012. *Neuro Oncol* **17 Suppl 4**, iv1-iv62 (2015).

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

4. Ostrom, Q.T. *et al.* CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2007–2011. *Neuro-Oncology* **16**, iv1-iv63 (2014).

5. Ostrom, Q.T. *et al.* The epidemiology of glioma in adults: a "state of the science" review. *Neuro Oncol* **16**, 896-913 (2014).

6. Hemminki, K., Tretli, S., Sundquist, J., Johannesen, T.B. & Granstrom, C. Familial risks in nervous-system tumours: a histology-specific analysis from Sweden and Norway. *Lancet Oncol* **10**, 481-8 (2009).

7. Sanson, M. *et al.* Chromosome 7p11.2 (EGFR) variation influences glioma risk. *Hum Mol Genet* **20**, 2897-904 (2011).

8. Shete, S. *et al.* Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet* **41**, 899-904 (2009).

9. Wrensch, M. *et al.* Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. *Nat Genet* **41**, 905-8 (2009).

10. Kinnersley, B. *et al.* Genome-wide association study identifies multiple susceptibility loci for glioma. *Nat Commun* **6**, 8559 (2015).

11. Walsh, K.M. *et al.* Variants near TERT and TERC influencing telomere length are associated with high-grade glioma risk. *Nat Genet* **46**, 731-5 (2014).

12. Jenkins, R.B. *et al.* A low-frequency variant at 8q24.21 is strongly associated with risk of oligodendroglial tumors and astrocytomas with IDH1 or IDH2 mutation. *Nat Genet* **44**, 1122-5 (2012).

13. Rajaraman, P. *et al.* Genome-wide association study of glioma and meta-analysis. *Hum Genet* **131**, 1877-88 (2012).

14. Stacey, S.N. *et al.* A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat Genet* **43**, 1098-103 (2011).

15. Kinnersley, B. *et al.* Quantifying the heritability of glioma using genome-wide complex trait analysis. *Sci Rep* **5**, 17267 (2015).

16. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580-585 (2013).

17. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics* **45**, 1238-U195 (2013).

18. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-487 (2016).

19. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-61 (2014).

20. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).

21. Ruark, E. *et al.* Identification of nine new susceptibility loci for testicular cancer, including variants near DAZL and PRDM14. *Nat Genet* **45**, 686-9 (2013).

22. Wu, C. *et al.* Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. *Nat Genet* **44**, 1090-7 (2012).

23. Riemenschneider, M.J. *et al.* Amplification and overexpression of the MDM4 (MDMX) gene from 1q32 in a subset of malignant gliomas without TP53 mutation or MDM2 amplification. *Cancer Res* **59**, 6091-6 (1999).

24. Boland, E. *et al.* Mapping of deletion and translocation breakpoints in 1q44 implicates the serine/threonine kinase AKT3 in postnatal microcephaly and agenesis of the corpus callosum. *Am J Hum Genet* **81**, 292-303 (2007).

25. Turner, K.M. *et al.* Genomically amplified Akt3 activates DNA repair pathway and promotes glioma progression. *Proc Natl Acad Sci U S A* **112**, 3421-6 (2015).

26. Gur, G. *et al.* LRIG1 restricts growth factor signaling by enhancing receptor ubiquitylation and degradation. *EMBO J* **23**, 3270-81 (2004).

27. Yang, J.A. *et al.* LRIG1 enhances the radiosensitivity of radioresistant human glioblastoma U251 cells via attenuation of the EGFR/Akt signaling pathway. *Int J Clin Exp Pathol* **8**, 3580-90 (2015).

28. Wei, J. *et al.* miR-20a mediates temozolomide-resistance in glioblastoma cells via negatively regulating LRIG1 expression. *Biomed Pharmacother* **71**, 112-8 (2015).

29. Watanabe, T., Nobusawa, S., Kleihues, P. & Ohgaki, H. IDH1 mutations are early events in the development of astrocytomas and oligodendrogliomas. *Am J Pathol* **174**, 1149-53 (2009).

30. Yan, H. *et al.* IDH1 and IDH2 mutations in gliomas. *N Engl J Med* **360**, 765-73 (2009).

31. Noushmehr, H. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510-22 (2010).

32. Christensen, B.C. *et al.* DNA methylation, isocitrate dehydrogenase mutation, and survival in glioma. *J Natl Cancer Inst* **103**, 143-53 (2011).

33. Sanson, M. *et al.* Isocitrate dehydrogenase 1 codon 132 mutation is an important prognostic biomarker in gliomas. *J Clin Oncol* **27**, 4150-4 (2009).

34. Eckel-Passow, J.E. *et al.* Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors. *N Engl J Med* **372**, 2499-508 (2015).

35. Walsh, K.M. *et al.* Telomere maintenance and the etiology of adult glioma. *Neuro Oncol* **17**, 1445-52 (2015).

36. Miyake, Y. *et al.* RPA-like mammalian Ctc1-Stn1-Ten1 complex binds to single-stranded DNA and protects telomeres independently of the Pot1 pathway. *Mol Cell* **36**, 193-206 (2009).

37. Chen, L.Y., Redon, S. & Lingner, J. The human CST complex is a terminator of telomerase activity. *Nature* **488**, 540-4 (2012).

38. Bainbridge, M.N. *et al.* Germline mutations in shelterin complex genes are associated with familial glioma. *J Natl Cancer Inst* **107**, 384 (2015).

39. Zhang, C. *et al.* Genetic determinants of telomere length and risk of common cancers: a Mendelian randomization study. *Hum Mol Genet* **24**, 5356-66 (2015).

40. Walsh, K.M. *et al.* Longer genotypically-estimated leukocyte telomere length is associated with increased adult glioma risk. *Oncotarget* **6**, 42468-77 (2015).

41. Ceccarelli, M. *et al.* Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **164**, 550-63 (2016).

42. Xipell, E. *et al.* Endoplasmic reticulum stress-inducing drugs sensitize glioma cells to temozolomide through downregulation of MGMT, MPG, and Rad51. *Neuro Oncol* (2016).

43. Nicolas, C.S. *et al.* The role of JAK-STAT signaling within the CNS. *JAKSTAT* **2**, e22925 (2013).

44. Cancer Genome Atlas Research, N. *et al.* Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N Engl J Med* **372**, 2481-98 (2015).

FIGURE LEGENDS

Figure 1: Genome-wide discovery-phase meta-analysis *P*-values ($-\log_{10}P$, y axis) plotted against their chromosomal positions (x axis): a) All Glioma, b) GBM, c) Non-GBM. The red horizontal line corresponds to a significance threshold of $P = 5.0 \times 10^{-8}$. New and known loci are labelled in red and blue respectively.

Figure 2: Relative impact of SNP associations at known and newly identified risk loci for GBM and non-GBM tumors. Odds ratios (ORs) derived with respect to the risk allele. Asterisks denote SNPs showing a significant difference between GBM and non-GBM from the case-only analysis as detailed in Supplementary Table 4.

							All glioma		GBM glioma		Non-GBM glioma	
Locus	**Subtype**	**SNP**	**Position**	**Alleles**	**RAF**	**INFO**	***P***	**OR (95% CI)**	***P***	**OR (95% CI)**	***P***	**OR (95% CI)**
1p31.3	GBM	rs12752552	65229299	**T**/C	0.870	0.992	**4.07×10^{-9}**	1.18 (1.11-1.24)	**2.04×10^{-9}**	1.22 (1.15-1.31)	4.78×10^{-3}	1.11 (1.03-1.18)
1q32.1	non-GBM	rs4252707	204508147	G/**A**	0.220	0.992	2.97×10^{-7}	1.12 (1.07-1.17)	0.015	1.07 (1.01-1.13)	**3.34×10^{-9}**	1.19 (1.12-1.26)
1q44	non-GBM	rs12076373	243851947	**G**/C	0.837	0.996	4.97×10^{-4}	1.09 (1.04-1.15)	0.846	0.99 (0.94-1.06)	**2.63×10^{-10}**	1.23 (1.16-1.32)
2q33.3	non-GBM	rs7572263	209051586	**A**/G	0.756	0.997	2.58×10^{-6}	1.11 (1.06-1.15)	0.019	1.06 (1.01-1.12)	**2.18×10^{-10}**	1.20 (1.13-1.26)
3p14.1	non-GBM	rs11706832	66502981	A/**C**	0.456	0.997	1.06×10^{-5}	1.08 (1.05-1.12)	0.158	1.03 (0.99-1.08)	**7.66×10^{-9}**	1.15 (1.09-1.20)
10q24.33	non-GBM	rs11598018	105661315	**C**/A	0.462	0.960	3.07×10^{-7}	1.10 (1.06-1.14)	0.0103	1.06 (1.01-1.11)	**3.39×10^{-8}**	1.14 (1.09-1.20)
11q14.1	GBM	rs11233250	82397014	**C**/T	0.868	0.990	5.40×10^{-6}	1.14 (1.08-1.21)	**9.95×10^{-10}**	1.24 (1.16-1.33)	0.592	0.98 (0.91-1.05)
11q21	non-GBM	rs7107785	95747337	**T**/C	0.479	0.997	2.96×10^{-4}	1.07 (1.03-1.11)	0.844	1.00 (0.95-1.04)	**3.87×10^{-10}**	1.16 (1.11-1.21)
14q12	non-GBM	rs10131032	33250081	**G**/A	0.916	0.991	2.33×10^{-6}	1.17 (1.09-1.24)	0.247	1.05 (0.97-1.13)	**5.07×10^{-11}**	1.33 (1.22-1.44)
16p13.3	GBM	rs2562152	123896	A/**T**	0.850	0.937	1.18×10^{-3}	1.09 (1.04-1.15)	**1.93×10^{-8}**	1.21 (1.13-1.29)	0.948	1.00 (0.93-1.07)
16p13.3	non-GBM	rs3751667	1004554	C/**T**	0.208	0.985	**8.75×10^{-10}**	1.14 (1.09-1.19)	5.95×10^{-6}	1.13 (1.07-1.19)	**2.61×10^{-9}**	1.18 (1.12-1.25)
16q12.1	GBM	rs10852606	50128872	T/**C**	0.713	0.990	**3.66×10^{-11}**	1.14 (1.10-1.19)	**1.29×10^{-11}**	1.18 (1.13-1.24)	2.42×10^{-3}	1.08 (1.03-1.14)
22q13.1	GBM	rs2235573	38477930	**G**/A	0.507	0.995	8.64×10^{-7}	1.09 (1.06-1.13)	**1.76×10^{-10}**	1.15 (1.10-1.20)	0.325	1.02 (0.97-1.07)

Table 1: Association statistics for the top SNP at each of the newly-reported glioma risk loci. Associations at $P<5\times10^{-8}$ are highlighted in bold. Odds ratios (ORs) were derived with respect to the risk allele underlined and highlighted in bold. Minor allele frequency (MAF) is according to European samples from 1000 genomes project. The INFO column indicates the average imputation info score across all studies, with a score of 1 indicating the SNP is directly genotyped in all studies. CI, confidence interval.

ONLINE METHODS

Ethics

Collection of patient samples and associated clinico-pathological information was undertaken with written informed consent and relevant ethical review board approval at respective study centers in accordance with the tenets of the Declaration of Helsinki. Specifically, UK: South-East Multicentre Research Ethics Committee (MREC) and the Scottish MREC; France: APHP ethical committee-CPP (comité de Protection des Personnes); Germany: Ethics Commission of the Medical Faculty of the University of Bonn and USA: US: University of Texas MD Anderson Cancer Institutional Review Board, the Mayo Clinic Office for Human Research Protection, the UCSF Committee on Human Research, the University Hospitals of Cleveland Institutional Review Board and the Cleveland Clinic Institutional Review Board (board for the Case Comprehensive Cancer Center). The diagnosis of glioma [ICDO-3 codes 9380-9480 or equivalent], was established through histology in all cases in accordance with World Health Organization guidelines. Every effort was made to classify tumors as GBM or non-GBM.

GWAS datasets

GICC, UK, French, German, MDA, SFAGS and GliomaScan

Studies participating in GICC are described in Amirian *et al*.[45] and in **Supplementary Table 1**. Briefly, they comprise 5,189 glioma cases and 3,827 controls ascertained through centers in the US, Denmark, Sweden and the UK. Cases had newly diagnosed glioma and controls had no personal history of central nervous tumor at ascertainment. Detailed information regarding recruitment protocol is given in Amirian *et al.*[45]. Cases and controls were genotyped using the Illumina Oncoarray according to the manufacturer's recommendations (Illumina Inc.). Individuals with call rate <99% as well as all individuals evaluated to be of non-European ancestry (<80% estimated European ancestry using the FastPop[46] procedure developed by the GAMEON consortium with HapMap version 2 CEU, JPT/CHB and YRI populations as a reference; **Supplementary Fig. 5**) were excluded. For apparent first-degree relative pairs, we removed the control from a case-control pair; otherwise, we excluded the individual with the lower call rate. SNPs with a call rate <95% were excluded as were those with a MAF<0.01 or displaying significant deviation from Hardy-Weinberg equilibrium (HWE) (*i.e.* $P<10^{-5}$).

The UK, French, German, MDA, SFAGS and GliomaScan GWAS of non-overlapping case-control series of Northern European ancestry, have been the subject of previous studies; Briefly: (1) The UK-GWAS[7,8,10] was based on 636 cases (401 males; mean age 46 years) ascertained through the INTERPHONE study[47]. Individuals from the 1958 Birth Cohort (n=2,930) served as a source of controls; (2) The French-GWAS[7,10] comprised 1,495 patients with glioma ascertained through the Service de Neurologie Mazarin, Groupe Hospitalier Pitié-Salpêtrière Paris. The controls (n=1,213) were ascertained from the SU.VI.MAX (SUpplementation en VItamines et MineralesAntioXydants) study of 12,735 healthy subjects (women aged 35–60 years; men aged 45–60 years)[48]; (3) The German-GWAS[10] comprised 880 patients who underwent surgery for a glioma at the Department of Neurosurgery, University of Bonn Medical Center, between 1996 and 2008. Control subjects were taken from three population studies: KORA (Co-operative Health Research in the Region of Augsburg; n=488)[49]; POPGEN (Population Genetic Cohort; n=678)[50] and from the Heinz Nixdorf Recall study (n=380)[51]. Standard, quality control measures were applied to the UK, French and German GWAS and have previously been reported. (4) The MDA-GWAS[8] was based on 1,281 cases (786 males; mean age 47 years) ascertained through the MD Anderson Cancer Center, Texas, between 1990 and 2008. Individuals from the Cancer Genetic Markers of Susceptibility (CGEMS, n=2,245) studies served as controls[52,53]. Quality control measures were applied as per the Primary GWAS. (5) The SFAGS-GWAS. The UCSF adult glioma case-control study includes participants of the San Francisco Bay Area Adult Glioma Study (AGS). Details of subject recruitment for AGS have been reported previously[9,12,34,54,55]. Briefly, cases were adults (>18 years of age) with newly diagnosed histologically confirmed glioma. Population-based cases diagnosed between 1991 and 2009 (Series 1-4) and residing in the six San Francisco Bay Area counties were ascertained using the Cancer Prevention Institute of California's early case ascertainment system. Clinic-based cases diagnosed 2002-2012, (Series 3-5) were recruited from the UCSF Neuro-oncology Clinic, regardless of place of residence. From 1991 to 2010, population-based controls from the same residential area as the population-based cases were identified using random digit dialling and were frequency matched to population-based cases on age, gender, and ethnicity. Between 2010 and 2012, all controls were selected from the UCSF general medicine phlebotomy clinic. Clinic-based controls were matched to clinic-based glioma cases on age, gender, and ethnicity. Consenting participants provided blood, buccal, and/or saliva specimens and information during in-person or telephone interviews. A total of 677 cases and 3,940 controls (including 3,347 iControls) were used in the current analysis. (6) The GliomaScan-GWAS[13] – in addition to the published analysis we excluded samples from the ATBC (Finnish study) and controls from NSHDS which were excluded due to exhibiting outlying

population ancestry after manual inspection of PCA plots. In total 1,653 cases and 2,725 controls were used in the current study.

GWAS data from the seven studies were imputed to >10 million SNPs with IMPUTE2 v2.3[56] software using a merged reference panel consisting of data from 1000 Genomes Project (phase 1 integrated release 3, March 2012)[57] and UK10K (ALSPAC, EGAS00001000090/EGAD00001000195 and TwinsUK EGAS00001000108/EGAS00001000194 studies). Genotypes were aligned to the positive strand in both imputation and genotyping. Imputation was conducted separately for each study, and in each, the data were pruned to a common set of SNPs between cases and controls before imputation. We set thresholds for imputation quality to retain potential risk variants with MAF>0.01. Poorly imputed SNPs defined by an information measure <0.40 with IMPUTE2 were excluded, as were SNPs exhibiting a significant deviation from hardy-weinberg equilibrium (P<1x10^{-8}) in controls. Test of association between imputed SNPs and glioma was performed using SNPTESTv2.5[58] under an additive frequentist model. The adequacy of the case-control matching and possibility of differential genotyping of cases and controls were formally evaluated using Q-Q plots of test statistics (**Supplementary Fig. 1**). Where appropriate, principal components, generated using common SNPs, were included in the analysis to limit the effects of cryptic population stratification that otherwise might cause inflation of test statistics. Principal components, based on genotyped SNPs were generated for GICC, GliomaScan, MDA-GWAS and SFAGS studies using PLINK[59]. Eigenvectors for the German-GWAS were inferred using smartpca (part of EIGENSOFTv2.4)[60] by merging cases and controls with Phase II HapMap samples[10]. PCA plots for all studies are provided in Supplementary Figure 4.

UCSF-Mayo GWAS

The UCSF-Mayo study comprised Mayo cases (*n*=945) and UCSF cases (*n*=574) and Mayo Clinic Biobank control (*n*=806) data. The Mayo Clinic case-control study has been described previously[9,34,61]. Briefly, adult cases (>18 years of age) were identified at diagnosis (diagnosed at Mayo Clinic) or at pathologic confirmation (diagnosed elsewhere and treated at Mayo Clinic), and had a surgical resection or biopsy between 1973 and 2014. Consenting participants provided blood, buccal, and/or saliva specimens and information during in-person or telephone interviews. This analysis used 574 non-overlapping cases from the UCSF adult glioma study described above. Mayo Clinic and UCSF cases were genotyped using the Illumina Oncoarray. The Mayo Clinic Biobank controls comprised volunteers who donated biological specimens, provide risk factor data, access to clinical data obtained from the medical record

and provide consent to participate in any study approved by the Access Committee. Recruitment for the Mayo Clinic Biobank took place from April 2009 through December 2015. While participants could be unselected volunteers, the vast majority of participants were contacted as part of a pre-scheduled medical examination in the Department of Medicine Divisions of Community Internal Medicine, Family Medicine, and General Internal Medicine at Mayo Clinic sites in Rochester, MN; Jacksonville, FL; and the Mayo Clinic Health System sites in La Crosse and Onalaska, WI. All were aged 18 years and older at time of consent. Illumina Omni Express genotyping arrays were run on the 806 Mayo Clinic Biobank participants.

Quality control analyses were performed on each cohort separately (Mayo cases; UCSF cases; Mayo Clinic Biobank controls). SNPs with call rates <95% were removed, followed by removal of subjects with call rates <95%. Concordance of replicate samples was assessed and the sample with the higher call rate was retained. Subject's sex was verified using the sex check option in PLINK. Relationship checking was performed by estimating the proportion of alleles shared identical by descent (IBD) for all pairs of subjects in PLINK[59]. STRUCTURE[62] was used to assess population admixture with 1000 Genomes as reference. Subjects indicated to be non-Caucasian were excluded. Prior to imputation, SNPs were tested for HWE and SNPs with HWE $P<10^{-6}$ removed. Mayo Clinic, UCSF and Mayo Clinic Biobank SNP data were each phased and imputed using the Michigan Imputation Server with the Haplotype Reference Consortium (release 1; http://www.haplotype-reference-consortium.org) as reference. Genotypes were forward-strand aligned to the 1000 genome reference and for ambiguous SNPs the Browning strand checking utility was used

(http://faculty.washington.edu/sguy/beagle/strand_switching/strand_switching.html). PCA was used to correct for population stratification using SNPs common to cases and controls. The first three principal components were significantly ($P<0.05$) associated with case-control status. An additive logistic regression model was used to assess the association between each SNP and disease status, with genotype coded as 0, 1, or 2 copies of the minor allele, adjusted for age, sex, and the first three principal components.

Meta-analysis and additional statistical analyses

Meta-analyses were performed using the fixed-effects inverse-variance method based on the β estimates and standard errors from each study using META v1.6[63]. Cochran's Q-statistic was used to test for heterogeneity and the I^2 statistic was used to quantify the proportion of the total variation

due to heterogeneity[64], taking I^2 values > 75 to indicate significant heterogeneity. Using the meta-analysis summary statistics and LD correlations from a reference panel of 1000 Genomes Project combined with UK10K we used GCTA[65,66] to perform conditional association analysis. Association statistics were calculated for all SNPs conditioning on the top SNP in each locus showing genome-wide significance. This is carried out in a step-wise fashion. We performed a case-only analysis to test for differences in SNP risk allele frequency between GBM and non-GBM tumors.

ENCODE and chromatin state dynamics

Risk SNPs and their proxies (*i.e.*, $r^2 > 0.8$ in the 1000 Genomes EUR reference panel) were annotated for putative functional effect using HaploReg v4[67], RegulomeDB[68] and SeattleSeq Annotation[69]. These servers make use of data from ENCODE, genomic evolutionary rate profiling (GERP) conservation metrics, combined annotation dependent depletion (CADD) scores and PolyPhen scores. We searched for overlap of associated SNPs with enhancers defined by the FANTOM5 enhancer atlas[19], annotating by overlap with ubiquitous, permissive and robust enhancers as well as enhancer-promoter correlations and enhancers specifically expressed in astrocytes, neuronal stem cells and brain tissue. Similarly we searched for overlap with "super-enhancer" regions as defined by Hnisz *et al*., 2013[20] restricting analysis to data from U87 GBM cells, astrocyte cells and brain tissue. We additionally made use of 15-state chromHMM data from H1- and H9-derived neuronal progenitor cells available from the Epigenome roadmap project[70]. Enhancer enrichment analysis was carried out using HaploReg v4.0[67]. Briefly, from a query list of variants, the overlap with enhancers in each of 107 cell types as predicted from Roadmap Epigenomics Project[70] chromatin state segmentations is calculated. A binomial test for enrichment was performed against a background set of all 1) 1000 Genomes variants with MAF > 0.05 and 2) all unique GWAS loci in the European population. We applied a cutoff of $P<3.94\text{x}10^{-4}$ corresponding to a Bonferroni correction for 127 cell lines/tissues.

Expression quantitative trait loci analysis

To examine the relationship between SNP genotype and gene expression we carried out Summary-data-based Mendelian Randomization (SMR) analysis as per Zhu *et al*., 2016[18] (at http://cnsgenomics.com/software/smr/index.html). We used publicly available brain tissue data from the GTEx[16] (http://www.gtexportal.org) v6p release. Briefly, GWAS summary statistics files were generated from the meta-analysis. Reference files were generated from merging 1000 genomes phase 3 and UK10K (ALSPAC and TwinsUK) vcfs. Summary eQTL files for GTEx samples were generated from

downloaded v6p "all_snpgene_pairs" files. Besd files were generated from these summary eQTL files using the –make-besd command. Additionally, we analyzed downloaded whole blood eQTL data from Westra *et al*., 2016[17]. Results from the SMR test for each of the 13 new glioma loci are reported in **Supplementary Data 3**. As previously advocated[18] only probes with at least one eQTL *P*-value of $<5.0\times10^{-8}$ were considered for SMR analysis. We set a threshold for the SMR test of $P_{SMR} < 1.06\times10^{-4}$ corresponding to a Bonferroni correction for 473 tests (473 probes with a top eQTL $P<5.0\times10^{-8}$ across the 13 loci, 10 brain regions and Westra dataset). For all genes passing this threshold we generated plots of the eQTL and GWAS associations at the locus, as well as plots of GWAS and eQTL effect sizes (*i.e.* corresponding to input for the HEIDI heterogeneity test). HEIDI test *P*-values < 0.05 were taken to indicate significant heterogeneity. Respective SMR plots for significant eQTLs are shown in **Supplementary Fig. 4**.

Additional statistical and bioinformatics analysis

Estimates of individual variance in risk associated with glioma risk SNPs was carried out using the method described in Pharoah, *et al*., 2008[71] assuming the familial risk of high-grade and low-grade glioma to be 1.76 and 1.54 respectively from analysis of the Swedish series in Scheurer *et al.*, 2010[72]. Briefly, for a single allele (*i*) of frequency *p*, relative risk *R* and ln risk *r*, the variance (V_i) of the risk distribution due to that allele is given by:

$$V_i = (1-p)^2E^2 + 2p(1-p)(r-E)^2 + p^2(2r-E)^2$$

Where *E* is the expected value of *r* given by:

$$E = 2p(1-p)r + 2p^2r$$

For multiple risk alleles the distribution of risk in the population tends towards the normal with variance:

$$V = \sum V_i$$

The total genetic variance (*V*) for all susceptibility alleles has been estimated to be √1.77. Thus the fraction of the genetic risk explained by a single allele is given by:

$$V_i/V$$

LD metrics were calculated in vcftools v0.1.12b[73] using UK10K data and plotted using visPIG[74]. LD blocks were defined on the basis of HapMap recombination rate (cM/Mb) as defined using the Oxford recombination hotspots and on the basis of distribution of confidence intervals defined by Gabriel *et al*.[75]

Data availability

Genotype data from the GICC GWAS are available from dbGaP (xx). Additionally, genotypes from the GliomaScan GWAS can be accessed through dbGaP accession phs000652.v1.p1. Data from other studies are available upon request.

ONLINE METHODS REFERENCES

45. Amirian, E.S. *et al.* The Glioma International Case-Control Study: A Report From the Genetic Epidemiology of Glioma International Consortium. *Am J Epidemiol* **183**, 85-91 (2016).

46. Li, Y. *et al.* FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data. *BMC Bioinformatics* **17**, 122 (2016).

47. Cardis, E. *et al.* The INTERPHONE study: design, epidemiological methods, and description of the study population. *Eur J Epidemiol* **22**, 647-64 (2007).

48. Hercberg, S. *et al.* The SU.VI.MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. *Arch Intern Med* **164**, 2335-42 (2004).

49. Wichmann, H.E., Gieger, C., Illig, T. & Group, M.K.S. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* **67 Suppl 1**, S26-30 (2005).

50. Krawczak, M. *et al.* PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet* **9**, 55-61 (2006).

51. Schmermund, A. *et al.* Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the Heinz Nixdorf RECALL Study. Risk Factors, Evaluation of Coronary Calcium and Lifestyle. *Am Heart J* **144**, 212-8 (2002).

52. Hunter, D.J. *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39**, 870-4 (2007).

53. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39**, 645-9 (2007).

54. Wiemels, J.L. *et al.* History of allergies among adults with glioma and controls. *Int J Cancer* **98**, 609-15 (2002).

55. Felini, M.J. *et al.* Reproductive factors and hormone use and risk of adult gliomas. *Cancer Causes Control* **20**, 87-96 (2009).

56. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).

57. Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).

58. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).

59. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

60. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).

61. Jenkins, R.B. *et al.* Distinct germ line polymorphisms underlie glioma morphologic heterogeneity. *Cancer Genet* **204**, 13-8 (2011).

62. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-59 (2000).

63. Liu, J.Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* **42**, 436-40 (2010).

64. Higgins, J.P., Thompson, S.G., Deeks, J.J. & Altman, D.G. Measuring inconsistency in meta-analyses. *BMJ* **327**, 557-60 (2003).

65. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).

66. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, S1-3 (2012).

67. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-4 (2012).

68. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-7 (2012).

69. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-6 (2009).

70. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).

71. Pharoah, P.D., Antoniou, A.C., Easton, D.F. & Ponder, B.A. Polygenes, risk prediction, and targeted prevention of breast cancer. *N Engl J Med* **358**, 2796-803 (2008).

72. Scheurer, M.E. *et al.* Familial aggregation of glioma: a pooled analysis. *Am J Epidemiol* **172**, 1099-107 (2010).

73. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8 (2011).

74. Scales, M., Jager, R., Migliorini, G., Houlston, R.S. & Henrion, M.Y. visPIG--a web tool for producing multi-region, multi-track, multi-scale plots of genetic data. *PLoS One* **9**, e107497 (2014).

75. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).

CHAPTER 4

Diffuse gliomas classified by 1p/19q co-deletion, TERT promoter and IDH mutation status are associated with specific genetic risk loci

6.1. Overview and rational

Despite glioma being an especially devastating malignancy little is known about its aetiology and aside from exposure to ionising radiation that accounts for very few cases no environmental or lifestyle factor has been unambiguously linked to risk [174]. Recent genome-wide association studies (GWAS) have however enlightened our understanding of glioma genetics identifying single-nucleotide polymorphisms (SNPs) at multiple independent loci influencing risk [85][87][89][92][88][91][175]. While understanding the functional basis of these risk loci offers the prospect of gaining insight into the development of glioma, few have been deciphered. Notable exceptions are the 17p13.1 locus, where the risk SNP rs78378222 disrupts *TP53* polyadenylation [88] and the 5p15.33 locus, where the risk SNP rs10069690 creates a splice-donor site leading to an alternate *TERT* splice isoform lacking telomerase activity [102].

Since the aetiological basis of glioma subtypes is likely to reflect different developmental pathways it is not perhaps surprising that subtype-specific associations have been shown for GBM (5p15.33, 7p11.2, 9p21.3, 11q14.1, 16p13.33, 16q12.1, 20q13.33 and 22q13.1) and for non-GBM glioma (1q44, 2q33.3, 3p14.1, 8q24.21, 10q25.2, 11q21, 11q23.2, 11q23.3, 12q21.2, 14q12 and 15q24.2) [175]. Recent large-scale sequencing projects have identified IDH mutation, *TERT* promoter mutation and 1p/19q co-deletion as cancer drivers in glioma. These findings have improved the subtyping of glioma [20][26][32][176] and this information has been incorporated into the revised 2016 WHO classification of glial tumours [16]. Since these mutations are early events in glioma development, any relationship between risk SNP and molecular profile should provide insight into glial oncogenesis. Evidence for the existence of such subtype specificity is already provided by the association of the 8q24.21 (rs55705857) risk variant with 1p/19q co-deletion, IDH mutated glioma [86]. Additionally, it has been proposed that associations may exist between risk SNPs at 5p15.33, 9p21.3 and 20q13.33 and

IDH wild-type glioma [177], as well as 17p13.1 and *TERT* promoter, IDH mutated glioma without 1p/19q co-deletion [32].

To gain a more comprehensive understanding of the relationship between the 25 glioma risk loci and tumour subtype I analysed three patient series totalling 2,648 cases. Since generically the functional basis of GWAS cancer risk loci appear primarily to be through regulatory effects [52], we analysed Hi-C and gene expression data to gain insight into the likely target gene/s of glioma risk SNPs.

The results of this Chapter have been published (APPENDIX 2). Therefore, due to the format, some data are available in the online version of the paper.

6.2. Methods

6.2.1 Patients, samples and datasets

We analysed data from three non-overlapping case series: TCGA, French GWAS, French sequencing. Details of these datasets are provided below and are summarised in Table 4.1.

TCGA

Raw genotyping files (.CEL) for the Affymetrix Genome-wide version 6 array were downloaded for germline (*i.e.* normal blood) glioma samples from The Cancer Genome Atlas (TCGA, dbGaP study accession: phs000178.v1.p1). Controls were from publicly accessible genotype data generated by the Wellcome Trust Case-Control Consortium 2 (WTCCC2) analysis of 2,699 individuals from the 1958 British birth cohort (1958-BC) [109]. Genotypes were generated using the Affymetrix Power Tools Release 1.20.5 using the Birdseed (v2) calling algorithm (https://www.affymetrix.com/support/developer/powertools/changelog/index.html) and PennCNV [178]. After quality control (Figure 4.1 and 4.2, Table 4.2) there were 521 TCGA glioma cases and 2,648 controls (Table 4.1). Glioma tumour molecular data (IDH mutation, 1p/19q co-deletion, *TERT* promoter mutation) were obtained from Ceccarelli *et al,* 2016 [179]. Further data (*EGFR* amplification/activating mutations, *CDKN2A* deletion) were obtained from the cBioportal for cancer genomics [180]. After adjustment for principal components there was minimal evidence of over-dispersion inflation (λ=1.01; Figure.4.2).

French GWAS

The French-GWAS [87][92] comprised 1,423 patients with newly diagnosed grade II to IV diffuse glioma attending the Service de Neurologie Mazarin, Groupe Hospitalier Pitié-Salpêtrière Paris. The controls (n=1,190) were ascertained from the SU.VI.MAX (SUpplementation en VItamines et MineralauxAntioXydants) study of 12,735 healthy subjects (women aged 35–60 years; men aged 45–60 years) [181]. Tumours from patients were snap-frozen in liquid nitrogen and DNA was extracted using the QIAmp DNA minikit, according to the manufacturer's instructions (Qiagen, Venlo, LN, USA). DNA was analysed for large-scale copy number variation by comparative genomic hybridisation (CGH) array as previously described [182][183]. For tumours not analysed by CGH array, 1p/19q co-deletion status was assigned using PCR microsatellites, and *EGFR*-amplification and *CDKN2A-p16-INK4a* homozygous deletion by quantitative PCR. *IDH1*, *IDH2* and *TERT* promoter mutation status was assigned by sequencing [26][184].

French sequencing

Eight hundred and fifteen patients newly diagnosed grade II to IV diffuse glioma were ascertained through the Service de Neurologie Mazarin, Groupe Hospitalier Pitié-Salpêtrière Paris. Genotypes for the 25 risk SNPs were obtained by universal-tailed amplicon sequencing in conjunction with Miseq technology (Illumina inc). Genotypes were called using GATK (Genome Analysis ToolKit, version 3.6-0-g89b7209) software. Duplicated samples and individuals with low call rate (<90%) were excluded (n=111). Molecular profiling of tumour samples was carried out as per the French GWAS.

			Case groupings																
			IDH status		*EGFR*		*CDKN2A*		Molecular subgroup						WHO 2016 classification					
Dataset	**Controls**	**Cases** (GBM/non-GBM)	**mut**	**wt**	**amp**	**wt**	**del**	**wt**	**IDH-only**	***TERT*-IDH**	***TERT*-only**	**Triple –ve**	**Triple +ve**	**Total**	**Astro IDH-mut**	**Astro IDH-wt**	**Oligo 1p19q**	**GBM IDH-mut**	**GBM IDH-wt**	**Total**
TCGA	2,648	521 (183/338)	293	228	246	270	254	262	100	4	45	10	65	224	166	51	116	10	171	514
French GWAS	1,190	1,423 (430/993)	366	498	118	628	173	573	169	46	309	141	85	750	188	214	95	27	233	757
French seq	5,527	704 (181/523)	427	277	101	592	144	549	181	28	185	92	199	685	178	114	218	31	148	689
Total	**9,365**	**2,648 (795/1,854)**	**1,086**	**1,003**	**465**	**1,490**	**571**	**1,384**	**450**	**78**	**539**	**243**	**349**	**1,659**	**532**	**379**	**429**	**68**	**552**	**1,960**

Table 6.1 Overview of TCGA, French GWAS and French Seq series and mutation status of tumours . Amp, amplified; astro, astrocytoma; del, deleted; mut, mutated; oligo, oligodendroglioma; wt, wildtype.

Unrelated French controls were obtained from the 3C Study (Group, 2003) [185] a population-based, prospective study of the relationship between vascular factors and dementia being carried out in Bordeaux , Montpellier, and Dijon. Genotyping of controls was performed using Illumina Human 610-Quad BeadChips. To recover untyped genotypes imputation using IMPUTE2 software was performed using 1000 Genomes multi-ethnic data (1000 G phase 1 integrated variant set release v3) as reference. SNPs genotypes were retained call rates were >98%, Hardy-Weinberg equilibrium (HWE) *P* value > $1x10^{-6}$, minor allele frequency (MAF) > 1%. After quality control, 704 cases and 5,527 controls were available for analysis (Table 4.1).

	TCGA		**French Sequencing**	
	Cases	**Controls**	**Cases**	**Controls**
Pre-quality control	754	2662	815	5,527
Sex discrepancy	24	-	-	-
Call rate <0.9	-	-	111	-
Heterozygosity rate	55	10	-	-
Related Individuals	-	5	-	-
Non-European Ancestry	179	-	-	-
Post-quality control†	521	2,648	704	5,527

Table 6.2 Details of the quality control filters applied to TCGA and FRENCH sequencing studies. Samples were excluded due to call rate (< 90% or failed genotyping), ethnicity (principle components analysis or other samples reported to be not of white, European descent), relatedness (any individuals found to be duplicated or related within or between data sets through identity by state) or sex discrepancy. TCGA, The Cancer Genome Atlas. † filters for quality control were performed simultaneously so numbers for each criteria may not sum to total removed.

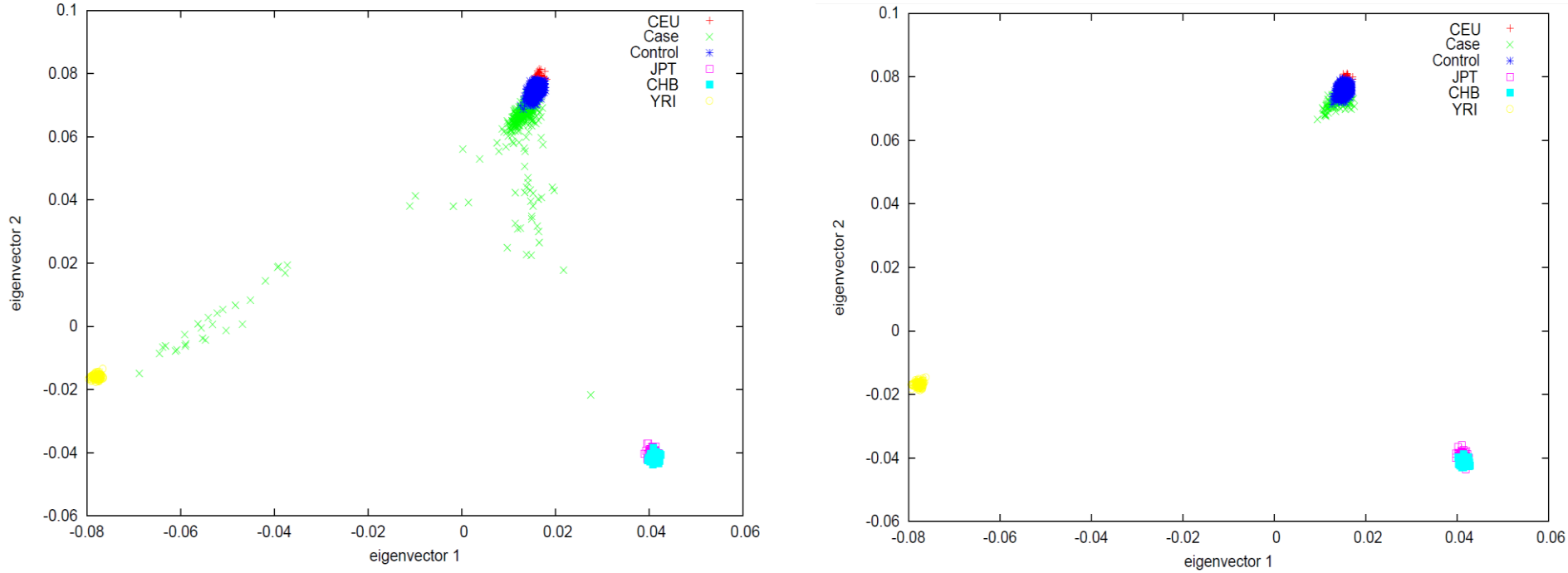


Figure 6.1 Identification of individuals of non-European ancestry in TCGA cases and WTCC controls. (a) before excluding non-European ancestry in cases and controls, and (b) after. The first two principal components of the analysis are plotted. HapMap CEU individuals are plotted in red, JPT individuals are plotted in pink, CHB are plotted in cyan, YRI are plotted in yellow. Cases are plotted in green and controls are plotted in blue.

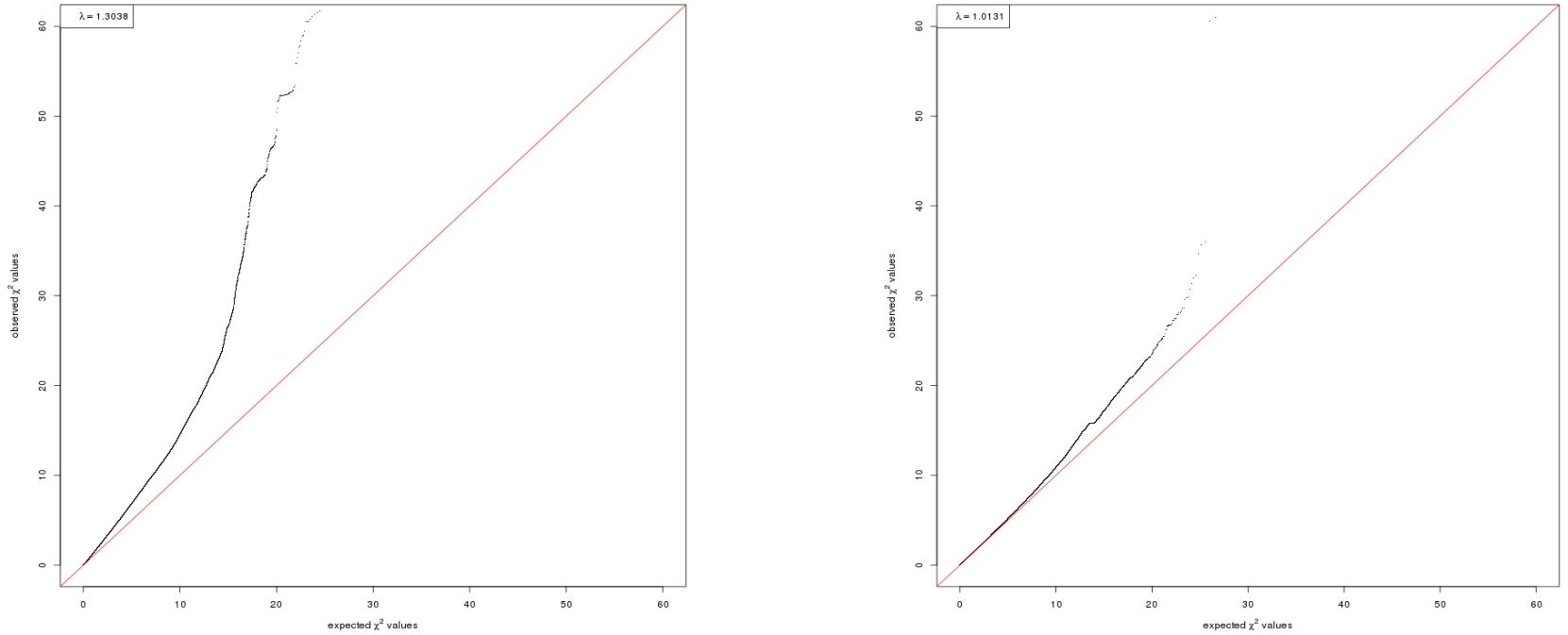


Figure 6.2 Quantile-Quantile (Q-Q) plots of observed and expected χ^2 values of association between SNP genotype and risk of glioma after imputation using TCGA cases and WTCC controls.. Before adjustment for population stratification in the left and after adjustment in the right. The red line represents the null hypothesis of no true association.

6.2.2 Statistical analysis

Test of association between SNP and glioma molecular subgroup was performed using SNPTESTv2.5 [144] under an additive frequentist model. Where appropriate, principal components, generated using common SNPs, were included in the analysis to limit the effects of cryptic population stratification that otherwise might cause inflation of test statistics. Eigenvectors for the TCGA study were inferred using smartpca (part of EIGENSOFTv2.4) [186] by merging cases and controls with Phase II HapMap samples [92].

To ensure reliability when restricting cases to per-group low sample counts, imputed genotypes were thresholded at a probability > 0.9 (*e.g.* –method threshold in SNPtest) for the TCGA and French-GWAS studies. For the French-sequence study we used –method expected, as we were comparing genotypes from directly sequenced cases against imputed controls. We compared control frequencies to those from European 1000 genomes project to ensure the validity of this approach.

Meta-analyses were performed using the fixed-effects inverse-variance method based on the β estimates and standard errors from each study using META v1.6 [145] Cochran's Q-statistic was used to test for heterogeneity [187].

Risk allele number and age at diagnosis

For imputed SNPs a genotype probability threshold > 0.9 was used. The age and survival distribution of cases carrying additive combinations of risk alleles were assessed for the 25 SNPs across the molecular subgroups. Trend lines were estimated using linear regression in R and plotted using the *ggplot2* package [173]. Association between risk allele number and age was assessed using Pearson correlation.

Survival analysis

Survival plots were generated using the *survfit* package in R which computes an estimate of a survival curve for censored data using the Kaplan-Meier method. Log-rank tests were used to compare curves between groups and power to demonstrate a relationship between different groups and overall survival was estimated using sample size formulae for comparative binomial trials. The Cox proportional-hazards regression model was used to investigate the association between survival and

age, grade, molecular group and number of risk alleles. Individuals were excluded if they died within a month of surgery. Date of surgery was used as a proxy for date of diagnosis.

Expression quantitative trait locus analysis

We searched for expression quantitative trait loci (eQTLs) in 10 brain regions using the V6p GTEx [188] portal (https://gtexportal.org/home/) as well as in whole blood using the blood eQTL browser [189] (https://molgenis58.target.rug.nl/bloodeqtlbrowser/).

Hi-C analysis

We examined for significant contacts between glioma risk SNPs and nearby genes using the HUGIn browser [161], which is based on analysis by Schmitt *et al*, 2016 [162]. We restricted analysis to Hi-C data generated on H1 Embryonic Stem Cell and Neuronal Progenitor cell lines, as originally described in Dixon *et al*, 2015 [163]. Plotted topologically associating domain (TAD) boundaries were obtained from the insulating score method [164] at 40-kb bin resolution. We searched for significant interactions between bins overlapping the glioma risk SNP and all other bins within 1Mb at each locus (*i.e.* "virtual 4C").

Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was carried out using version 3.0 with gene sets from Molecular Signatures Database (MSigDB) v6.0 [190][130], restricted to the C2 canonical pathways sets (*n*=1,329). Analysis was carried out using default settings, with the exception of removing restrictions on gene set size. RSEM normalised mRNASeq expression data for 20,501 genes in 676 glioma cases from TCGA were downloaded from the Broad Institute TCGA GDAC (http://gdac.broadinstitute.org/). These were assigned molecular groupings using sample information from Supplementary Table 1 of Ceccarelli *et al*, 2016 [179].

6.3 . Results

Descriptive characteristics of datasets

I studied three non-overlapping glioma case-control series of Northern European ancestry totalling 2,648 cases and 9,365 controls (Table 4.1). For 1,659 of the 2,648 cases information on tumour, 1p/19q co-deletion, *TERT* promoter and IDH mutation status was available (Figure 4.3). Using these data allowed definition of five molecular subgroups of glioma: Triple-positive (IDH mutated, 1p/19q co-deletion, *TERT* promoter mutated); *TERT*-IDH (IDH mutated, *TERT* promoter mutated, 1p/19q-wild-type); IDH-only (IDH mutated, 1p/19q wild-type, *TERT* promoter wild-type);*TERT*-only (*TERT* promoter mutated, IDH wild-type, 1p/19q wild-type) and Triple-negative (IDH wild-type, 1p/19q wild-type, *TERT* promoter wild-type).

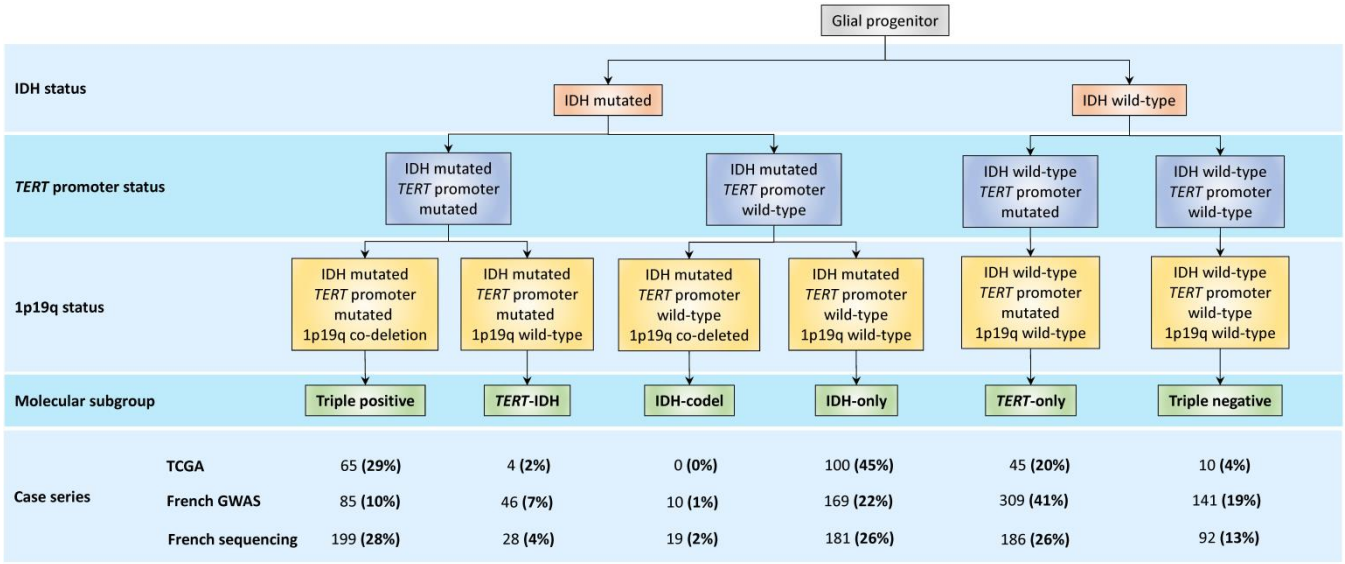


Figure 6.3 Molecular classification of diffuse glioma and frequency of each subgroup in the TCGA, French-GWAS and French sequencing case series.

As only 29 cases were classified as IDH mutation, 1p/19q co-deletion and *TERT* promoter wild-type, we restricted subsequent analyses to the five groups as above. Table 4.1 also shows grouping of the 1,960 cases adopting the WHO 2016 classification of glial tumours into five categories (Astrocytoma with IDH mutation, IDH wild-type astrocytoma, Oligodendroglioma with 1p/19q co-deletion, GBM with IDH mutation and IDH wild-type GBM) (APPENDIX 1; *page: 151*).

SNP selection

We analysed 25 SNPs, which had been reported to show the strongest genome-wide significant association with glioma in Chapter 3 meta-analysis of 12,496 cases and 18,190 controls [175] (Table

4.3). In the current analysis all of the SNPs exhibited a consistent direction of effect with that previously reported, albeit some weakly (APPENDIX 1; *page: 152*, online resource; Supplementary Table 3).

Relationship between risk SNP and molecular subgroup

In the first instance we examined whether the associations at the 25 risk loci were broadly defined by IDH status. We observed significant association for IDH mutated group with 1q44 (rs12076373), 2q33.3 (rs7572263), 3p14.1 (rs11706832), 8q24.21 (rs55705857), 11q21 (rs7107785), 11q23.3 (rs12803321), 14q12 (rs10131032), 15q24.2 (rs77633900) and 17p13.1 (rs78378222) risk SNPs. In addition, we found strong associations for the IDH wild-type glioma with 5p15.33 (rs10069690), 7p11.2 (rs75061358), 9p21.3 (rs634537), and 20q13.33 (rs2297440) (APPENDIX 1; *page: 152*, online resource; Supplementary Table 3). Of particular note was the finding that many of the risk loci recently discovered which were reported to be associated with non-GBM (1q44, 2q33.3, 3p14.1, 11q21, 14q12, 15q24.2) [175] showed a strong association with IDH mutant glioma.

Following on from this we performed a more detailed stratified analysis based on classifying the glioma tumours into the five molecularly defined groups. We found a strong association with IDH mutated tumours at 8q24.21 (rs55705857), in particular with Triple-positive glioma (P=1.27x10^{-37}, OR=9.30 [6.61-13.08]), which corresponds to the WHO 2016 oligodendroglioma classification (APPENDIX 1; *page 153*, online resource; Supplementary Table 3). Furthermore, we confirmed the previously reported associations at 5p15.33 (rs10069690), 9p21.3 (rs634537), 17p13.1 (rs78378222) and 20q13.33 (rs2297440) with *TERT*-only glioma in each of the three series [32]. Finally, we found suggestive evidence for an association between 22q13.1 (rs2235573) with *TERT*-only glioma, as well as 11q21 (rs7107785), 11q23.2 (rs648044), and 12q21.2 (rs1275600) with Triple-positive glioma (Figure 4.4**,** online resource; Supplementary Table 3).

Locus	SNP	Alleles	RAF	Reported subtype
1p31.3	rs12752552	**T**/C	0.87	GBM
1q32.1	rs4252707	G/**A**	0.22	Non-GBM
1q44	rs12076373	**G**/C	0.84	Non-GBM
2q33.3	rs7572263	**A**/G	0.76	Non-GBM
3p14.1	rs11706832	A/**C**	0.46	Non-GBM
5p15.33	rs10069690	C/**T**	0.28	GBM
7p11.2	rs75061358	T/**G**	0.10	GBM
7p11.2	rs11979158	**A**/G	0.83	GBM
8q24.21	rs55705857	A/**G**	0.06	Non-GBM
9p21.3	rs634537	T/**G**	0.41	GBM
10q24.33	rs11598018	**C**/A	0.46	Non-GBM
10q25.2	rs11196067	A/**T**	0.58	Non-GBM
11q14.1	rs11233250	**C**/T	0.87	GBM
11q21	rs7107785	**T**/C	0.48	Non-GBM
11q23.2	rs648044	**A**/G	0.39	Non-GBM
11q23.3	rs12803321	**G**/C	0.64	Non-GBM
12q21.2	rs1275600	**T**/A	0.60	Non-GBM
14q12	rs10131032	**G**/A	0.92	Non-GBM
15q24.2	rs77633900	G/**C**	0.09	Non-GBM
16p13.3	rs2562152	A/**T**	0.85	GBM
16p13.3	rs3751667	C/**T**	0.21	Non-GBM
16q12.1	rs10852606	T/**C**	0.71	GBM
17p13.1	rs78378222	T/**G**	0.01	All
20q13.33	rs2297440	T/**C**	0.80	GBM
22q13.1	rs2235573	**G**/A	0.51	GBM

Table 6.3 Overview of glioma risk SNPs at the 25 loci. The risk allele is emboldened and the minor allele underlined. The risk allele frequency (RAF) is from European samples from 1000 genomes project. Note: At 10q25.2, rs115997751 [175] failed sequencing so the originally reported SNP rs111960672 [92] was used.

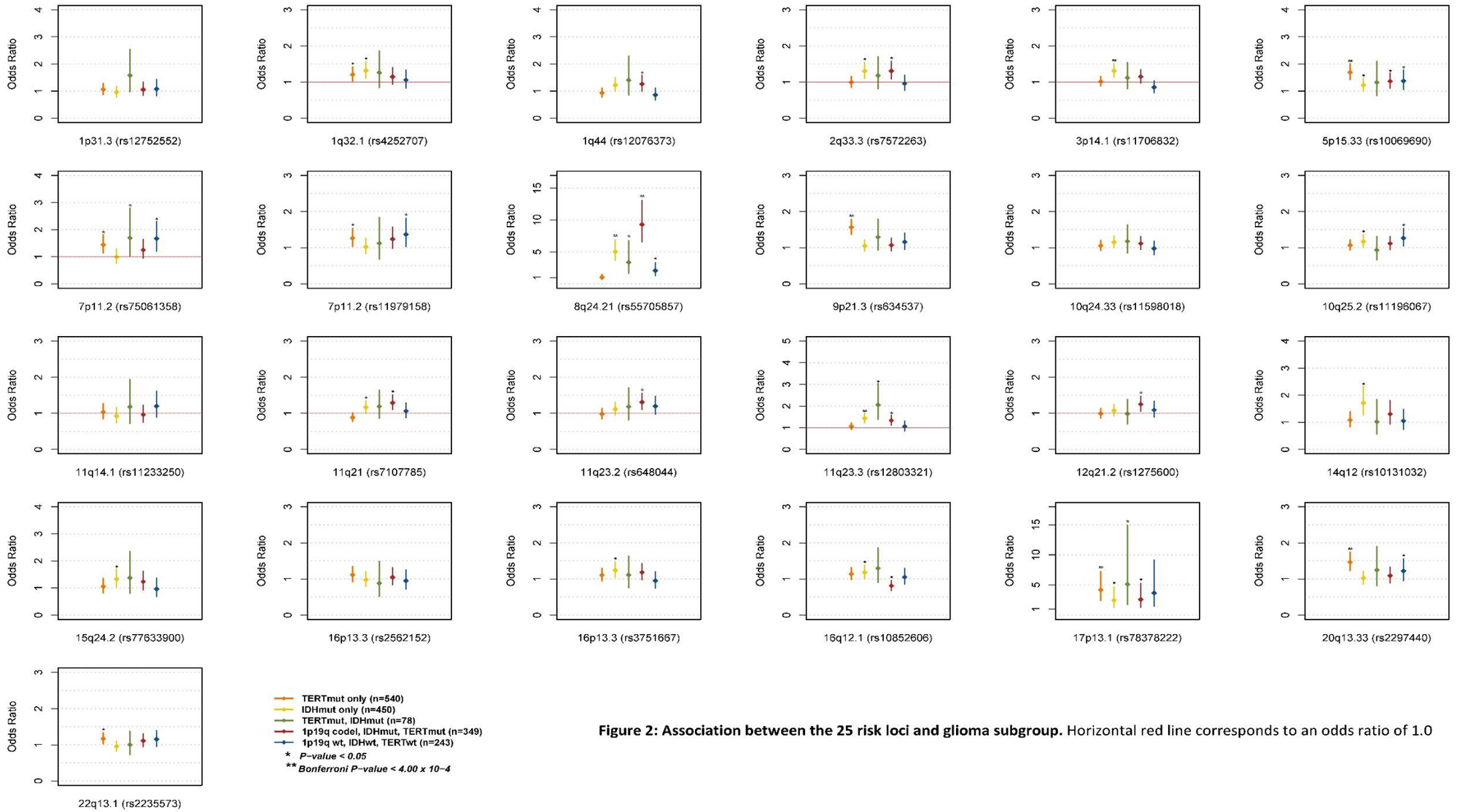


Figure 2: Association between the 25 risk loci and glioma subgroup. Horizontal red line corresponds to an odds ratio of 1.0

Figure 6.4 Association between the 25 risk loci and glioma molecular subgroup. Horizontal red line corresponds to an odds ratio of 1.0.

In addition to data on 1p/19q co-deletion, *TERT* promoter and IDH mutation, for 1,955 of the tumours we had information on *EGFR* amplification and *CDKN2A* deletion status (Table 4.1). Using these data we examined for an association with *EGFR* amplification and *CDKN2A* deletion, particularly focusing on the 7p11.2 (rs75061358 and rs11979158) and 9p21.3 (rs634537) risk SNPs in view of the fact that these loci map in or near *EGFR* and *CDKN2A* respectively (APPENDIX 1; *page 154-155*, online resource; Supplementary Table 3). At 7p11.2, the intergenic variant rs75061358, which is located in the genomic vicinity of *EGFR,* was associated with *EGFR* amplified tumours and not those without amplification. There was a less strong association with EGFR amplification seen with the second independent signal at the locus defined by rs11979158, which is intronic within *EGFR* itself. At 9p21.3 rs634537, which is intronic within *CDKN2B-AS1* and in the vicinity of *CDKN2A* and *CDKN2B*, was not associated with *CDKN2A* deletion status. Low grade gliomas tend to be *EGFR* wild-type and *p16* wild-type tumours, and therefore as anticipated many non-GBM risk SNPs were most strongly associated with these tumours; notably 2q33.3 (rs7572263), 3p14.1 (rs11706832), 8q24.21 (rs55705857), 10q25.2 (rs11196067), 11q23.3 (rs12803321) (APPENDIX 1; *page 154-155*, online resource; Supplementary Table 3).

Polygenic contribution to age at diagnosis and patient survival

Patient survival by molecular subgroup in each of the three series was consistent with previous published reports [20][32]; specifically, patients with Triple-positive tumours had the best prognosis whilst those with *TERT*-only tumours had the worst outcome (APPENDIX 1; *page 148-150*). We investigated whether an increased burden of glioma risk alleles might be associated with earlier age at diagnosis (*i.e*. indicative of influence on glioma initiation) or survival (indicative of influence on glioma progression). There was a slight albeit, non-significant trend towards decreased age at diagnosis with increased risk allele number in the IDH-only, *TERT*-only and Triple-positive molecular subgroup, but with decreased risk allele number in the *TERT*-IDH and Triple-negative tumours (APPENDIX 1; *page 156-158*). We found no overall relationship between age and risk allele number, or for the individual molecular groups (APPENDIX 1; *page 174*). Examining each SNP individually, only rs55705857 at 8q24.21 was nominally associated with age (APPENDIX 1; *page 174*)

We used Cox Proportional-Hazards Regression to investigate whether burden of glioma risk was associated with survival, with each risk allele coded as 0, 1 or 2. As expected, age, grade and all molecular group (Triple-negative, Triple-positive, *TERT*-only, IDH-only and *TERT*-IDH) were strongly associated with decreased survival. Intriguingly, the number of risk alleles was associated with increased survival (APPENDIX 1; *page 175*; $P<10^{-4}$) with 1q32.1 (rs4252707), 11q23.3 (rs12803321) and

11q21 (rs7107785) each being nominally associated with survival, independent of age and molecular subgroup. Considering the relationship between burden of glioma risk alleles and survival in each molecular subgroup a consistent association with increased survival was shown in Triple-positive, Triple-negative and *TERT*-only molecular groups but not in IDH-only and *TERT*-IDH groups.

Biological inference of risk loci

Since genomic spatial proximity and chromatin looping interactions are fundamental for regulation of gene expression [191], we interrogated physical interactions at respective risk loci in embryonic stem cells and neuronal progenitor cells using Hi-C data. We also sought to gain insight into the possible biological mechanisms for associations by performing expression quantitative trait locus (eQTL) analysis using mRNA expression data in 10 brain regions using the GTEx portal.

We identified significant Hi-C contacts from the genomic regions which encompass 14 of the 25 risk loci implicating a number of presumptive candidate genes. For two of these, candidacy was supported by eQTL data. (Table 4.4; Online Resource,Supplementary Table 6). Notably at 2q33.3, there was a significant looping interaction between the risk SNP and *IDH1*/*IDH1-AS1* and *LRIG1* at 3p14.1 (Figure 4.5), as well as with *EGFR/EGFR-AS1* at 7p11.2, *CDKN2A/CDKN2B* at 9p21.3, *NFASC* at 1q32.1 (APPENDIX 1, *page 159-172*). At the 8q24.21 gene desert Hi-C data revealed a significant interaction between the risk SNP rs55705857 and *MYC*, as well as lincRNAs in the region such as *PCAT1/PCAT2*. Additionally, the risk SNP rs12803321 at 11q23.3 was significantly associated with *PHLDB1* expression in the brain.

Pathway analysis

To potentially gain further insight into the biological basis of subtype associations, we performed a gene-set enrichment analysis (GSEA) analysing gene expression data from TCGA (online resource; Supplementary Table 7) While we did not identify any significantly altered gene sets (at FDR *q*-value <0.1), the most significantly expressed genes in subgroups was upregulation of PI3K signalling shown in 1p/19q co-deleted tumours (online resource; Supplementary Table 7).

		Glioma molecular classification grouping						
Locus	**SNP**	**Molecular subgroup**	**IDH group**	***EGFR* group**	***CDKN2A* group**	**eQTL**	**Hi-C**	**Commentary**
1p31.3	rs12752552	*TERT*-IDH (ns)	-	-	-	*JAK1* (brain)	*RAVER2* *JAK1* *UBE2U* *CACHD1*	JAK1 is involved in actomyosin contractility in tumour cells and stroma to aid metastasis [192].
1q32.1	rs4252707	*TERT*-only* IDH-only*	IDHmut*	*EGFR*wt*	*CDKN2A*wt*	-	*NFASC*	NFASC is a cell adhesion molecule involved in axon subcellular targeting and synapse formation during neural development [193].
1q44	rs12076373	TP*	IDHmut**	-	-	-	*AKT3* *ZBTB18* *SDCCAG8*	AKT3 is highly expressed in brain, regulates cell signalling in response to insulin and growth factors [194], involved in regulation of normal brain size [195].
2q33.3	rs7572263	IDH-only* TP*	IDHmut**	*EGFR*wt*	*CDKN2A*wt*	-	*IDH1* *IDH1-AS1*	Overexpression of IDH mutant proteins renders glioma cells more sensitive to radiation [196].
3p14.1	rs11706832	IDH-only**	IDHmut**	*EGFR*wt*	*CDKN2A*wt*	*LRIG1* (blood) *SLC25A26* (blood)	*LRIG1*	-
5p15.33	rs10069690	*TERT*-only** IDH-only* TP* TN*	IDHmut* IDHwt**	*EGFR*amp** *EGFR*wt*	*CDKN2A*del* *CDKN2A*wt**	-	-	rs10069690 affects *TERT* splicing [102].
7p11.2	rs75061358	*TERT*-only* *TERT*-IDH* TN*	IDHwt**	*EGFR*amp**	*CDKN2A*wt*	-	-	-
7p11.2	rs11979158	*TERT*-only* TN*	IDHwt*	*EGFR*amp* *EGFR*wt*	*CDKN2A*del* *CDKN2A*wt*	-	*EGFR* *EGFR-AS1*	-
8q24.21	rs55705857	IDH-only** *TERT*-IDH* TP** TN*	IDHmut**	*EGFRwt***	*CDKN2A*wt** *CDKN2A*del**	-	*PCAT1* *PCAT2* *CASC8* *CASC11* *MYC* *PVT1*	-
9p21.3	rs634537	*TERT*-only**	IDHwt**	*EGFR*amp* *EGFR*wt*	*CDKN2A*del* *CDKN2A*wt**	-	*CDKN2A* *CDKN2B-AS1*	-

10q24.33	rs11598018	-	IDHmut*	*EGFR*wt*	-	-	*GSTO1* *GSTO2* *SH3PXD2A*	Correlated SNP to rs11598018 associated with telomere length likely through *OBFC1* [197].
10q25.2	rs11196067	IDH-only* TN*	IDHmut* IDHwt*	*EGFR*wt*	*CDKN2A*wt*	-	*TCF7L2* *VTI1A* *HABP2*	TCF7L2 modifies beta-catenin signalling and controls oligodendrocyte differentiation [198].
11q14.1	rs11233250	-	-	-	-	-	-	-
11q21	rs7107785	IDH-only* TP*	IDHmut**	*EGFR*wt*	*CDKN2A*del*	*RP11-712B9.2* (brain)	-	-
11q23.2	rs648044	TP*	IDHmut*	*EGFR*wt**	*CDKN2A*wt**	-	*NNMT* *ZBTB16*	NNMT is upregulated in GBM, NAD metabolism important in glioma [199].
11q23.3	rs12803321	IDH-only** *TERT*-IDH* TP*	IDHmut**	*EGFR*wt**	*CDKN2A*wt** *CDKN2A*del*	*PHLDB1* (brain)	-	PHLDB1 is an insulin-responsive protein that enhances Akt activation [200].
12q21.2	rs1275600	TP*	IDHmut*	*EGFR*wt*	*CDKN2A*del*		*KRR1* *GLIPR1*	GLIPR1 is targeted by TP53 [201].
14q12	rs10131032	IDH-only*	IDHmut**	*EGFR*wt*	*CDKN2A*del* *CDKN2A*wt*		*NPAS3*	NPAS3 is a tumour suppressor for astrocytoma [202].
15q24.2	rs77633900	IDH-only*	IDHmut**	*EGFR*wt*	*CDKN2A*wt*	-	*SCAPER*	-
16p13.3	rs2562152	-	-	-	-	-	-	-
16p13.3	rs3751667	IDH-only*	IDHmut*	*EGFR*amp* *EGFR*wt*	*CDKN2A*wt*	*RP11-161M6.2* (brain) *SOX8* (blood)	-	SOX8 is strongly expressed in brain and may be involved in neural development [203].
16q12.1	rs10852606	IDH-only* TP* (-ve)	-	-	-	*HEATR3*	-	HEATR3 may be involved in NOD2-mediated NF-kappa B signalling [204].
17p13.1	rs78378222	*TERT*-only** IDH-only* *TERT*-IDH* TP*	IDHmut** IDHwt*	*EGFR*amp* *EGFR*wt**	*CDKN2A*wt** *CDKN2A*del*	-	-	SNP rs78378222 affects *TP53* 3'UTR poly-adenylation processing [88].
20q13.33	rs2297440	*TERT*-only** TN*	IDHwt**	*EGFR*amp** *EGFR*wt*	*CDKN2A*del* *CDKN2A*wt*	*STMN3* (brain) *LIME1* (blood) *ZGPAT* (blood) *EEF1A2* (blood)	-	Overexpression of STMN3 promotes growth in GBM cells [205].
22q13.1	rs2235573	*TERT*-only*	IDHwt*	-	-	*CTA-228A9.3* (brain)	-	-

Table 6.4 Candidate gene basis of glioma risk loci. ns, non-significant; PMID, PubMed identifier; TN, triple negative (*i.e.* IDH-wildtype, *TERT* promoter wildtype, 1p/19q wildtype); TP, triple positive (*i.e.* IDH-mutation, *TERT* promoter mutation and 1p/19q co-deletion). * $P<0.05$; ** significant after adjustment for multiple comparisons.

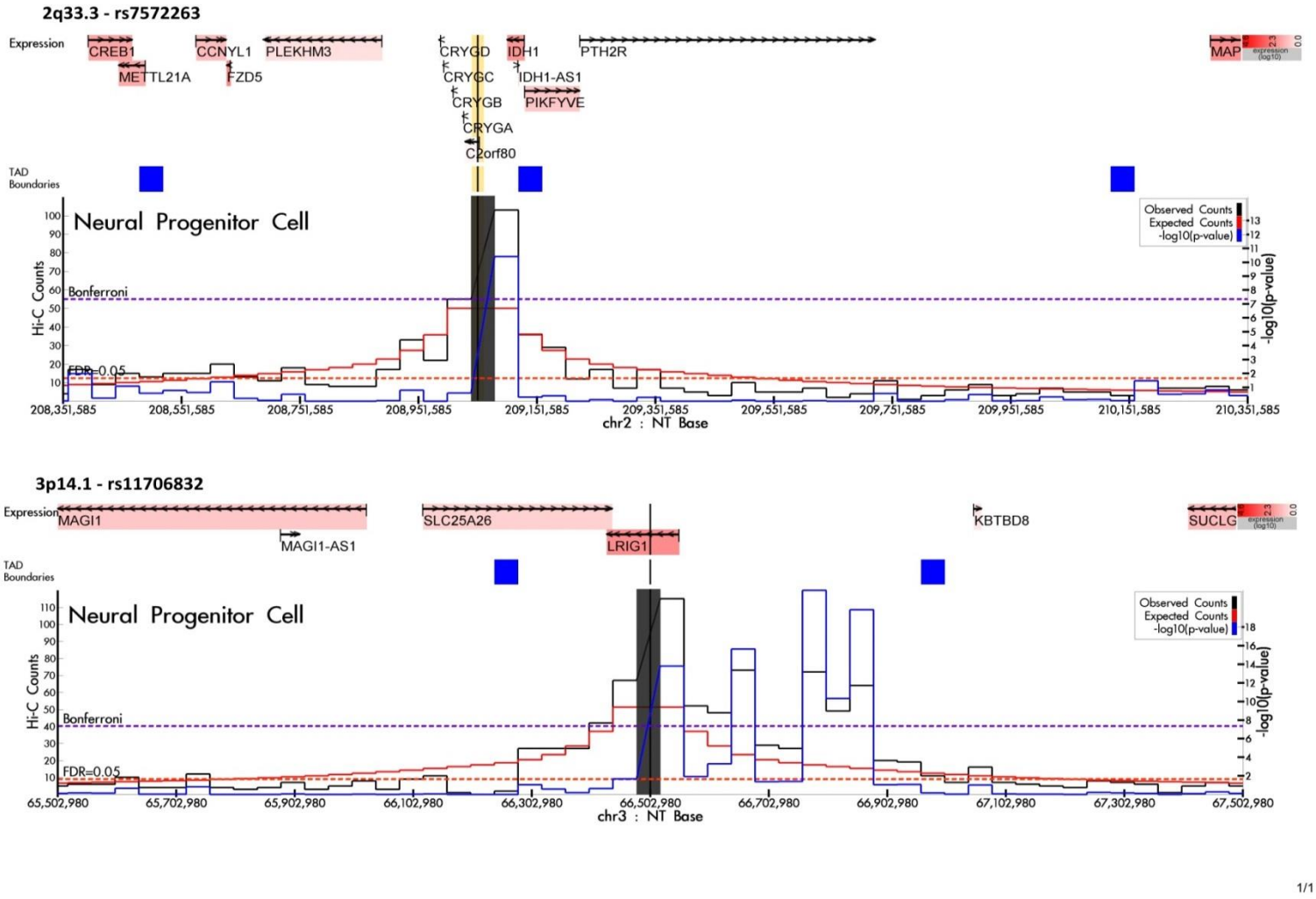


Figure 6.5 Plots of Hi-C interactions in H1 neuronal progenitor cells at the 2q33.3 and 3p14.1 risk loci. Plots were generated using the HUGIn browser [161]. Each plot shows a "virtual 4C" of all Hi-C interactions with "bait" fragments overlapping the glioma risk SNP of interest (indicated by the shaded rectangle).Topologically associating domain (TAD) boundaries are plotted as filled blue rectangles. The purple dotted line represents the Bonferroni threshold, with interactions exceeding this threshold treated as statistically significant.

6.4 .Discussion

These findings provide further support for subtype specific associations for glioma risk loci. Specifically, we confirm the strong relationship between the 8q24.21 (rs55705857) risk variant and Triple-positive glioma. Moreover, we substantiate the proposed specific associations between 5p15.33 (rs10069690) and 20q13.33 (rs2297440) variants with *TERT* promoter mutations, 9p21.3 (rs634537) with *TERT*-only glioma, as well as 17p13.1 (rs78378222) with *TERT*-IDH glioma. Other loci such as 1q32.1 (rs4252707) and 10q25.2 (rs11196067) appear to have more generic effects.

Although preliminary, and in part speculative, our analysis delineates potential candidate disease mechanisms across the 25 glioma risk loci (Table 4.4; Figure. 4.5). Firstly, maintenance of telomeres is central to cell immortalization [206], and is generally considered to require mutually exclusive mutations in either the *TERT* promoter or *ATRX*. The risk alleles at 5p15.33 (*TERT*) and 10q24.33 (*OBFC1*) are associated with increased leukocyte telomere length, thereby supporting a relationship between SNP genotype and biology [206][207][208]. While dysregulation of the telomere gene *RTEL1* has traditionally been assumed to represent the functional basis of the 20q13.33 locus, the glioma risk SNP does not map to the locus associated with telomere length [175][197] Intriguingly, our analysis instead implicates *STMN3* at 20q13.33, whose over-expression promotes growth in GBM cells [205] suggesting an alternative mechanism by which the risk SNP influences glioma development. With respect to the 5p15.33 (*TERT*) and 10q24.33 (*OBFC1*) loci, it is unclear whether the effect on glioma risk is solely due to telomeres or is pleiotropic and involves multiple factors. For example, rs10069690 at 5p15.33 is strongly associated with *TERT*-only glioma, yet the *TERT* promoter mutation increases telomerase activity without necessarily affecting telomere length [179]. An intriguing hypothesis to test would therefore be to examine the impact of allele-specific effects of rs10069690 on telomere length in the context of gliomas carrying the *TERT* promoter mutation.

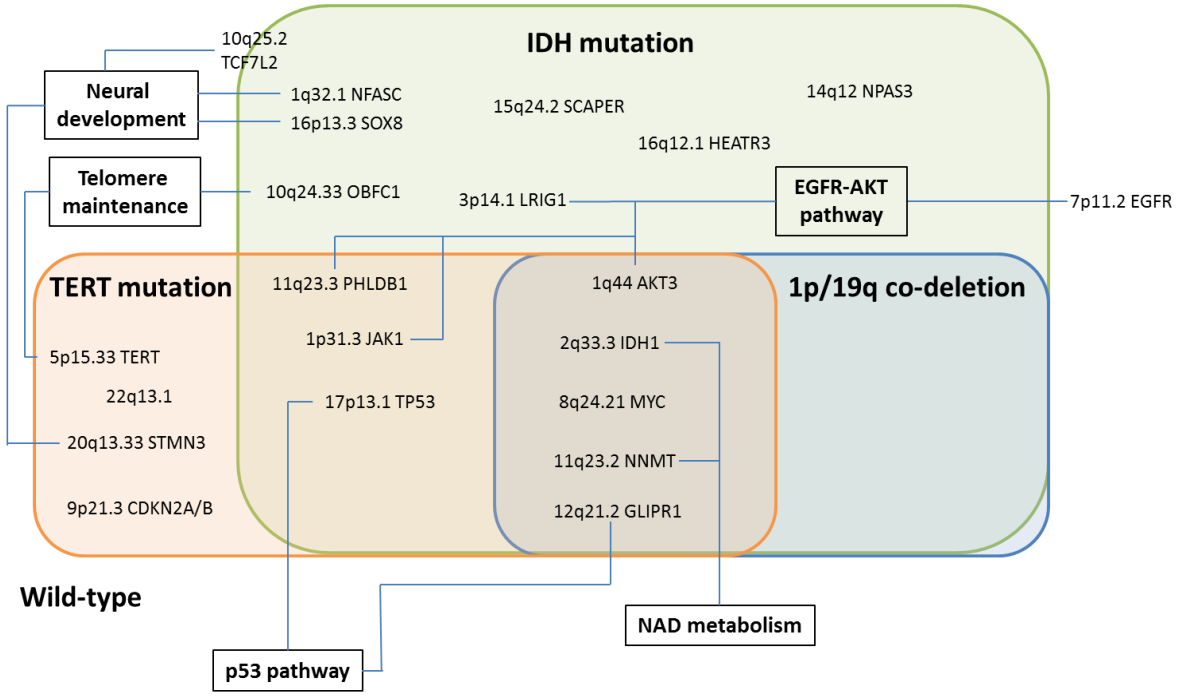


Figure 6.6 Summary of the relationship between glioma risk with molecular subgroup and associated biological pathways. The extent of the evidence supporting each candidate gene (ranging from an established role in glioma to largely speculative) is summarised in Table 4.4.

Secondly, the *EGFR-AKT* pathway involves *EGFR* at 7p11.2, *LRIG1* at 3p14.1, *PHLDB1* at 11q23.3 and *AKT3* at 1q44. We showed a significant interaction between the risk SNP rs11979158 at 7p11.2 and *EGFR*, consistent with a cis-regulatory effect on gene expression. Although the mechanistic basis of the 7p11.2 locus has long been suspected to involve *EGFR* and is highly associated with classical GBM, emerging evidence suggests that additional components of the *EGFR-AKT* signalling pathway are implicated by non-GBM SNPs. At the IDH-only associated locus 3p14.1, *LRIG1* is highly expressed in the brain and negatively regulates the epidermal growth factor receptor (EGFR) signalling pathway [209]. Reduced *LRIG1* expression is linked to tumour aggressiveness, temozolomide resistance and radio-resistance [210][211]. Downstream components of *EGFR-AKT* signalling are implicated at 11q23.3 via *PHLDB1*, as well as 1p31.3 via *JAK1* and 1q44 via *AKT3*. The risk allele of rs12803321 is associated with increased expression of *PHLDB1*, an insulin-responsive protein that enhances Akt activation [200]. *AKT3* at 1q44 is highly expressed in the brain and appears to respond to EGF in a PI3K dependent manner [212], with GBM cells containing amplified AKT3 having enhanced DNA repair and resistance to radiation and temozolomide [213]. The risk allele of rs12752552 at 1p31.3 is associated with increased *JAK1* expression in brain tissue. Since *JAK1* can be activated by EGF phosphorylation, it may be involved in astrocyte formation [214][215][216]. The 3p14.1 and 11q23.3

loci are strongly associated with *EGFR* amplification negative gliomas, with a consistent albeit non-significant trend at 1p31.3 and 1q44, consistent with elevated upstream *EGFR* activation masking their functional effects.

Thirdly, the *NAD* pathway involves *IDH1* at 2q33.3 and *NNMT* at 11q23.2. At 2q33.3 we detected a significant Hi-C interaction between the glioma risk SNP rs7572263 and *IDH1*/*IDH1-AS1.* Overexpression of *IDH1* mutant proteins has been reported to sensitize glioma cells to radiation [196] providing an interesting mechanism to test the allele-specific effects of this SNP. IDH mutation causes de-regulation of NAD signalling [17]. Interestingly therefore, at 11q23.2 which is strongly associated with IDH mutated gliomas, the most convincing molecular mechanism is via *NNMT*, which encodes nicotinamide N-methyltransferase and is highly expressed in GBM relative to normal brain, causing methionine depletion-mediated DNA hypomethylation and accelerated tumour growth [199][217].

Fourthly, genes with established roles in neural development may be involved. While the risk SNP rs4252707 at 1q32.1 is within the intron of *MDM4*, the strongest evidence for a mechanistic effect was with *NFASC*. Neurofascin is involved in synapse formation during neural development [193] and therefore represents an attractive functional candidate for the association with glioma. Additionally at 16p13.3 and 20q13.33, implicated genes *SOX8* and *STMN3* are strongly expressed in the brain and thought to play a role in neural development [203][205]. At 10q25.2, implicated gene *TCF7L2* modifies beta-catenin signalling and controls oligodendrocyte differentiation [198]. Intriguingly, 10q25.2 has previously been reported to be a risk locus for colorectal cancer [218], a tumour driven by wnt signalling, however the risk SNP is not correlated with rs11196067 raising the possibility of tissue-specific regulation across the wider region.
Finally, the p53 pathway is involved at 17p13.1, where the risk SNP rs7837222 affects *TP53* 3'UTR poly-adenylation processing. In addition the p53 target GLIPR1 [201] is implicated at 12q21.2. Moreover, 12q21.2 is most strongly associated with Triple-positive glioma, which does not feature *TP53* mutation, consistent with wild-type p53 protein being required for the SNP to exert a functional effect.

As with many cancers, the exact point at which the risk SNPs exert their functional impact on glioma oncogenesis still remains to be elucidated, and we did not demonstrate a relationship between

increased risk allele number and age at diagnosis. Surprisingly we found a significant association between increasing risk allele number and improved outcome. This result was consistent across the prognostic molecular groups, consistent with our observations not being due to an over-representation of the more favourable prognostic groups among patients with a higher burden of risk alleles. In addition, the distribution of risk allele numbers did not differ across the four groups (*P*=0.3, ANOVA test). Examining the impact of an individual SNP's impact on survival did not reveal any loci strongly associated with outcome. Collectively our findings suggest that, independent of other prognostic factors, the greater the number of risk alleles carried, the better the outcome.

In conclusion, we performed the most comprehensive association study between molecular subgroup and the 25 recently identified glioma risk loci to date. While confirming previous observations, we show that the majority of risk loci are associated with IDH mutation. Through integration of Hi-C and eQTL data we have additionally sought to define candidate target genes underlying the associations. Collectively our observations highlight pathways critical to glioma susceptibility, notably neural development and NAD metabolism, as well as EGFR-AKT signalling. Intriguingly, we show here that the number of risk alleles is consistently associated with better outcome. Functional investigation in tumour and neural progenitor-based systems will be required to more fully elucidate these molecular mechanisms. Notably, IDH mutant tumours have been shown to reshape 3D chromatin organisation and may reveal new regulatory interactions [219].

CHAPTER 5

TCF12 is mutated in anaplastic oligodendroglioma

7.1 .Overview and rational

Anaplastic oligodendrogliomas (AO; World Health Organization grade III oligodendrogliomas) are rare primary malignant brain tumours with a highly variable overall prognosis. The genomic instability in cancer is characterised by somatic genetic mutation, including single-nucleotide variants (SNV), small-scale insertion-deletions (indels), large-scale somatic copy number alterations (sCNA) and genomic translocations.

The emblematic molecular alteration in oligodendrogliomas is 1p/19q co-deletion, which is associated with a better prognosis and response to early chemotherapy with procarbazine, lomustine and vincristine (PVC) [27][41][220]. Prior to the work presented in this thesis, recent high-throughput sequencing approaches have identified IDH (*IDH1* and *IDH2*), *CIC*, *FUBP1* and *TERT* promoter mutations in oligodendroglioma (75%, 50%, 10% and 75%, respectively) [27][28][221]; IDH mutation status typically being associated with a better clinical outcome [17]. Identifying additional driver genes and altered pathways in oligodendroglioma offers the prospect of developing more effective therapies and biomarkers to predict individual patient outcome.

Here I performed whole-exome and transcriptome sequencing analysis of AO to search for additional tumour driver mutations and pathways disrupted.

The results of this Chapter have been published (APPENDIX 2). Therefore, due to the format, some data are available in the online version of the paper.

7.2 .Methods

7.2.1 Patients, samples and datasets

The Exome sequencing was conducted on samples from 51 AO patients (33 male; median age 49 years at diagnosis, range 27-81), as detailed in 2.1.3 and 2.2.6. For targeted follow-up analyses we studied the tumours from an additional 83 AO patients and 75 patients with grade II tumours. A summary of each of the tumour cohorts and respective pathological information on the patients is provided in online resource: Supplementary Data 1.

Additionally, to explore the mutational spectra of AO in an independent series I made use of data generated by The Cancer Genome Atlas (TCGA) study of low grade glioma, which provides exome sequencing data on a further 43 AO tumours.

7.2.2 Statistical and bioinformatics analysis

Sequence alignment, mapping, and variant calling performed using BWA/Stampy/GATK/MuTect software, as detailed in 2.3.2.3 and 2.3.6. Copy number variation (CNV) analysis was conducted using SNP array as detailed in 2.3.6.2. Pathway analysis was performed as described in 2.3.6.1 using Oncodrive-fm [167] as implemented within the IntOGen-mutations platform [168], using all SNVs and indel mutations called across the 51 tumours.

Gene expression profiles of 71 samples were analysed using Affymetrix Human Genome U133 Plus 2.0 arrays. All samples were normalized in batch using the RMA algorithm (Bioconductor *affy* package), and probe set intensities were then averaged per gene symbol.

To identify the significantly mutated pathways, gene set member lists were retrieved online from MSigDB33, GO34 and SMD35 databases. We searched for gene sets harbouring more damaging mutations than expected by chance. Given the set G of all the genes sequenced with sufficient coverage, the set S of tumour samples (of size n) and any gene set P, we calculated the probability of observing a number of mutations equal or greater to that observed in P across the n samples according to a binomial law B(k, p), with $k = n \times L(P)$ and the mutation rate $p = A(G,S) / (n \times L(G))$, where L(X) is the sum of the lengths (in bp) of all genes/exons from a gene set X, and A(G, S) is the total number of mutations observed in all the targeted sequences across all the samples from S.

7.3 .Results

In accordance with conventional clinical practice I considered three molecular subtypes for our analyses: (i) IDH mutated 1p/19q co-deleted (IDHmut-codel); (ii) IDH mutated 1p/19q non-co-deleted (IDHmut-non-codel) and (iii) IDH-wildtype (IDHwt)[7]. Assignment of IDH mutated (defined by *IDH1* R132 or *IDH2* R172 mutations), 1p/19q and *TERT* promoter mutation (defined by C228T or C250T) status in tumours was determined using conventional sequencing and SNP array methods.

Mutational landscape

Whole exome sequencing of 51 AO tumours and matched germline DNA were performed, targeting 318,362 exons from 18,901 genes. The mean sequencing coverage across targeted bases was 57x, with 80% of target bases above 20x coverage (Figure 5.1). We identified a total of 4,733 mutations (with a mean of 37 non-silent mutations per sample) equating to a mean somatic mutation rate of 1.62 mutations per megabase (Mb) (Figure 5.2). Although the tumours of two patients (3063 and 3149) had high rates of mutation (9.1 and 12.4 respectively) this was not reflective of tumour site (both frontal lesions as were 68% of the whole series) or treatment. Excluding these two cases the mean rate of non-silent mutations per tumour was 33±14, which is similar to the number found in most common adult brain tumours. The mutation spectrum in AO tumours was characterized by a predominance of C>T transitions, as observed in most solid cancers (Figure 5.2) [166][222]. While few of the tumours were IDH*wt,* these did not harbour a significantly higher number of mutations compared to IDH*mut-1p/19q co-deleted* and IDH*mut-non-1p/19q co-deleted* tumours (Figure 5.2). Intriguingly one tumour (2688) was co-mutated for IDH1 (R132H) and IDH2 (P162S), but exhibited no distinguishing phenotype in terms of clinico-pathology or mutation rate.

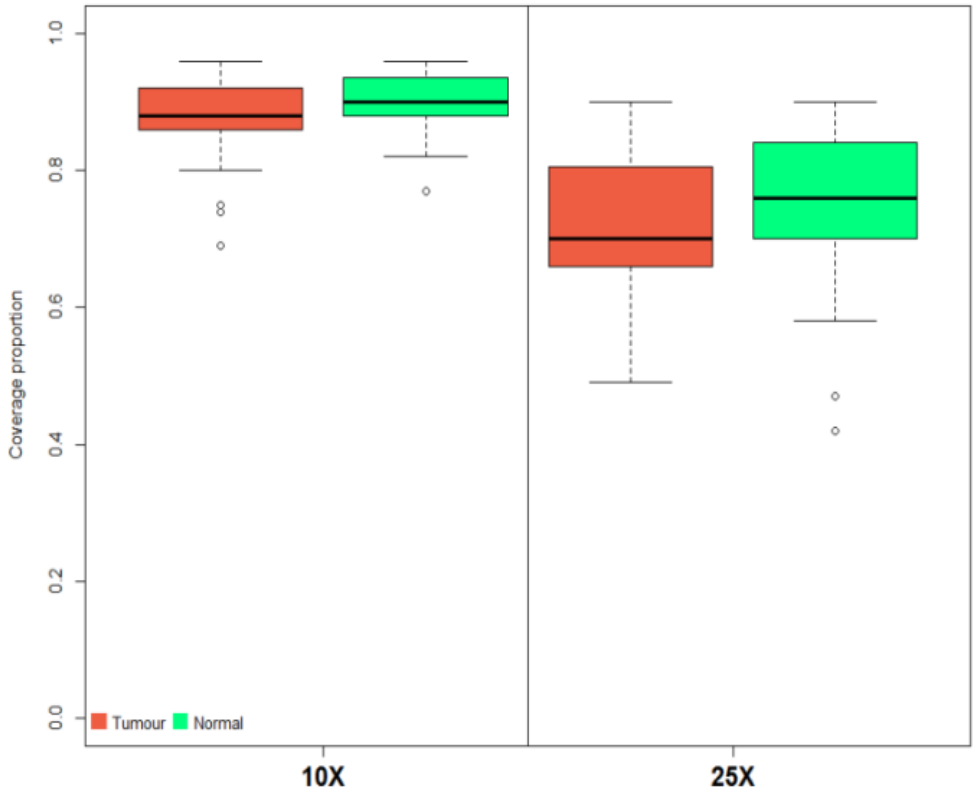


Figure 7.1 Coverage of exome sequencing. Proportion of bases in targeted exons sequenced at a depth of 10× and 25× for 51 AOs tumours and their normal counterparts. Boxes divided by median values. Length of boxes corresponds to interquartile range and whiskers correspond to 1.5 interquartile ranges.

I used MutSigCV version 1.4 [166] to identify genes harbouring more non-synonymous mutations than expected by chance given gene size, sequence context and mutation rate of each tumour for the three molecular subtypes, respectively. As expected we observed frequent mutations of the tumour suppressors *FUBP1* (22%) located on 1p, and *CIC* (32%) located on 19q, which have been reported in the context of 1p/19q co-deletion; these were not mutually exclusive events (Figure 5.2). Also within the IDHmut-codel group, 37 of tumours tested carried *TERT* C228T or C250T promoter mutations (72%); none of which also carried an *ATRX* mutation, concordant with the previously reported finding that these are mutually exclusive events [27].

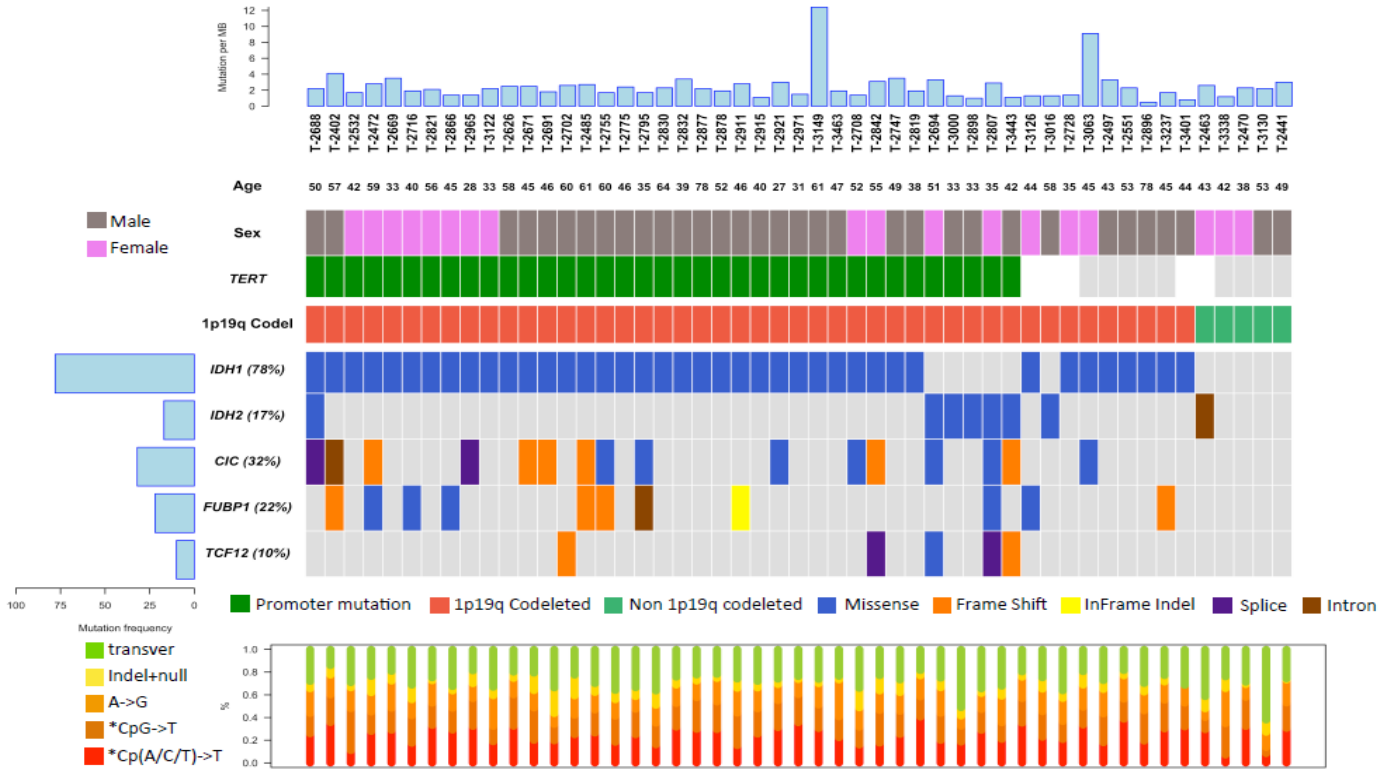


Figure 7.2 Significantly mutated genes in anaplastic oligodendroglioma by molecular subtype. Significantly mutated genes (Q-value<0.1) identified by exome sequencing are listed by Q-value. The percentage of AO samples with mutation detected by automated calling is detailed on the left. Samples are displayed as columns, with the mutation rate plotted at the top. Samples are arranged to emphasize mutual exclusivity. Mutation types are indicated in different colours (see legend). White colour indicates no information available. Also shown is the relative proportion of base-pair substitutions within mutation categories for each tumour.

In addition to mutation of *IDH1* (78%), *IDH2* (17%), *CIC* (32%), and *FUBP1* (22%), *TCF12* was also significantly mutated (*Q* value <0.1; Figure 5.2; Online resource: Supplementary Data 2). Heterozygous somatic mutations in *TCF12,* which encodes the basic helix-loop-helix (bHLH) transcription factor 12 (aliases *HEB*, *HTF4*, *ALF1*) were identified in five (1 missense, R602M; 2 splice-site, c.825+5G>T, c.1979-3_1979-delTA and 2 frameshift, E548fs*13, S682fs*14) of the 46 IDHmutated-1p/19q co-deleted (Figure 5.3). Intriguingly germline mutations of residues E548R and 602M have been previously shown to cause coronal craniosynostosis [223].

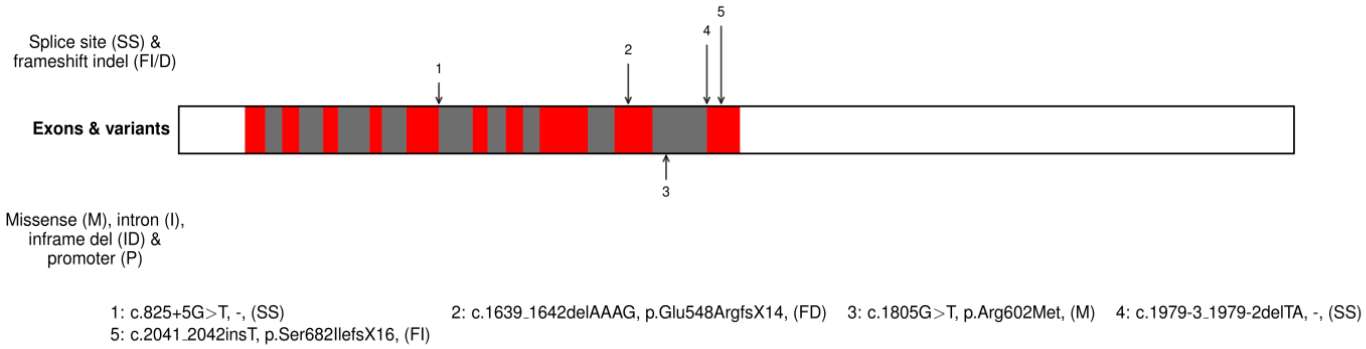


Figure 7.3 Location of mutations of TCF12 in AO. Transcripts are plotted 5' to 3'; untranslated regions are not colored; coding regions of exons are shown in alternating red and gray. The variants track shows the distribution of mutations.

The availability of high quality tumour material allowed us to generate SNP array and expression data on 31 of the cases exome sequenced. In addition to co-deletion of chromosome arms 1p/19q we identified several other recurrent genomic alterations - mainly loses of chromosomes 4 (29%), 9p (28%), and 14q (19%) (Figure 5.4; Table5.1). Notably, tumours featuring mutation of Notch-pathway genes showed significant chromosome 4 loss (*P*=0.02, Chi squared test).

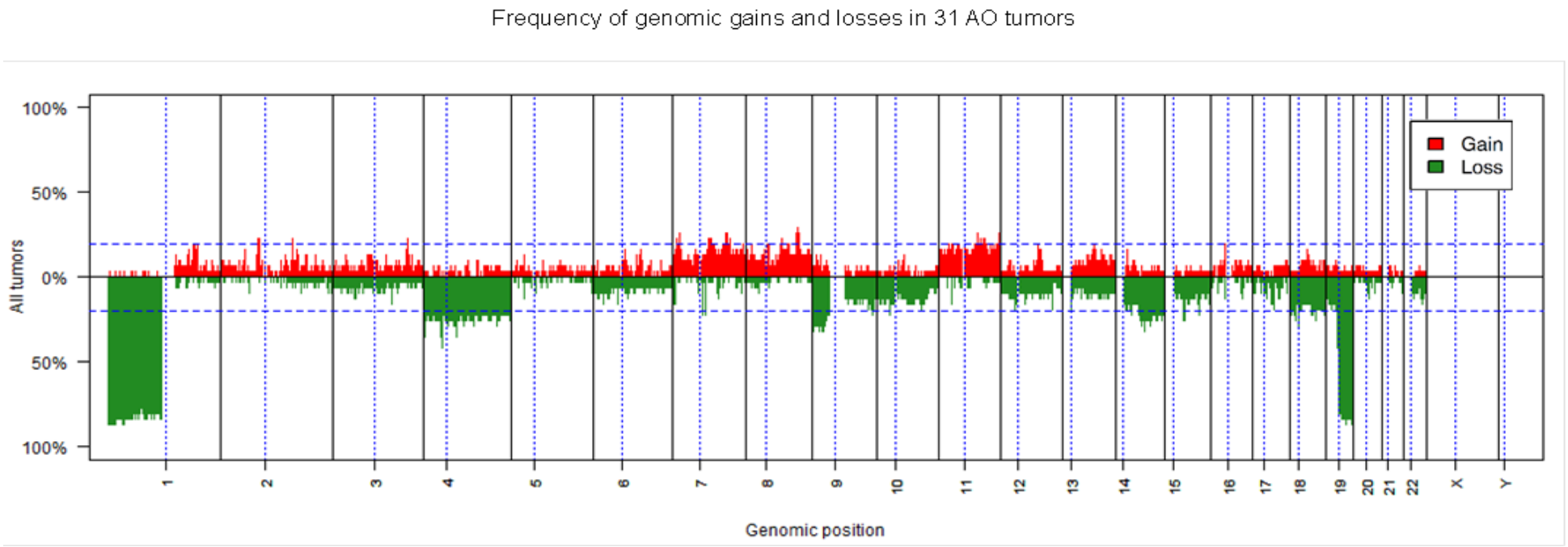


Figure 7.4 Frequency of genomic gains and losses in 31 AO samples. Vertical solid lines separate chromosomes, and vertical dashed lines indicate centromeres positions. Gains and losses frequency peaks were computed for each genomic position targeted by SNP arrays (excluding sexual chromosomes and positions within known frequent germline CNVs).

Arm	# Genes	Amp frequency	Amp z-score	Amp q-value	Del Frequency	Del z-score	Del q-value
1p	2121	0	-0.689	0.951	0.84	14.9	0
1q	1955	0.08	-0.177	0.951	0.17	1.64	0.194
2p	924	0.13	0.84	0.712	0	-1.6	0.951
2q	1556	0.13	0.838	0.712	0	-1.6	0.951
3p	1062	0.03	-0.998	0.951	0.07	-0.388	0.951
3q	1139	0.03	-0.998	0.951	0.07	-0.388	0.951
4p	489	0	-1.44	0.951	0.29	4.04	0.000215
4q	1049	0	-1.44	0.951	0.29	4.03	0.000215
5p	270	0.1	0.204	0.951	0	-1.63	0.951
5q	1427	0.06	-0.438	0.951	0	-1.66	0.951
6p	1173	0	-1.66	0.951	0.06	-0.437	0.951
6q	839	0	-1.66	0.951	0.06	-0.437	0.951
7p	641	0.16	1.48	0.453	0	-1.57	0.951
7q	1277	0.16	1.48	0.453	0	-1.57	0.951
8p	580	0.16	1.48	0.453	0	-1.57	0.951
8q	859	0.16	1.48	0.453	0	-1.57	0.951
9p	422	0.09	0.00747	0.951	0.28	3.63	0.000932
9q	1113	0.08	-0.057	0.951	0.24	2.96	0.00848
10p	409	0	-1.57	0.951	0.16	1.48	0.194
10q	1268	0	-1.57	0.951	0.16	1.48	0.194
11p	862	0.23	2.76	0.0574	0	-1.51	0.951
11q	1515	0.23	2.75	0.0574	0	-1.51	0.951
12p	575	0	-1.63	0.951	0.1	0.203	0.82
12q	1447	0	-1.63	0.951	0.1	0.2	0.82
13q	654	0.14	1.06	0.712	0.11	0.454	0.745
14q	1341	0	-1.54	0.951	0.19	2.12	0.0744
15q	1355	0.04	-0.873	0.951	0.17	1.56	0.194
16p	872	0.14	0.983	0.712	0.07	-0.231	0.951
16q	702	0.1	0.262	0.951	0.04	-0.957	0.951
17p	683	0.07	-0.231	0.951	0.14	0.984	0.398
17q	1592	0.07	-0.233	0.951	0.14	0.981	0.398
18p	143	0.04	-0.781	0.951	0.23	2.86	0.0103
18q	446	0.04	-0.872	0.951	0.17	1.56	0.194
19p	995	0.05	-0.686	0.951	0.3	4.16	0.00021
19q	1709	0.2	0.901	0.712	0.87	15.2	0
20p	355	0.06	-0.436	0.951	0	-1.66	0.951
20q	753	0.1	0.202	0.951	0	-1.63	0.951
21q	509	0.06	-0.436	0.951	0	-1.66	0.951
22q	921	0.04	-0.957	0.951	0.1	0.261	0.82

Table 7.1 Significantly recurrent broad copy number changes identified by GISTIC2.0 analysis

To identify fusion transcripts, we analysed RNAseq data, which was available for 36 of the 51 tumours. After filtering, the only chimeric transcript identified was the predicted driver *FGFR3-TACC3* fusion, previously described in IDH wild type gliomas [224][225][226], which was seen in 2 of the IDHwt-non-1p/19q co-deleted tumours - Patients 2463 and 2441; Of note was that Patient 2463 carried an *IDH2* intron 5 mutation (c.679-28C>T).

Incorporation of TCGA mutation data

To explore the mutational spectra of AO in an independent series, we made use of data generated by The Cancer Genome Atlas (TCGA) study of low-grade glioma, which provides exome sequencing data on a further 43 AO tumours.Two of the analysed 43 tumours harboured frameshift mutations in *TCF12* (E548R and D171fs) (Online resource: Supplementary Data 2). As with our series, these *TCF12* mutations were exclusive to IDH-1p/19q co-deleted tumours. In a combined analysis, mutations in *PI3KCA, NOTCH1* and *TP53* were significantly overrepresented when analyzed using MutSigCV (*Q* value <0.1; online resource: Supplementary Data 2). Additionally mutation of *ATRX* and *RBPJ* were of borderline significance.

A bias towards variants with functional impact (FM) is a feature of cancer drivers [167] To increase our ability to identify cancer drivers and delineate associated oncogenic pathways for AO, we incorporated mutation data from multiple tumour types using Oncodrive-fm [167] implemented within the IntOGen-mutations platform [168] (Figure 5.5). The most recurrently mutated genes according to MutSig were also detected by Oncodrive-fm as significantly mutated (Q-value<0.05). Oncodrive-fm also identified a number of other important mutated genes (that is, displaying high FM bias) including *SETD2*, *NOTCH2*, *RBPJ*, *ARID1A*, *ARID1B*, *HDAC2* and *SMARCA4* (Figure 5.5).

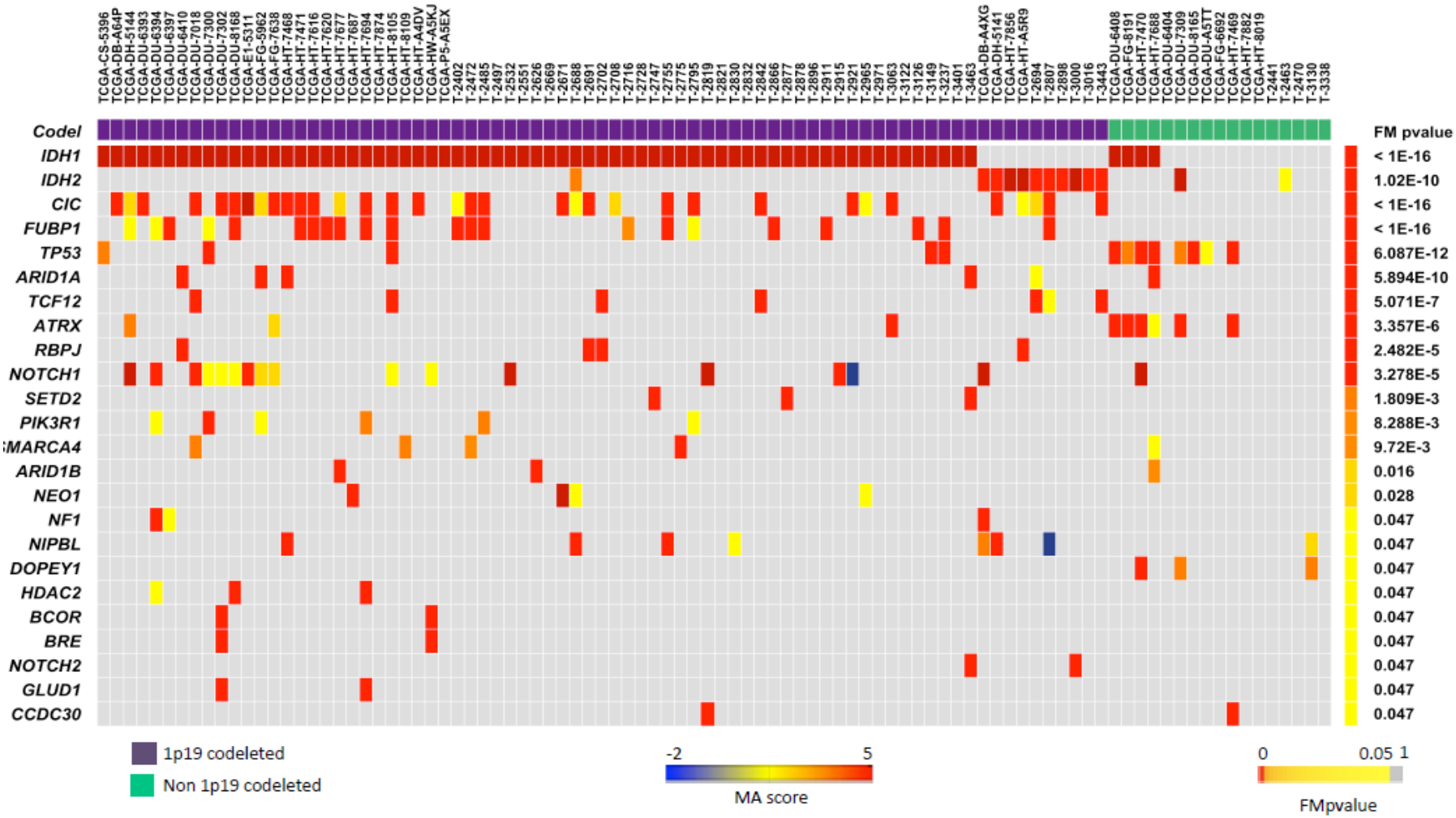


Figure 7.5 FM-biased genes and gene modules in AO identified by Oncodrive-fm using data from this study and tumours profiled by TCGA. Heatmap shows tumours in columns and genes in rows, the colour reflecting the MutationAssessor (MA) scores of somatic mutations. FM ext. qv, corrected P values of the FM bias analysis using the external null distribution.

Using all mutation results, we performed an analysis to identify pathways or gene ontologies that were significantly enriched in mutated genes. As expected the most significantly altered pathways were linked to TCA cycle and isocitrate metabolic process as a consequence of IDH mutation. Consistent with the other genes that were found significantly mutated by MutSigCV and Oncodrive-fm analysis, Notch-signaling pathway (P=1.0x10^{-5}, Binomial test), genes involved in neuron differentiation (P=2.0x10^{-5}, Binomial test), and genes involved in chromatin organization (P=0.02, Binomial test) were also significantly enriched for mutations (Table 5.2).

a)

Rank	*Gene Set*	*p-value*
43	***ONDER_CDH1_TARGETS_2_UP***	***1,23E-08***
138	***CUI_TCF21_TARGETS_2_DN***	***2,61E-06***
443	***NUYTTEN_EZH2_TARGETS_UP***	***3,57E-04***
1418	***CUI_TCF21_TARGETS_2_UP***	***1,13E-02***
1594	***WIEDERSCHAIN_TARGETS_OF_BMI1_AND_PCGF2***	***1,57E-02***
2060	***ONDER_CDH1_TARGETS_1_UP***	***2,91E-02***
2225	***BMI1_DN_MEL18_DN.V1_DN***	***3,38E-02***

b)

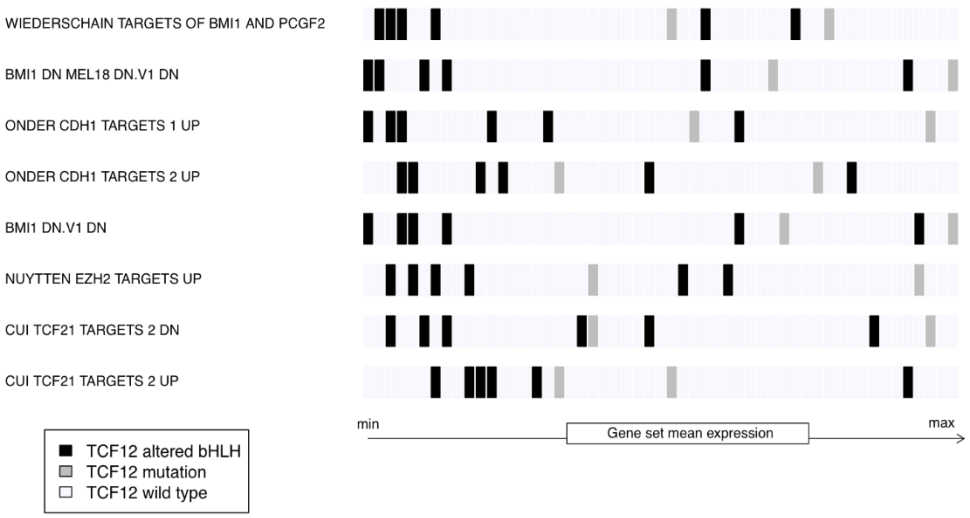


Table 7.2 Downregulation of pathways regulated by TCF12 partners in tumors with altbHLH TCF12m mutants. (a) Target gene sets of CDH1, TCF21, EZH2 and BMI1 are significantly enriched in differentially expressed genes between TCF12 bHLH altered samples and TCF12 wild type tumors. Gene set ranks refer to the p-value ranks among the 19591 gene sets that were tested. CDH1, TCF21, EZH2 and BMI1 target gene set members were retrieved from MSigDB (see ref and methods). (b) Visualisation of samples ranked according to their value of mean gene expression for each gene set. Each row corresponds to the gene set listed on the left, and each rectangle corresponds to a tumour with a color indicating its TCF12 status (wt, altbHLH mutant, or other mutations). Samples with the lowest global expression of all the target genes (whether or not they were initially found differentially expressed in TCF12 bHLH altered samples) are on the left hand side. Reciprocally, samples with the highest global expression of all the target genes are on the right hand side.

Gene set	Source	Number of genes in the gene set	Number of mutated genes	Number of mutations in the gene set	Main genes mutated (number of mutations)	Binomial test p-value	BH adjusted q-value
MSIGDB.C2.CP___KEGG_CITRATE_CYCLE_TCA_CYCLE	KEGG	32	2	49	*IDH1 (n=42); IDH2 (n=7) ;*	0	0
GO:0048699=generation of neurons	GO	1221	90	219	*AATK (n=2); ATG7 (n=2); COL6A1 (n=5); DIAPH1 (n=4); KIDINS220 (n=3); NEO1 (n=3); NOTCH1 (n=6); NOTCH3 (n=3); PSD3; TCF12 (n=5)*	8.90E-06	0.00035499
MSIGDB.C2.CP___KEGG_NOTCH_SIGNALING_PATHWAY	KEGG	47	7	20	*CREBBP (n=3);MAML2 (n=2);NCOR2 (n=2); NOTCH1 (n=6); NOTCH2 (n=2); NOTCH3 (n=3); RBPJ (n=2)*	1.04E-05	0.00040244
GO:0022008=neurogenesis	GO	1296	95	230	*AATK (n=2); ATG7 (n=2); COL6A1 (n=5); DIAPH1 (n=4); KIDINS220 (n=3); NEO1 (n=3); NOTCH1 (n=6); NOTCH3 (n=3); PSD3; TCF12 (n=5)*	1.12E-05	0.00042855
GO:0030182=neuron differentiation	GO	1126	82	202	*AATK (n=2); ATG7 (n=2); COL6A1 (n=5); DIAPH1 (n=4;KIDINS220 (n=3); NEO1 (n=3);NOTCH1 (n=6);NOTCH3 (n=3);PSD3; TCF12 (n=5)*	2.39E-05	0.00080623
GO:0006325=chromatin organization	GO	618	35	85	*ARID1A (n=2);ATRX (n=4);ATXN7 (n=2);BRCA2 (n=2); CREBBP (n=3); NIPBL (n=5); SETD2 (n=3); SMARCA4 (n=2); TRRAP (n=4)*	0.02110324	0.17069518
GO:0016568=chromatin modification	GO	495	31	77	*ARID1A (n=2);ATRX (n=4);ATXN7 (n=2);BRCA2 (n=2); CREBBP (n=3); NIPBL (n=5); SETD2 (n=3); SMARCA4 (n=2); TRRAP (n=4)*	0.02728121	0.20473476

Table 7.3 Significantly mutated gene sets.(a) Gene sets harbouring significantly more mutations than expected by chance are indicated. (b) Gene sets highlighted in the study, with detailed number of mutations among the main mutated genes of each set (identified as significantly mutated through MutSigCV or as significantly biased through Oncodrive-fm).

Validation of *TCF12* in an additional series of AO

To identify additional *TCF12* mutated AO tumours we conducted targeted sequencing of a further 83 AOs. Five tumours harboured *TCF12* mutations - G48fs*38, M260fs*5, R326S, D455fs*59 and delN606 (Online resource: Supplementary Data 1). Based on our combined sample of 134 tumours the mutation frequency of *TCF12* in AO is 7.5% (95% confidence interval 3.6-13.2%). No significant difference in patient survival in 1p/19q co-deleted AOs was associated with *TCF12* mutation in 69 patients (Figure 5.6). While our power to demonstrate a statistically significant relationship was limited (*i.e.* ~40% for a hazard ratio of 2.0, stipulating *P*=0.05) we noted that patients having either *TCF12* mutated or *TCF12* LOH tended to be associated with shorter survival (Figure 5.6). To gain further insight into the role of *TCF12* mutation in oligodendroglioma we sequenced 75 grade II tumours identifying one mutation carrier (P212fs*31; Online resource: Supplementary Data 1). The observation that the frequency of *TCF12* mutations is higher in AO as compared with grade II tumours (*P*=0.049, Chi squared test) is compatible with *TCF12* participating in the generation of a more aggressive phenotype.

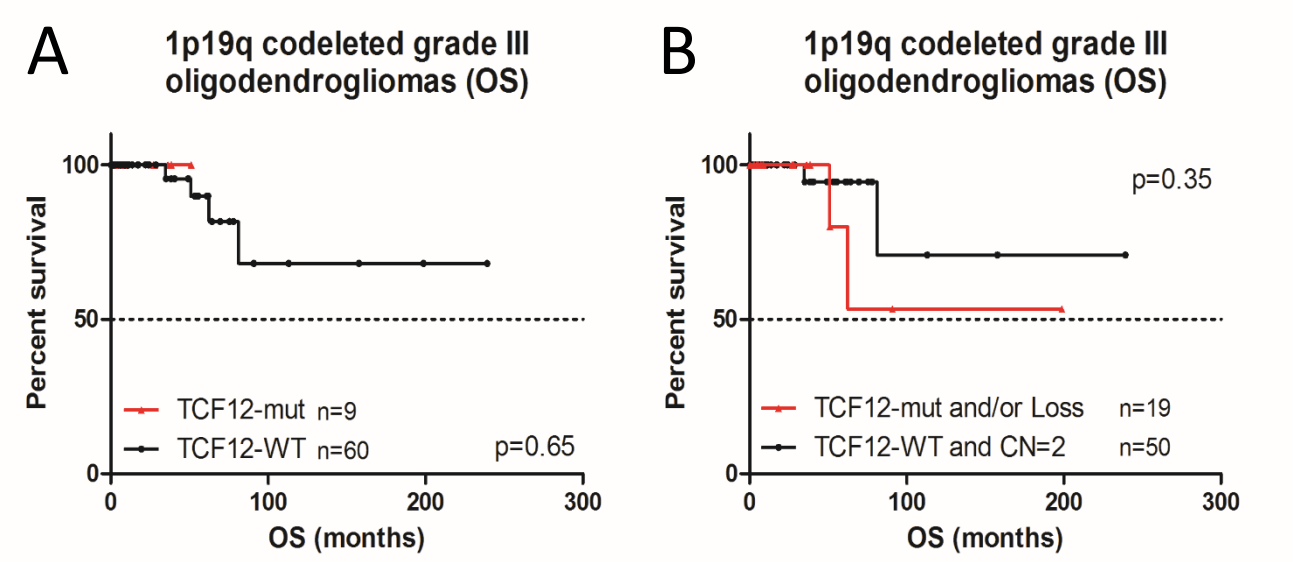


Figure 7.6 Overall survival from of 1p/19q co-deleted anaplastic oligodendrogliomas according to TCF12 mutation status. Overall survival analysis of (a) TCF12 mutant (red line) and TCF12 wild-type glioma patients (black line), (b) TCF12 mutant ± TCF12 loss of heterozygosity (LOH; red line) and TCF12 wild-type patients without any copy number change (black line). The median follow-up was 35 months. Log-rank (Mantel-Cox) test was used to evaluate the significance of differences.

TCF12 bHLH mutants compromised transactivation

To explore the functional consequences of *TCF12* mutation, we tested the transcriptional activity of several mutants (Figure 5.7). We tested the frameshift mutations M260fs*5 and E548fs*13, which in the germline cause coronal craniosynostosis [223] and S682fs*14, since introduction of a C-terminal premature stop codon may result in escape from non-sense mediated decay. We also tested the missense mutation R602M, which is predicted to destabilize the bHLH domain required for DNA binding and dimerization (Figure 5.7) and whose adjacent residue (R603) has been found recurrently mutated in colon cancer [227]. Finally, we tested the missense mutation R326S, since mutations of adjacent G327 have been reported in lung adenocarcinoma [228]. The frame-shift mutants M260fs*5 and E548fs*13 completely abolished TCF12 transactivation consistent with the lack of bHLH DNA binding domain (Figure 5.7). R602M retained only 34% of WT transcriptional activity (*P*=0.0018, Student's t-test; Figure 5.7). We did not observe significant modulation of transactivation for the R326S and S682fs*14 mutants although the latter consistently showed decreased activity (Figure 5.7).

Down-regulation of pathways in *TCF12* bHLH mutants

We profiled gene expression in 8 *TCF12* mutated and 45 wild-type tumours within 1p/19q co-deleted samples (Table 5.1). *TCF12* mutation was associated with significant enrichment of immune response pathways (Table 5.3). Restricting the analysis to tumours with *TCF12* altered bHLH domain (n=6), we found down regulation of pathways featuring known partners of TCF12, such as TCF21, EZH2 and BMI1 [229] (Table 5.2). Interestingly, we found decreased activity of genes sets related to E-cadherin (*CDH1*), which is a TCF12 target gene associated with tumour phenotype [229]. Since the promoter sequences of *CDH1* and *BMI1* feature E-box motifs and are modulated by the bHLH binding [230][231], this provides a mechanistic basis for change in gene expression associated with mutant *TCF12*.

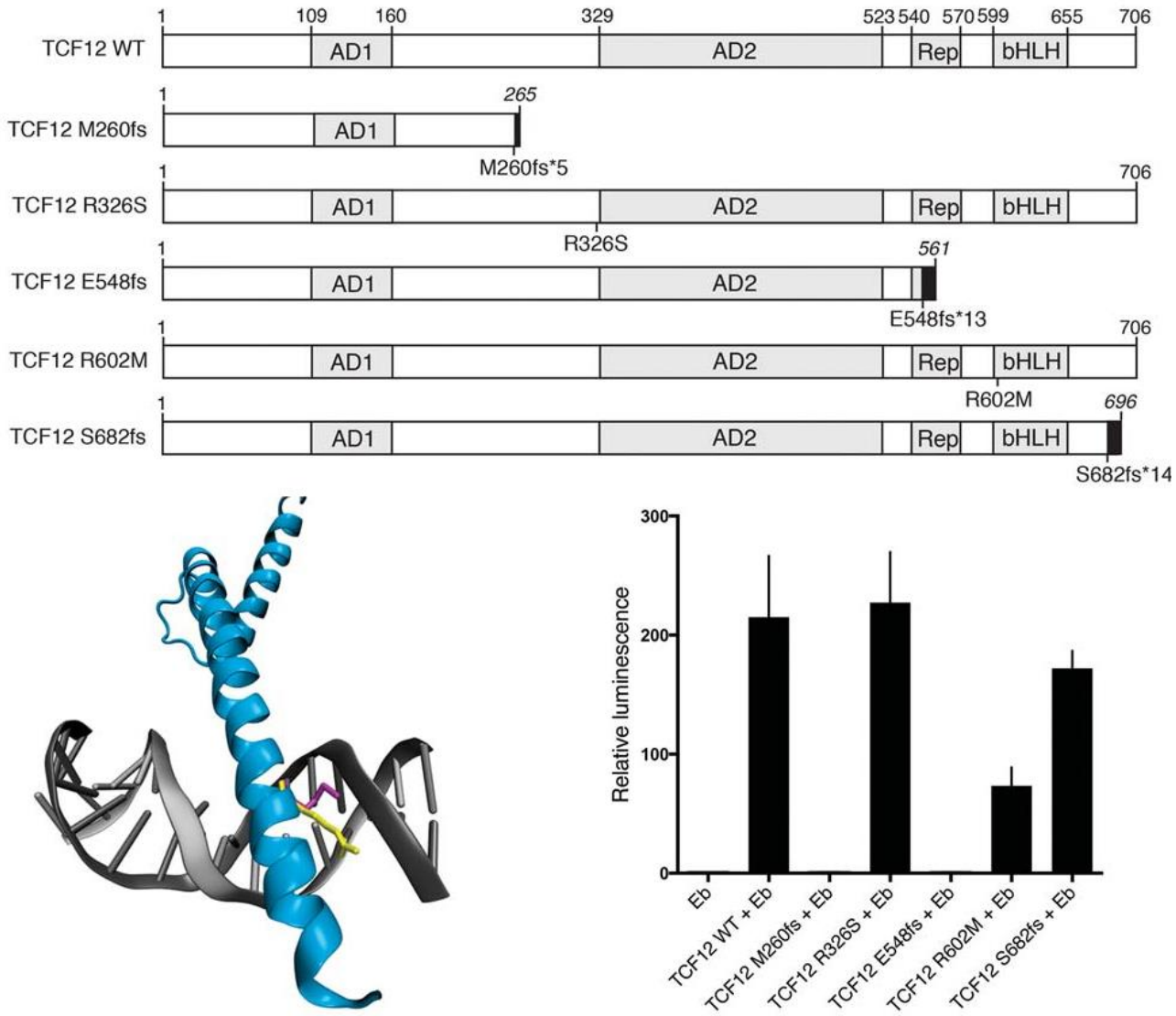


Figure 7.7 TCF12 mutations altering the bHLH domain result in impaired transactivation. (a) Schematic view of the wild-type and mutant TCF12 proteins for which the transactivation capacity has been assessed. Upper panel: wild-type human TCF12, functional domains in grey—activation domain 1 (AD1), activation domain 2 (AD2), repressor domain (Rep) and bHLH domain (bHLH). Lower panel: resulting truncated proteins. Black boxes indicate non-related amino-acid sequences resulting from frameshift mutations (fs), and truncated proteins size is in italic. (b) Schematic structure of the bHLH domain of TCF12 (blue) bound to DNA (grey). WT R602 (yellow) and mutant M602 (purple) residues are indicated. (c) E-box-luciferase reporter plasmid (Eb) was transfected alone or in combination with TCF12 wild-type or mutant expression plasmids. Both frameshift mutants that lack the bHLH DNA binding domain completely abolish TCF12 transcriptional activity. All samples were run in triplicate in four independent experiments. Data were normalized to control renilla luciferase. Values are mean±s.d. ***P=0.0002, **P=0.0018 (Student's t-test).

Mutant TCF12 proteins show subcellular localization changes

We evaluated TCF12 expression and subcellular localization for all of our 11 *TCF12*-mutated tumours (10 AO and 1 Oligodendroglioma grade II) and 11 *TCF12* wild-type tumours by immunohistochemistry. All *TCF12* wild-type tumours showed nuclear expression in a heterogeneous cell population (Figure 5.8; Figure 5.9), whereas *TCF12* mutated tumours showed nuclear and cytoplasmic staining (Figure 5.8). Interestingly, mutations abolishing transcriptional activity were associated with increased staining, suggesting mutant protein accumulation.

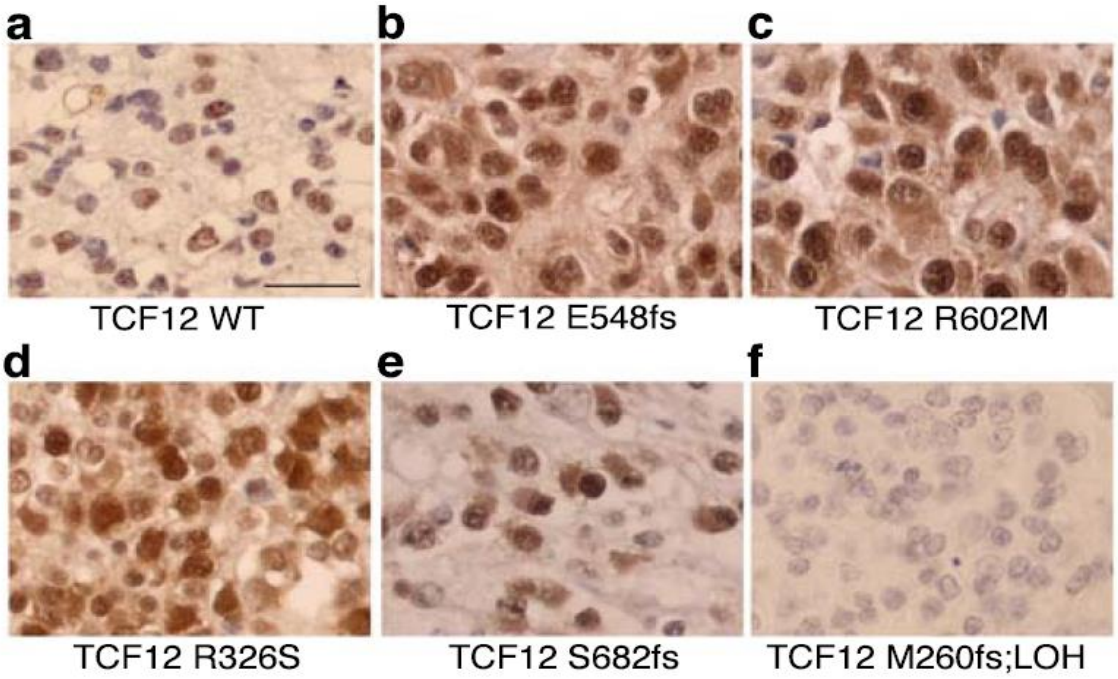


Figure 7.8 TCF12 is highly expressed in a subset of anaplastic oligodendroglioma. Representative TCF12 immunostainings are shown: (a) wild-type TCF12 tumours show nuclear staining in a heterogeneous cell population. (b–e) Mutant TCF12 tumours show strong nuclear and cytoplasmic staining. (f) Mutant M260fs (resulting in a truncated protein) is associated with 15q21.3 LOH and shows no staining. Scale bar, 50 µm.

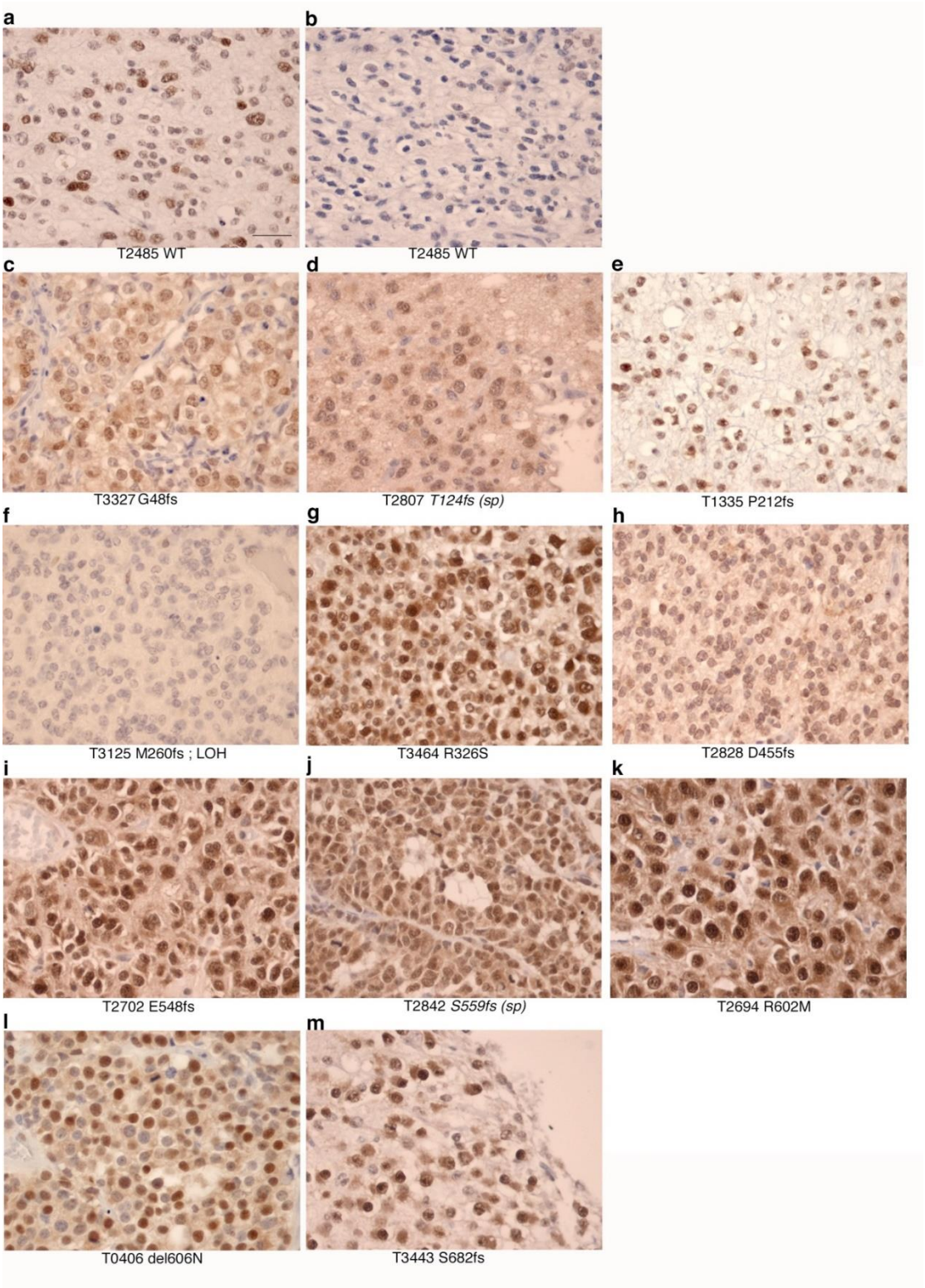


Figure.7.9 TCF12 protein expression in anaplastic oligodendroglioma. (a-m) TCF12 immunostaining on paraffin sections of 1p/19q co-deleted AO. (a) Representative IHC of wild type TCF12 tumor shows nuclear staining in a heterogeneous cell population, the scale bar corresponds to 5u (b) TCF12 negative field from the same tumor, (c-e,h) N-terminal heterozygous frame shift (fs) mutants show reduced positive staining, corresponding only to the residual wild type allele, (f) N-terminal frame shift mutant M260fs with loss of heterozygosity at 15q21.3 stains negative, (g,i-k,m) C-terminal TCF12 mutants show a characteristic strong nuclear and cytoplasmic staining. The in-frame deletion in (l) showing only nuclear staining is the exception

TCF12 mutations associate with aggressive tumour phenotype

We profiled the extent of necrosis, microvascular proliferation and the mitotic index available for *TCF12* wild type or mutated tumours. A significant increase in palisading necrosis (Figure 5.10) as well as a trend towards a higher mitotic index was associated with *TCF12* mutation, consistent with a more aggressive phenotype (Figure 5.10). Intriguingly, tumours harboring disruptive bHLH domain mutations exhibited the highest proportion of palisading necrosis and mitotic figures.

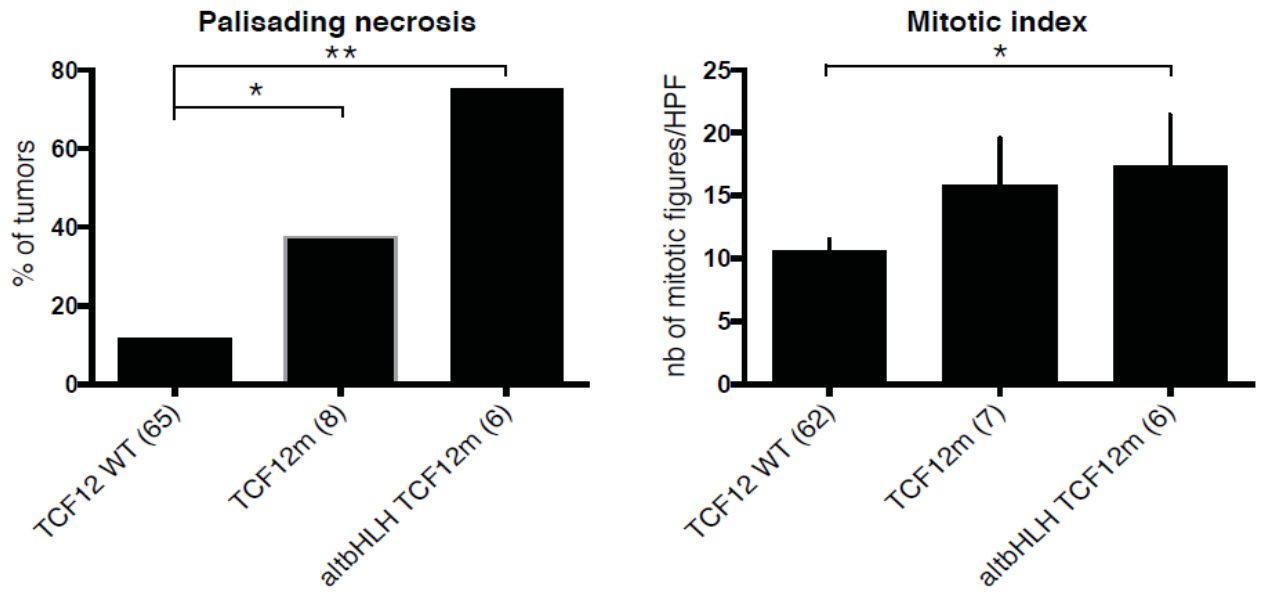


Figure 7.10 TCF12 mutation correlates with a higher necrotic and mitotic index. (a) Percentage of palisading necrosis in tumours with wild-type TCF12, all tumours mutated for TCF12 or only altered bHLH TCF12 mutants; **P*=0.02, ***P*=0.004. (b) Mitotic index in TCF12 wild-type, TCF12-mutated and altered bHLH TCF12 mutants; **P*=0.039, mean±s.e.m. CN, copy number; LOH, loss of heterozygosity; HPF, high-power field. The number of samples is indicated in parenthesis.

7.4 .Discussion

These whole exome sequencing of AO has confirmed the mutually exclusive mutational profile in IDHmut-1p/19q co-deleted and IDHmut non-1p/19q co-deleted tumour subtypes, which reflect distinct molecular mechanisms of oncogenesis - consistent with the requirement for either 1p/19q co-deletion or TP53 mutation post IDH-mutation. Moreover, as previously proposed, the genomic abnormalities in IDHmut- 1p/19p co-deleted tumours are consistent with one common mechanism of tumour initiation being through 1p/19q loss, mutation of IDH1 or IDH2, and TERT activation through promoter mutation [27], which in turn predisposes to deactivation of CIC, FUBP1, NOTCH and activating mutations/amplifications in the PI3K-pathway.

I identified and replicated mutations in TCF12, a bHLH transcription factor that mediates transcription by forming homo- or heterodimers with other bHLH transcription factors. Tcf12 is highly expressed in neural progenitor cells during neural development [232] and in cells of the oligodendrocyte lineage [233] .

We found that mutations generating truncated TCF12 lacking the bHLH DNA binding domain abrogate the transcriptional activity of TCF12. In addition, single residue substitutions such as R602M within the bHLH domain also dramatically reduce TCF12 transcriptional ability. Finally, we found that the loss of TCF12 transcriptional activity was associated with a more aggressive tumour phenotype. Although speculative, our expression data provides evidence that the effects of TCF12 mutation on AO development may be mediated in part through E-cadherin related pathway. Indeed, this was one of the pathways down-regulated in mutated tumours and intriguingly CDH1 has been implicated in metastatic behavior in a number of cancers [229][234]. It is likely that some TCF12 mutations may have subtle effects on bHLH function or act through independent pathways. Irrespective of the downstream effects of TCF12 mutation on glioma our data are compatible with TCF12 having haplo-insufficient tumour suppressor function. TCF12 haploinsufficiency has previously been reported in patients with coronal craniosynostosis and in their unaffected relatives [223]. Strikingly, 3 of the 11 mutations we identified in AO, that concern residues M260, E548 and R602 cause coronal craniosynostosis [223][235]. Although speculative collectively these data raise the possibility that carriers of germline TCF12 mutations may be at an increased risk of developing AO.

CHAPTER 6

General discussion, future work and conclusion

8.1 .Glioma inherited predisposition

A major focus of this thesis has been on glioma germline genetic susceptibility, and the results of this work can be summarised as follows. The identification of thirteen new risk loci for glioma in chapter 3 provides additional evidence that genetic susceptibility to glioma is polygenic. This study, which is the largest glioma GWAS to date, provides strong evidence for specific associations between risk SNPs and different histological glioma subtypes, presumably resulting from different etiological pathways. In the combined meta-analysis, among previously published glioma risk SNPs, those for all glioma at 17p13.1 (*TP53*), for GBM at 5p15.33 (*TERT*), 7p11.2 (*EGFR*), 9p21.3 (*CDKN2B–AS1*) and 20q13.33 (*RTEL1*), and for non-GBM tumours at 8q24.21 (*CCDC26*), 11q23.2, 11q23.3 (*PHLDB1*) and 15q24.2 (*ETFA*) showed even greater evidence for association. SNPs at 10q25.2 and 12q12.1 for non-GBM tumours retained genome-wide significance. Associations at the previously reported 3q26.2 (near *TERC*) [90] and 12q23.33 (*POLR3B*) [92] loci for GBM did not retain statistical significance. In addition to previously reported loci, we identified genome-wide significant associations marking new risk loci for GBM at 1p31.3 (*RAVER2*), 11q14.1, 16p13.3 (near *MPG*), 16q12.1 (*HEART3*) and 22q13.1 (SLC16A8) and for non-GBM tumours at 1q32.1 (*MDM4*), 1q44 (*AKT3*), 2q33.3 (near *IDH1*), 3p14.1 (*LRIG1*), 10q24.33 (*OBCF1*), 11q21 (*MAML2*), 14q12 (*AKAP6*) and 16p13.3 (*LMF1*).

As demonstrated in Chapter 3, meta-analysis of GWAS studies with genotype imputation using a UK10K and 1000 genomes project reference panel is a robust method for investigating low-penetrance genetic susceptibility to glioma. However, the 25 identified risk SNPs for glioma account for, at best, ∼27% and ∼37% of the familial risk of GBM and non-GBM tumours, respectively. Therefore, further GWAS-based analyses should lead to additional insights into the biology and etiological basis of the different glioma histologies. Notably, such information can inform gene discovery initiatives and thus have a measurable effect on the successful development of new therapeutic agents. Regarding future studies of glioma germline genetics, an important step forward is the continued large collaborative efforts such as the Glioma International Case Control (GICC) consortium to increase detection power for common alleles.

In the course of this thesis the WHO 2016 CNS classification has emerged to provide better definition and more precise categorisation of distinct brain tumours. This new classification integrates molecular markers with histology, consistent with results in chapter 4 where a majority of risk loci show evidence of molecular subtype specificity notably for 5p15.33, 9p21.3, 17p13.1 and 20q13.33 with *TERT* promoter mutated only glioma as well as 8q24.21 for glioma with IDH mutation, *TERT* promoter mutation and 1p/19q co-deletion.

This analysis was based on defining glioma subgroups using only three primary markers. Integration of additional genetic markers to molecular sub-grouping of glioma resulting from ongoing large-scale tumour sequencing projects is likely to provide for further insights into glial oncogenesis and ultimately may suggest targets for novel therapeutic strategies.

The functional basis of most GWAS risk loci is through regulatory effects, and results in chapter 3 and 4 demonstrate that the use of publicly available eQTL, chromatin state and Hi-C data to identify candidate regulatory elements and target genes. However, such data is limited and it is extremely important to use the most appropriate model systems to investigate these loci. Future studies therefore will benefit from more extensive reference data, for example to enable exploration of chromatin architecture differences between IDH mutated and wild-type gliomas, as well as at different stages of gliogenesis,

8.2 .Somatic genetic studies of Anaplastic Oligodentroglioma OA

To our knowledge the study in chapter 5 represents the largest sequencing study of AO conducted to date. *TCF12* was shown to be a driver gene with mutations compromising *TCF12* transcriptional activity and resulting in a more aggressive tumour type. However, given the number of tumour-normal pairs we have analysed and the mutational frequency in AO, we were only well powered to identify genes which have a high frequency of mutations (i.e. >10%). Hence further insights into the biology of AO should be forthcoming through additional sequencing initiatives and meta-analyses of these data.

8.3 .Overall conclusion

Understanding the molecular basis of glioma predisposition is likely to derive insight into tumour biology and potentially identify novel targets or pathways for therapeutic intervention. While progress in this area has been limited so far, initiatives providing integrated molecular and clinico-pathological data on glioma are likely to accelerate advances. The collective findings from this thesis suggest future efforts in genetic predisposition to glioma are likely to involve further GWAS as well as functional studies to identify the molecular mechanisms by which risk loci influence disease risk.

During my thesis, I also, studied the genetic susceptibility in Primary central nervous system lymphoma (PCNSL) which is a rare form of Hodgkin lymphoma. I performed a meta-analysis of two new genome-wide association studies of PCNSL totaling 475 cases and 1,134 controls of European ancestry. These study led to the identification of independent risk loci at 3p22.1 (rs41289586, *ANO10*, $P = 2.17 \times 10^{-8}$) and 6p25.3 near *EXOC2* (rs116446171, $P = 1.95 \times 10^{-13}$). These data provided for the first time, insight into inherited predisposition to PCNSL (Labreche *et al"A genome-wide association study identifies susceptibility loci for primary central nervous system lymphoma at 6p25.3 and 3p22.1: a LOC network study group"* Nature Communication 2018, in review).

In addition, I contributed on work that has led to the identification of a novel recurrent gene fusion ETV6-IgH in PCNSL. Overall, ETV6-IgH was found in 13 out of 72 PCNSL (18%). ETV6 was significantly underexpressed at the gene level. ETV6-IgH is a new potential surrogate marker of PCNSL with favorable prognosis, with ETV6 haploinsuffiency as a possible mechanism.

The great opportunity provided by working in two prestigious multidisciplinary research institutes, has offered me the chance to be implicated in diverse studies both in terms of tumours type as well as the technologies and analytical methods I employed. The peer reviewed publications I contributed to during my thesis are listed in the APPENDIX 2.

References

[1] C. Rouse, H. Gittleman, Q. T. Ostrom, C. Kruchko, and J. S. Barnholtz-Sloan, "Years of potential life lost for brain and CNS tumors relative to other cancers in adults in the United States, 2010," *Neuro. Oncol.*, vol. 18, no. 1, pp. 70–77, 2016.

[2] Q. T. Ostrom *et al*, "CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014," *Neuro. Oncol.*, vol. 19, no. suppl_5, pp. v1–v88, 2017.

[3] J. Ferlay, D. M. Parkin, and E. Steliarova-Foucher, "Estimates of cancer incidence and mortality in Europe in 2008," *Eur. J. Cancer*, vol. 46, no. 4, pp. 765–781, 2010.

[4] M. R. Gilbert *et al*, "A Randomized Trial of Bevacizumab for Newly Diagnosed Glioblastoma," *N. Engl. J. Med.*, vol. 370, no. 8, pp. 699–708, 2014.

[5] O. L. Chinot *et al*, "Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma.," *N Engl J Med*, vol. 370, no. 8, pp. 709–22, 2014.

[6] M. R. Gilbert *et al*, "Dose-dense temozolomide for newly diagnosed glioblastoma: a randomized phase III clinical trial.," *J. Clin. Oncol.*, vol. 31, no. 32, pp. 4085–4091, 2013.

[7] D. Gramatzki *et al*, "Glioblastoma in the Canton of Zurich, Switzerland revisited: 2005 to 2009," *Cancer*, vol. 122, no. 14, pp. 2206–2215, 2016.

[8] M. Weller *et al*, "Glioma," *Nature Reviews Disease Primers*, vol. 1. 2015.

[9] J. T. Huse and E. C. Holland, "Targeting brain cancer: Advances in the molecular pathology of malignant glioma and medulloblastoma," *Nature Reviews Cancer*, vol. 10, no. 5. pp. 319–331, 2010.

[10] D. N. Louis *et al*, "The 2007 WHO classification of tumours of the central nervous system," *Acta Neuropathologica*, vol. 114, no. 2. pp. 97–109, 2007.

[11] H. Ohgaki and P. Kleihues, "The definition of primary and secondary glioblastoma," *Clinical Cancer Research*. 2013.

[12] A. Darlix *et al*, "Epidemiology for primary brain tumors: a nationwide population-based study," *Journal of Neuro-Oncology*, pp. 1–22, 2016.

[13] H. Nakamura, K. Makino, S. Yano, and J. I. Kuratsu, "Epidemiological study of primary intracranial tumors: A regional survey in Kumamoto prefecture in southern Japan-20-year study," *Int. J. Clin. Oncol.*, vol. 16, no. 4, pp. 314–321, 2011.

[14] M. Kallio, "The incidence of intracranial gliomas in southern Finland," *Acta Neurol. Scand.*, vol. 78, no. 6, pp. 480–483, 1988.

[15] M. Sant *et al*, "Survival of European patients with central nervous system tumors," *Int J Cancer*, vol. 131, no. 1, pp. 173–185, 2012.

[16] D. N. Louis *et al*, "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary," *Acta Neuropathologica*. 2016.

[17] H. Yan *et al*, "*IDH1* and *IDH2* Mutations in Gliomas," *N. Engl. J. Med.*, 2009.

[18] D. W. Parsons *et al*, "An integrated genomic analysis of human glioblastoma multiforme," *Science (80-.).*, 2008.

[19] H. Suzuki *et al*, "Mutational landscape and clonal architecture in grade II and III gliomas," *Nat. Genet.*, 2015.

[20] C. G. A. R. Network *et al*, "Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas.," *N. Engl. J. Med.*, 2015.

[21] T. Watanabe, S. Nobusawa, P. Kleihues, and H. Ohgaki, "IDH1 mutations are early events in the development of astrocytomas and oligodendrogliomas," *Am. J. Pathol.*, 2009.

[22] M. Sasaki *et al*, "IDH1(R132H) mutation increases murine haematopoietic progenitors and alters epigenetics," *Nature*, 2012.

[23] J. A. Losman and W. G. Kaelin, "What a difference a hydroxyl makes: Mutant IDH, (R)-2-hydroxyglutarate, and cancer," *Genes and Development*. 2013.

[24] H. Noushmehr *et al*, "Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma," *Cancer Cell*, 2010.

[25] R. B. Jenkins *et al*, "A t(1;19)(q10;p10) mediates the combined deletions of 1p and 19q and predicts a better prognosis of patients with oligodendroglioma," *Cancer Res.*, 2006.

[26] M. Labussière *et al*, "TERT promoter mutations in gliomas, genetic associations and clinico-pathological correlations," *Br. J. Cancer*, 2014.

[27] P. J. Killela *et al*, "TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal," *Proc. Natl. Acad. Sci.*, 2013.

[28] C. Bettegowda *et al*, "Mutations in CIC and FUBP1 contribute to human oligodendroglioma," *Science (80-.).*, 2011.

[29] S. et al A. N. (2015) 129: 829. doi:10. 1007/s0040.-015-1432-1 Aldape, K., Zadeh, G., Mansouri, "Glioblastoma : pathology , molecular mechanisms and markers," *Acta Neuropathol.*, 2015.

[30] K. D. Aldape *et al*, "Immunohistochemical detection of EGFRvIII in high malignancy grade astrocytomas and evaluation of prognostic significance," *J. Neuropathol. Exp. Neurol.*, 2004.

[31] et al McLendon, R, Friedman, A, Bigner, D, Van Meir, EG, Brat, DJ, Mastrogianakis, GM, Olson, JJ, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, 2008.

[32] J. E. Eckel-Passow *et al*, "Glioma Groups Based on 1p/19q, *IDH* , and *TERT* Promoter Mutations in Tumors," *N. Engl. J. Med.*, 2015.

[33] R. G. W. Verhaak *et al*, "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, no. 1, pp. 98–110, 2010.

[34] T. M. Kim, W. Huang, R. Park, P. J. Park, and M. D. Johnson, "A developmental taxonomy of glioblastoma defined and maintained by microRNAs," *Cancer Res.*, vol. 71, no. 9, pp. 3387–3399, 2011.

[35] W. Wick *et al*, "Temozolomide chemotherapy alone versus radiotherapy alone for malignant astrocytoma in the elderly: The NOA-08 randomised, phase 3 trial," *Lancet Oncol.*, vol. 13, no. 7, pp. 707–715, 2012.

[36] A. Malmström *et al*, "Temozolomide versus standard 6-week radiotherapy versus hypofractionated radiotherapy in patients older than 60 years with glioblastoma: The Nordic randomised, phase 3 trial,"

Lancet Oncol., vol. 13, no. 9, pp. 916–926, 2012.

[37] R. Stupp *et al*, "Radiotherapy plus Concomitant\nand Adjuvant Temozolomide for Glioblastoma," *N. Engl. J. Med.*, 2005.

[38] M. J. Van Den Bent *et al*, "Adjuvant procarbazine, lomustine, and vincristine chemotherapy in newly diagnosed anaplastic oligodendroglioma: Long-term follow-up of EORTC brain tumor group study 26951," *J. Clin. Oncol.*, 2013.

[39] M. J. van den Bent *et al*, "Interim results from the CATNON trial (EORTC study 26053-22054) of treatment with concurrent and adjuvant temozolomide for 1p/19q non-co-deleted anaplastic glioma: a phase 3, randomised, open-label intergroup study," *Lancet*, 2017.

[40] J. G. Cairncross *et al*, "Chemotherapy plus radiotherapy (CT-RT) versus RT alone for patients with anaplastic oligodendroglioma: Long-term results of the RTOG 9402 phase III study," *J. Clin. Oncol. Conf.*, 2012.

[41] G. Cairncross *et al*, "Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: Long-term results of RTOG 9402," *J. Clin. Oncol.*, 2013.

[42] H. Duffau and L. Taillandier, "New concepts in the management of diffuse low-grade glioma: Proposal of a multistage and individualized therapeutic approach," *Neuro. Oncol.*, 2015.

[43] O. L. Chinot *et al*, "Bevacizumab plus Radiotherapy–Temozolomide for Newly Diagnosed Glioblastoma," *N. Engl. J. Med.*, 2014.

[44] W. Wick *et al*, "Lomustine and Bevacizumab in Progressive Glioblastoma," *N. Engl. J. Med.*, 2017.

[45] T. E. Taylor, F. B. Furnari, and W. K. Cavenee, "Targeting EGFR for Treatment of Glioblastoma: Molecular Basis to Overcome Resistance," *Curr. Cancer Drug Targets*, vol. 12, no. 3, pp. 197–209, 2012.

[46] T. Schumacher *et al*, "A vaccine targeting mutant IDH1 induces antitumour immunity," *Nature*, vol. 512, no. 7514, pp. 324–327, 2014.

[47] A. DE, "Genetic study of breast cancer: identification of a high risk group," *Cancer*, vol. 34, no. 4, pp. 1090–7, 1974.

[48] Y. Miki *et al*, "Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1," *Science (80-.).*, vol. 266, pp. 66–71, 1994.

[49] C. J. Hussussian *et al*, "Germline p16 mutations in familial melanoma," *Nat. Genet.*, vol. 8, no. 1, pp. 15–21, 1994.

[50] R. Wooster *et al*, "Identification of the breast cancer susceptibility gene BRCA2.," *Nature*, vol. 378, no. 6559, pp. 789–92, 1995.

[51] F. S. Leach *et al*, "Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer," *Cell*, vol. 75, no. 6, pp. 1215–1225, 1993.

[52] A. Sud, B. Kinnersley, and R. S. Houlston, "Genome-wide association studies of cancer: Current insights and future perspectives," *Nature Reviews Cancer*. 2017.

[53] F. P. Li *et al*, "A Cancer Family Syndrome in Twenty-four Kindreds," *Cancer Res.*, vol. 48, no. 18, pp. 5358–5362, 1988.

[54] A. Iavarone, K. K. Matthay, T. M. Steinkirchner, and M. A. Israel, "Germ-line and somatic p53 gene

mutations in multifocal osteogenic sarcoma," *Proc Natl Acad Sci U S A*, vol. 89, no. 9, pp. 4207–4209, 1992.

[55] D. Malkin *et al*, "Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms," *Science (80-.).*, vol. 250, no. 4985, pp. 1233–1238, 1990.

[56] D. Malkin *et al*, "Germline mutations of the p53 tumor-suppressor gene in children and young adults with second malignant neoplasms.," *N. Engl. J. Med.*, vol. 326, no. 20, pp. 1309–15, 1992.

[57] D. Sidransky *et al*, "Inherited p53 gene mutations in breast cancer," *Cancer Res.*, vol. 52, no. 10, pp. 2984–2986, 1992.

[58] A. P. Kyritsis *et al*, "Germline p53 gene mutations in subsets of glioma patients," *J. Natl. Cancer Inst.*, vol. 86, no. 5, pp. 344–349, 1994.

[59] P. Chen, a Iavarone, J. Fick, M. Edwards, M. Prados, and M. a Israel, "Constitutional p53 mutations associated with brain tumors in young adults," *Cancer Genet.*, vol. 82, no. 2, pp. 106–115, 1995.

[60] D. G. R. Evans, "Neurofibromatosis type 2," *J. Med. Genet.*, vol. 37, no. 12, pp. 897–904, 2000.

[61] D. H. Gutmann *et al*, "Molecular analysis of astrocytomas presenting after age 10 in individuals with NF1," *Neurology*, vol. 61, no. 10, pp. 1397–1400, 2003.

[62] J. W. Poley *et al*, "Biallelic germline mutations of mismatch-repair genes: A possible cause for multiple pediatric malignancies," *Cancer*, vol. 109, no. 11, pp. 2349–2356, 2007.

[63] J. A. Randerson-Moor *et al*, "A germline deletion of p14(ARF) but not CDKN2A in a melanoma-neural system tumour syndrome family.," *Hum. Mol. Genet.*, vol. 10, no. 1, pp. 55–62, 2001.

[64] I. Tachibana, J. S. Smith, K. Sato, S. M. Hosek, D. W. Kimmel, and R. B. Jenkins, "Investigation of germline PTEN, p53, p16(INK4A)/p14(ARF), and CDK4 alterations in familial glioma," *Am. J. Med. Genet.*, vol. 92, no. 2, pp. 136–141, 2000.

[65] S. R. Hamilton *et al*, "The Molecular Basis of Turcot's Syndrome," *N. Engl. J. Med.*, vol. 332, no. 13, pp. 839–847, 1995.

[66] S. B. Elmariah, J. Huse, B. Mason, P. LeRoux, and R. A. Lustig, "Multicentric glioblastoma multiforme in a patient with BRCA-1 invasive breast cancer," *Breast J.*, vol. 12, no. 5, pp. 470–474, 2006.

[67] A. G. Knudson, "Mutation and Cancer: Statistical Study of Retinoblastoma," *Proc. Natl. Acad. Sci.*, vol. 68, no. 4, pp. 820–823, 1971.

[68] K. Hemminki, S. Tretli, J. Sundquist, T. B. Johannesen, and C. Granström, "Familial risks in nervous-system tumours: a histology-specific analysis from Sweden and Norway," *Lancet Oncol.*, vol. 10, no. 5, pp. 481–488, 2009.

[69] N. Paunu *et al*, "A novel low-penetrance locus for familial glioma at 15q23-q26.3," *Cancer Res.*, vol. 62, no. 13, pp. 3798–3802, 2002.

[70] S. Shete *et al*, "Genome-wide high-density SNP linkage search for glioma susceptibility loci: Results from the gliogene consortium," *Cancer Res.*, vol. 71, no. 24, pp. 7568–7575, 2011.

[71] A. Jalali *et al*, "Targeted Sequencing in Chromosome 17q Linkage Region Identifies Familial Glioma Candidates in the Gliogene Consortium," *Sci. Rep.*, vol. 5, 2015.

[72] Y. Liu *et al*, "Insight in glioma susceptibility through an analysis of 6p22.3, 12p13.33-12.1, 17q22-23.2

and 18q23 SNP genotypes in familial and non-familial glioma," *Hum. Genet.*, vol. 131, no. 9, pp. 1507–1517, 2012.

[73] N. Risch and K. Merikangas, "The Future of Genetic Studies of Complex Human Diseases," *Science (80-.).*, vol. 273, no. 5281, pp. 1516–1517, 1996.

[74] W. Bodmer and C. Bonilla, "Common and rare variants in multifactorial susceptibility to common diseases," *Nature Genetics*, vol. 40, no. 6. pp. 695–701, 2008.

[75] K. M. Walsh *et al*, "Analysis of 60 Reported Glioma Risk SNPs Replicates Published GWAS Findings but Fails to Replicate Associations From Published Candidate-Gene Studies," *Genet. Epidemiol.*, vol. 37, no. 2, pp. 222–228, 2013.

[76] D. E. Reich and E. S. Lander, "On the allelic spectrum of human disease," *Trends in Genetics*, vol. 17, no. 9. pp. 502–510, 2001.

[77] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease," *Nature Genetics*, vol. 33, no. 3S. pp. 228–237, 2003.

[78] H. RS and P. J, "The future of association studies of common cancers," *Hum Genet*, vol. 112, no. 4, pp. 434–5, 2003.

[79] D. Welter *et al*, "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic Acids Res.*, vol. 42, no. D1, 2014.

[80] B. Howie, J. Marchini, and M. Stephens, "Genotype Imputation with Thousands of Genomes," *G3: Genes|Genomes|Genetics*, vol. 1, no. 6, pp. 457–470, 2011.

[81] J. Huang *et al*, "Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel," *Nat. Commun.*, vol. 6, 2015.

[82] Y. Wang *et al*, "Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer," *Nat. Genet.*, vol. 46, no. 7, pp. 736–741, 2014.

[83] V. Enciso-Mora *et al*, "Low penetrance susceptibility to glioma is caused by the TP53 variant rs78378222," *Br. J. Cancer*, vol. 108, no. 10, pp. 2178–2185, 2013.

[84] J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies," *Nature Reviews Genetics*, vol. 11, no. 7. pp. 499–511, 2010.

[85] S. Shete *et al*, "Genome-wide association study identifies five susceptibility loci for glioma," *Nat. Genet.*, vol. 41, no. 8, pp. 899–904, 2009.

[86] V. Enciso-Mora *et al*, "Deciphering the 8q24.21 association for glioma," *Hum. Mol. Genet.*, vol. 22, no. 11, pp. 2293–2302, 2013.

[87] M. Sanson *et al*, "Chromosome 7p11.2 (EGFR) variation influences glioma risk," *Hum. Mol. Genet.*, vol. 20, no. 14, pp. 2897–2904, 2011.

[88] S. N. Stacey *et al*, "A germline variant in the TP53 polyadenylation signal confers cancer susceptibility," *Nat. Genet.*, vol. 43, no. 11, pp. 1098–1103, 2011.

[89] M. Wrensch *et al*, "Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility," *Nat. Genet.*, vol. 41, no. 8, pp. 905–908, 2009.

[90] K. M. Walsh *et al*, "Variants near TERT and TERC influencing telomere length are associated with high-grade glioma risk," *Nat. Genet.*, vol. 46, no. 7, pp. 731–735, 2014.

[91] R. B. Jenkins *et al*, "A low-frequency variant at 8q24.21 is strongly associated with risk of oligodendroglial tumors and astrocytomas with IDH1 or IDH2 mutation," *Nat. Genet.*, vol. 44, no. 10, pp. 1122–1125, 2012.

[92] B. Kinnersley *et al*, "Genome-wide association study identifies multiple susceptibility loci for glioma," *Nat. Commun.*, vol. 6, no. May, p. 8559, 2015.

[93] E. Cardis *et al*, "The INTERPHONE study: Design, epidemiological methods, and description of the study population," *Eur. J. Epidemiol.*, vol. 22, no. 9, pp. 647–664, 2007.

[94] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: A tool for genome-wide complex trait analysis," *Am. J. Hum. Genet.*, vol. 88, no. 1, pp. 76–82, 2011.

[95] Y. Lu *et al*, "Most common 'sporadic' cancers have a significant germline genetic component," *Hum. Mol. Genet.*, vol. 23, no. 22, pp. 6112–6118, 2014.

[96] B. Kinnersley *et al*, "Quantifying the heritability of glioma using genome-wide complex trait analysis," *Sci. Rep.*, vol. 5, 2015.

[97] O. Fletcher and R. S. Houlston, "Architecture of inherited susceptibility to common cancer," *Nature Reviews Cancer*, vol. 10, no. 5. pp. 353–361, 2010.

[98] M. Li, C. Li, and W. Guan, "Evaluation of coverage variation of SNP chips for genome-wide association studies," *Eur. J. Hum. Genet.*, vol. 16, no. 5, pp. 635–643, 2008.

[99] J. W. Kent, "Rare variants, common markers: Synthetic association and beyond," *Genet. Epidemiol.*, vol. 35, no. SUPPL. 1, 2011.

[100] C. I. Amos *et al*, "The oncoarray consortium: A network for understanding the genetic architecture of common cancers," *Cancer Epidemiol. Biomarkers Prev.*, vol. 26, no. 1, pp. 126–135, 2017.

[101] K. Michailidou *et al*, "Large-scale genotyping identifies 41 new loci associated with breast cancer risk," *Nat. Genet.*, vol. 45, no. 4, pp. 353–361, 2013.

[102] A. Killedar *et al*, "A Common Cancer Risk-Associated Allele in the hTERT Locus Encodes a Dominant Negative Inhibitor of Telomerase," *PLoS Genet.*, vol. 11, no. 6, 2015.

[103] E. Grundberg *et al*, "Mapping cis- and trans-regulatory effects across multiple tissues in twins," *Nat. Genet.*, vol. 44, no. 10, pp. 1084–1089, 2012.

[104] Z. Zhu *et al*, "Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets," *Nat. Genet.*, vol. 48, no. 5, pp. 481–487, 2016.

[105] E. P. Consortium *et al*, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[106] N. H. Dryden *et al*, "Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C," *Genome Res.*, vol. 24, no. 11, pp. 1854–1868, 2014.

[107] R. Jäger *et al*, "Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci," *Nat. Commun.*, vol. 6, 2015.

[108] E. S. Amirian *et al*, "The Glioma International Case-Control Study: A Report from the Genetic

Epidemiology of Glioma International Consortium," *Am. J. Epidemiol.*, vol. 183, no. 2, pp. 85–91, 2016.

[109] C. Power and J. Elliott, "Cohort profile: 1958 British birth cohort (National Child Development Study)," *Int. J. Epidemiol.*, vol. 35, no. 1, pp. 34–41, 2006.

[110] H.-E. Wichmann, C. Gieger, and T. Illig, "KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes.," *Gesundheitswesen*, vol. 67 Suppl 1, pp. S26-30, 2005.

[111] M. Krawczak, S. Nikolaus, H. Von Eberstein, P. J. P. Croucher, N. E. El Mokhtari, and S. Schreiber, "PopGen: Population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships," in *Community Genetics*, 2006, vol. 9, no. 1, pp. 55–61.

[112] A. Schmermund *et al*, "Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: Rationale and design of the Heinz Nixdorf RECALL study," *Am. Heart J.*, vol. 144, no. 2, pp. 212–218, 2002.

[113] D. J. Hunter *et al*, "A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer," *Nat. Genet.*, vol. 39, no. 7, pp. 870–874, 2007.

[114] M. Yeager *et al*, "Genome-wide association study of prostate cancer identifies a second risk locus at 8q24," *Nat. Genet.*, vol. 39, no. 5, pp. 645–649, 2007.

[115] J. L. Wiemels, J. K. Wiencke, J. D. Sison, R. Miike, A. McMillan, and M. Wrensch, "History of allergies among adults with glioma and controls," *Int. J. Cancer*, vol. 98, no. 4, pp. 609–615, 2002.

[116] M. J. Felini *et al*, "Reproductive factors and hormone use and risk of adult gliomas," *Cancer Causes Control*, vol. 20, no. 1, pp. 87–96, 2009.

[117] P. Rajaraman *et al*, "Genome-wide association study of glioma and meta-analysis," *Hum. Genet.*, vol. 131, no. 12, pp. 1877–1888, 2012.

[118] C. W. K. Kleihues P, "World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Nervous System," in *Pathology and genetics of the tumors of the nervous system*, 2000, p. 314.

[119] V. Chaitankar, G. Karakülah, R. Ratnapriya, F. O. Giuste, M. J. Brooks, and A. Swaroop, "Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research," *Progress in Retinal and Eye Research*, vol. 55. pp. 1–31, 2016.

[120] R. R Development Core Team, *R: A Language and Environment for Statistical Computing*. 2011.

[121] S. Purcell *et al*, "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *Am. J. Hum. Genet.*, 2007.

[122] I. Pe'er, R. Yelensky, D. Altshuler, and M. J. Daly, "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants," *Genet. Epidemiol.*, 2008.

[123] F. Dudbridge and A. Gusnanto, "Estimation of significance thresholds for genomewide association scans," *Genet. Epidemiol.*, 2008.

[124] C. J. Hoggart, T. G. Clark, M. De Iorio, J. C. Whittaker, and D. J. Balding, "Genome-wide significance for dense SNP and resequencing data," *Genet. Epidemiol.*, 2008.

[125] P. C. Sham and S. M. Purcell, "Statistical power and significance testing in large-scale genetic studies," *Nature Reviews Genetics*. 2014.

[126] E. W. Sayers *et al*, “Database resources of the National Center for Biotechnology Information,” *Nucleic Acids Res.*, 2009.

[127] T. 1000 G. P. 1000 Genomes Project Consortium *et al*, “An integrated map of genetic variation from 1,092 human genomes.,” *Nature*, 2012.

[128] UK10K Consortium, “The UK10K project identifies rare variants in health and disease.,” *Nature*. 2015.

[129] P. Flicek *et al*, “Ensembl 2014,” *Nucleic Acids Res.*, 2014.

[130] A. Subramanian *et al*, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proc. Natl. Acad. Sci.*, vol. 102, no. 43, pp. 15545–15550, 2005.

[131] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, “The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants,” *Nucleic Acids Res.*, 2009.

[132] H. Li *et al*, “The Sequence Alignment / Map format and SAMtools,” *Bioinformatics*, 2009.

[133] P. Danecek *et al*, “The variant call format and VCFtools,” *Bioinformatics*, 2011.

[134] A. Rimmer *et al*, “Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications,” *Nat. Genet.*, 2014.

[135] G. Lunter and M. Goodson, “Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads,” *Genome Res.*, 2011.

[136] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform,” *Bioinformatics*, 2009.

[137] A. McKenna *et al*, “The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data,” *Genome Res.*, 2010.

[138] G. A. Van der Auwera *et al*, “From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline,” *Curr. Protoc. Bioinforma.*, 2013.

[139] I. A. Adzhubei *et al*, “A method and server for predicting damaging missense mutations,” *Nature Methods*. 2010.

[140] P. Kumar, S. Henikoff, and P. C. Ng, “Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm,” *Nat. Protoc.*, 2009.

[141] A. González-Pérez and N. López-Bigas, “Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel,” *Am. J. Hum. Genet.*, 2011.

[142] L. A. Hindorff *et al*, “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits,” *Proc. Natl. Acad. Sci.*, 2009.

[143] T. H. Emigh, “A Comparison of Tests for Hardy-Weinberg Equilibrium,” *Biometrics*, 1980.

[144] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, “A new multipoint method for genome-wide association studies by imputation of genotypes,” *Nat. Genet.*, vol. 39, no. 7, pp. 906–913, 2007.

[145] J. Z. Liu *et al*, “Meta-analysis and imputation refines the association of 15q25 with smoking quantity,” *Nat. Genet.*, vol. 42, no. 5, pp. 436–440, 2010.

[146] J. P. T. Higgins and S. G. Thompson, “Quantifying heterogeneity in a meta-analysis,” *Stat. Med.*, 2002.

[147] D. G. Clayton *et al*, “Population structure, differential bias and genomic control in a large-scale, case-control association study,” *Nat. Genet.*, 2005.

[148] B. Devlin and N. Risch, “A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mappingitle,” *Genomics*, 1995.

[149] L. B. Jorde, “Linkage Disequilibrium and the Search for Complex Disease Genes,” *Genome Res.*, 2000.

[150] K. G. Ardlie, L. Kruglyak, and M. Seielstad, “Patterns of linkage disequilibrium in the human genome,” *Nature Reviews Genetics*. 2002.

[151] S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly, “Genetics: A fine-scale map of recombination rates and hotspots across the human genome,” *Science (80-.).*, 2005.

[152] O. Delaneau, J. Marchini, and J. F. Zagury, “A linear complexity phasing method for thousands of genomes,” *Nat. Methods*, 2012.

[153] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis, “Fast and accurate genotype imputation in genome-wide association studies through pre-phasing,” *Nat. Genet.*, 2012.

[154] G. M. Cooper, E. A. Stone, G. Asimenos, E. D. Green, S. Batzoglou, and A. Sidow, “Distribution and intensity of constraint in mammalian genomic sequence,” *Genome Res.*, 2005.

[155] J. Felsenstein and G. A. Churchill, “A Hidden Markov Model approach to variation among sites in rate of evolution,” *Mol. Biol. Evol.*, 1996.

[156] J. Ernst and M. Kellis, “Discovery and Characterization of Chromatin States for Systematic Annotation of the Human Genome,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011.

[157] L. D. Ward and M. Kellis, “HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants,” *Nucleic Acids Res.*, 2012.

[158] A. P. Boyle *et al*, “Annotation of functional variation in personal genomes using RegulomeDB,” *Genome Res.*, 2012.

[159] D. Hnisz *et al*, “Super-enhancers in the control of cell identity and disease.,” *Cell*, 2013.

[160] Roadmap Epigenomics Consortium *et al*, “Integrative analysis of 111 reference human epigenomes,” *Nature*, 2015.

[161] J. S. Martin *et al*, “HUGIn: Hi-C unifying genomic interrogator,” *Bioinformatics*, vol. 33, no. 23, pp. 3793–3795, 2017.

[162] A. D. Schmitt *et al*, “A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome,” *Cell Rep.*, vol. 17, no. 8, pp. 2042–2059, 2016.

[163] J. R. Dixon *et al*, “Chromatin architecture reorganization during stem cell differentiation,” *Nature*, vol. 518, no. 7539, pp. 331–336, 2015.

[164] E. Crane *et al*, “Condensin-driven remodelling of X chromosome topology during dosage compensation,” *Nature*, vol. 523, no. 7559, pp. 240–244, 2015.

[165] M. Costello *et al*, “Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation,” *Nucleic Acids Res.*, 2013.

[166] M. S. Lawrence *et al*, “Mutational heterogeneity in cancer and the search for new cancer-associated genes,” *Nature*, 2013.

[167] A. Gonzalez-Perez and N. Lopez-Bigas, “Functional impact bias reveals cancer drivers,” *Nucleic Acids Res.*, 2012.

[168] A. Gonzalez-Perez *et al*, “IntOGen-mutations identifies cancer drivers across tumor types,” *Nat. Methods*, 2013.

[169] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz, “GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers,” *Genome Biol.*, 2011.

[170] M. K. Iyer, A. M. Chinnaiyan, and C. A. Maher, “ChimeraScan: A tool for identifying chimeric transcription in sequencing data,” *Bioinformatics*, 2011.

[171] C. A. Maher *et al*, “Chimeric transcript discovery by paired-end transcriptome sequencing,” *Proc. Natl. Acad. Sci.*, 2009.

[172] M. Scales, R. Jäger, G. Migliorini, R. S. Houlston, and M. Y. R. Henrion, “VisPIG - A web tool for producing multi-region, multi-track, multi-scale plots of genetic data,” *PLoS One*, 2014.

[173] L. Wilkinson, “ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H.,” *Biometrics*, vol. 67, no. 2, pp. 678–679, 2011.

[174] M. L. Bondy *et al*, “Brain tumor epidemiology: Consensus from the Brain Tumor Epidemiology Consortium,” *Cancer*, vol. 113, no. 7. pp. 1953–1968, 2008.

[175] B. S. Melin *et al*, “Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors,” *Nat. Genet.*, vol. 49, no. 5, pp. 789–794, 2017.

[176] M. Labussière *et al*, “All the 1p19q codeleted gliomas are mutated on IDH1 or IDH2,” *Neurology*, vol. 74, no. 23, pp. 1886–1890, 2010.

[177] A. L. Di Stefano *et al*, “Association between glioma susceptibility loci and tumour pathology defines specific molecular etiologies,” *Neuro. Oncol.*, vol. 15, no. 5, pp. 542–547, 2013.

[178] K. Wang *et al*, “PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data,” *Genome Res.*, vol. 17, no. 11, pp. 1665–1674, 2007.

[179] M. Ceccarelli *et al*, “Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma,” *Cell*, vol. 164, no. 3, pp. 550–563, 2016.

[180] J. Gao *et al*, “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal,” *Sci. Signal.*, vol. 6, no. 269, 2013.

[181] S. Hercberg *et al*, “The SU. VI. MAX study a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals,” *Arch. Intern. Med.*, vol. 164, no. 21, pp. 2335–2342, 2004.

[182] A. Gonzalez-Aguilar *et al*, “Recurrent mutations of MYD88 and TBL1XR1 in primary central nervous system lymphomas,” *Clin. Cancer Res.*, vol. 18, no. 19, pp. 5203–5211, 2012.

[183] A. Idbaih *et al*, “BAC array CGH distinguishes mutually exclusive alterations that define clinicogenetic subtypes of gliomas,” *Int. J. Cancer*, vol. 122, no. 8, pp. 1778–1786, 2008.

[184] M. Sanson *et al*, "Isocitrate dehydrogenase 1 codon 132 mutation is an important prognostic biomarker in gliomas," *J. Clin. Oncol.*, vol. 27, no. 25, pp. 4150–4154, 2009.

[185] 3C Study Group, "Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population.," *Neuroepidemiology*, vol. 22, no. 6, pp. 316–325, 2003.

[186] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS Genet.*, vol. 2, no. 12, pp. 2074–2093, 2006.

[187] J. P. T. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman, "Measuring inconsistency in meta-analyses," *BMJ Br. Med. J.*, vol. 327, no. 7414, pp. 557–560, 2003.

[188] J. Lonsdale *et al*, "The Genotype-Tissue Expression (GTEx) project," *Nature Genetics*, vol. 45, no. 6. pp. 580–585, 2013.

[189] H. J. Westra *et al*, "Systematic identification of trans eQTLs as putative drivers of known disease associations," *Nat. Genet.*, vol. 45, no. 10, pp. 1238–1243, 2013.

[190] V. K. Mootha *et al*, "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nat. Genet.*, vol. 34, no. 3, pp. 267–273, 2003.

[191] S. S. P. Rao *et al*, "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, no. 7, pp. 1665–1680, 2014.

[192] V. Sanz-Moreno *et al*, "ROCK and JAK1 Signaling Cooperate to Control Actomyosin Contractility in Tumor Cells and Stroma," *Cancer Cell*, 2011.

[193] F. Ango, G. Di Cristo, H. Higashiyama, V. Bennett, P. Wu, and Z. J. Huang, "Ankyrin-based subcellular gradient of neurofascin, an immunoglobulin family protein, directs GABAergic innervation at Purkinje axon initial segment," *Cell*, 2004.

[194] D. Brodbeck, P. Cron, and B. A. Hemmings, "A human protein kinase Bgamma with regulatory phosphorylation sites in the activation loop and in the C-terminal hydrophobic domain.," *J Biol Chem.*, 1999.

[195] J. H. Lee *et al*, "De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly," *Nat. Genet.*, 2012.

[196] S. Li *et al*, "Overexpression of isocitrate dehydrogenase mutant proteins renders glioma cells more sensitive to radiation," *Neuro. Oncol.*, 2013.

[197] V. Codd *et al*, "Identification of seven loci affecting mean telomere length and their association with disease," *Nat. Genet.*, 2013.

[198] C. Zhao *et al*, "Dual regulatory switch through interactions of Tcf7l2/Tcf4 with stage-specific partners propels oligodendroglial maturation," *Nat. Commun.*, 2016.

[199] J. Jung *et al*, "Nicotinamide metabolism regulates glioblastoma stem cell maintenance," *J. Clin. Invest.*, 2017.

[200] Q. L. Zhou *et al*, "A novel pleckstrin homology domain-containing protein enhances insulin-stimulated Akt phosphorylation and GLUT4 translocation in adipocytes," *J. Biol. Chem.*, 2010.

[201] C. Ren, C. H. Ren, L. Li, A. A. Goltsov, and T. C. Thompson, "Identification and characterization of RTVP1/GLIPR1-like genes, a novel p53 target gene cluster," *Genomics*, 2006.

[202] F. Moreira *et al*, "NPAS3 demonstrates features of a tumor suppressive role in driving the progression of astrocytomas," *Am. J. Pathol.*, 2011.

[203] G. E. Schepers, M. Bullejos, B. M. Hosking, and P. Koopman, "Cloning and characterisation of the Sry-related transcription factor gene Sox8.," *Nucleic Acids Res.*, 2000.

[204] W. Zhang *et al*, "Extended haplotype association study in Crohn's disease identifies a novel, Ashkenazi Jewish-specific missense mutation in the NF-κB pathway gene, HEATR3," *Genes Immun.*, 2013.

[205] Y. Zhang *et al*, "Overexpression of SCLIP promotes growth and motility in glioblastoma cells," *Cancer Biol. Ther.*, 2015.

[206] K. M. Walsh *et al*, "Telomere maintenance and the etiology of adult glioma," *Neuro-Oncology*. 2015.

[207] K. M. Walsh *et al*, "Longer genotypically-estimated leukocyte telomere length is associated with increased adult glioma risk.," *Oncotarget*, 2015.

[208] C. Zhang *et al*, "Genetic determinants of telomere length and risk of common cancers: A Mendelian randomization study," *Hum. Mol. Genet.*, 2015.

[209] G. Gur *et al*, "LRIG1 restricts growth factor signaling by enhancing receptor ubiquitylation and degradation," *EMBO J.*, 2004.

[210] J. Wei *et al*, "miR-20a mediates temozolomide-resistance in glioblastoma cells via negatively regulating LRIG1 expression," *Biomed Pharmacother*, 2015.

[211] J. A. Yang *et al*, "LRIG1 enhances the radiosensitivity of radioresistant human glioblastoma U251 cells via attenuation of the EGFR/Akt signaling pathway," *Int. J. Clin. Exp. Pathol.*, 2015.

[212] J. Okano, I. Gaslightwala, M. J. Birnbaum, a K. Rustgi, and H. Nakagawa, "Akt/protein kinase B isoforms are differentially regulated by epidermal growth factor stimulation.," *J. Biol. Chem.*, 2000.

[213] K. M. Turner *et al*, "Genomically amplified Akt3 activates DNA repair pathway and promotes glioma progression," *Proc. Natl. Acad. Sci.*, 2015.

[214] A. Bonni *et al*, "Regulation of gliogenesis in the central nervous system by the JAK-STAT signaling pathway," *Science (80-.).*, 1997.

[215] O. K. Park, T. S. Schaefer, and D. Nathans, "In vitro activation of Stat3 by epidermal growth factor receptor kinase.," *Proc. Natl. Acad. Sci. U. S. A.*, 1996.

[216] K. Shuai *et al*, "Polypeptide signalling to the nucleus through tyrosine phosphorylation of Jak and Stat proteins.," *Nature*, 1993.

[217] O. A. Ulanovskaya, A. M. Zuhl, and B. F. Cravatt, "NNMT promotes epigenetic remodeling in cancer by creating a metabolic methylation sink," *Nat. Chem. Biol.*, 2013.

[218] H. Wang *et al*, "Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A," *Nat. Commun.*, 2014.

[219] W. A. Flavahan *et al*, "Insulator dysfunction and oncogene activation in IDH mutant gliomas," *Nature*, 2016.

[220] M. J. Riemenschneider, T. H. Koy, and G. Reifenberger, "Expression of oligodendrocyte lineage genes in oligodendroglial and astrocytic gliomas," *Acta Neuropathol.*, 2004.

[221] S. Yip *et al*, “Concurrent CIC mutations, IDH mutations, and 1p/19q loss distinguish oligodendrogliomas from other cancers,” *J. Pathol.*, 2012.

[222] C. Greenman *et al*, “Patterns of somatic mutation in human cancer genomes,” *Nature*, 2007.

[223] V. P. Sharma *et al*, “Mutations in TCF12, encoding a basic helix-loop-helix partner of TWIST1, are a frequent cause of coronal craniosynostosis,” *Nat. Genet.*, 2013.

[224] C. W. Brennan *et al*, “The somatic genomic landscape of glioblastoma,” *Cell*, 2013.

[225] D. Singh *et al*, “Transforming fusions of FGFR and TACC genes in human glioblastoma,” *Science (80-.).*, 2012.

[226] A. L. Di Stefano *et al*, “Detection, characterization, and inhibition of FGFR-TACC fusions in IDH wild-type glioma,” *Clin. Cancer Res.*, 2015.

[227] S. Seshagiri *et al*, “Recurrent R-spondin fusions in colon cancer,” *Nature*, 2012.

[228] M. Imielinski *et al*, “Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing,” *Cell*, 2012.

[229] C. C. Lee *et al*, “TCF12 protein functions as transcriptional repressor of E-cadherin, and its overexpression is correlated with metastasis of colorectal cancer,” *J. Biol. Chem.*, 2012.

[230] M. Sideridou *et al*, “Cdc6 expression represses E-cadherin transcription and activates adjacent replication origins,” *J. Cell Biol.*, 2011.

[231] M. H. Yang *et al*, “Bmi1 is essential in Twist1-induced epithelial-mesenchymal transition,” *Nat. Cell Biol.*, 2010.

[232] M. Uittenbogaard and a Chiaramello, “Expression of the bHLH transcription factor Tcf12 (ME1) gene is linked to the expansion of precursor cell populations during neurogenesis.,” *Brain Res. Gene Expr. Patterns*, 2002.

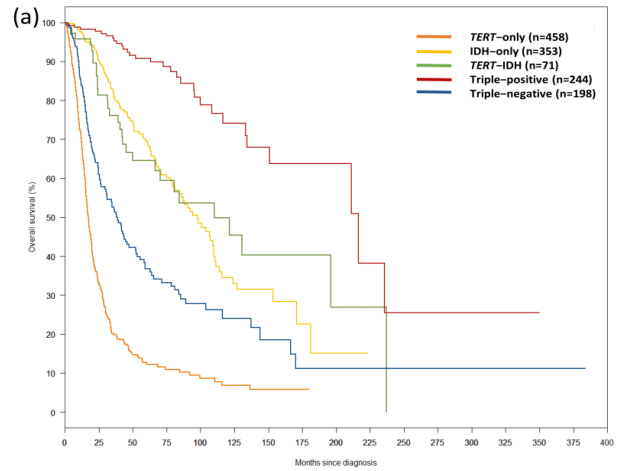
[233] H. Fu *et al*, “A Genome-Wide Screen for Spatially Restricted Expression Patterns Identifies Transcription Factors That Regulate Glial Development,” *J. Neurosci.*, 2009.

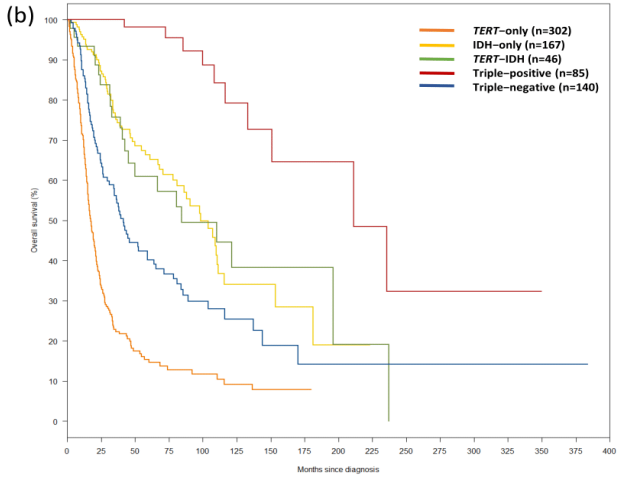
[234] J. Paredes *et al*, “Epithelial E- and P-cadherins: Role and clinical significance in cancer,” *Biochimica et Biophysica Acta - Reviews on Cancer*. 2012.

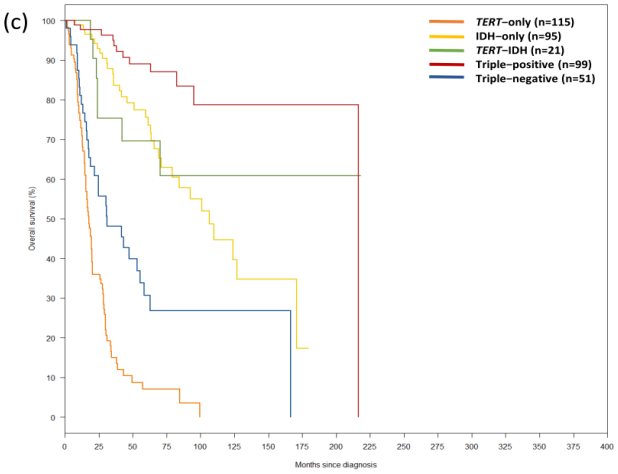
[235] B. Paumard-Hernández *et al*, “Expanding the mutation spectrum in 182 Spanish probands with craniosynostosis: Identification and characterization of novel TCF12 variants,” *Eur. J. Hum. Genet.*, 2015.

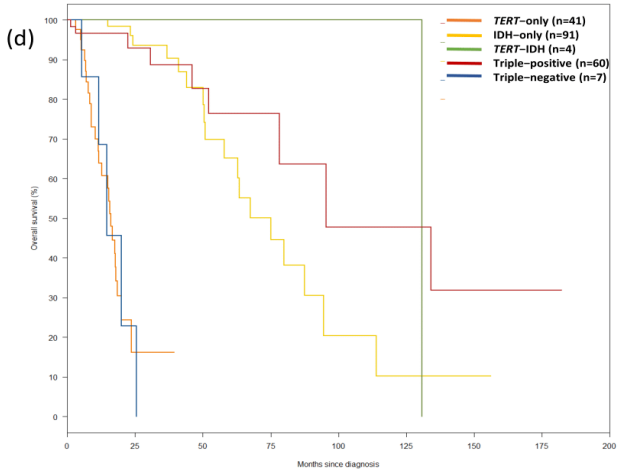
APPENDIX 1

Supplementary Figure 3: Survival by molecular subgroup. (a) All patients cohorts combined – all glioma; (b) French GWAS – all glioma; (c) French sequenced – all glioma; (d) TCGA- all glioma; (e) All dataset combined – GBM; (f) French GWAS – GBM; (g) French sequenced – GBM; (h) TCGA – GBM; (i) All dataset combined – nonGBM; (j) French GWAS – non-GBM; (k) French sequenced – non-GBM; (l) TCGA non-GBM

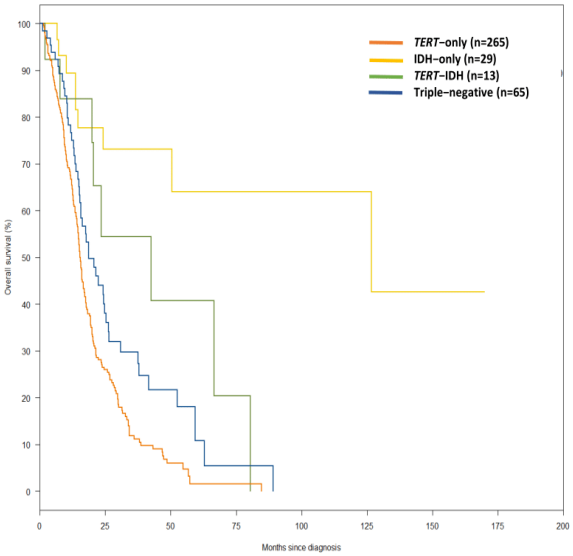




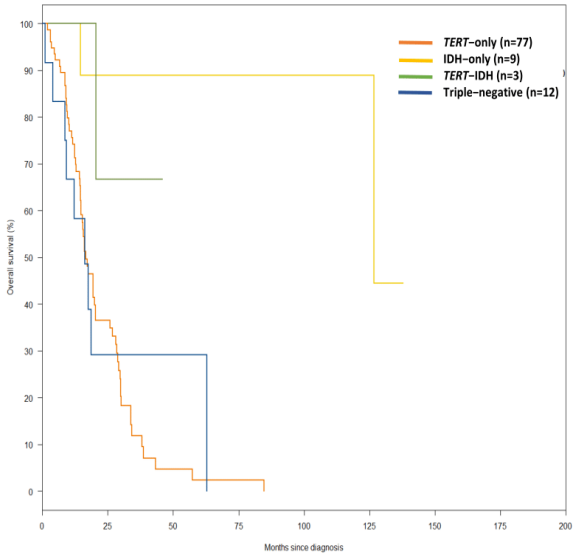




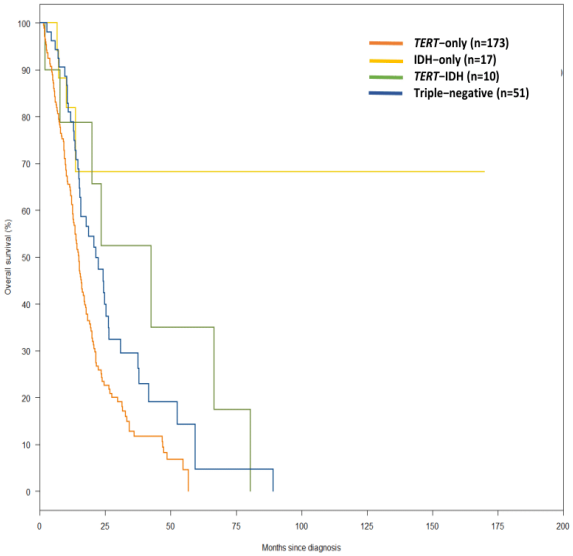
(e)


TERT–only (n=265)
IDH–only (n=29)
TERT–IDH (n=13)
Triple–negative (n=65)
Overall survival (%)
Months since diagnosis

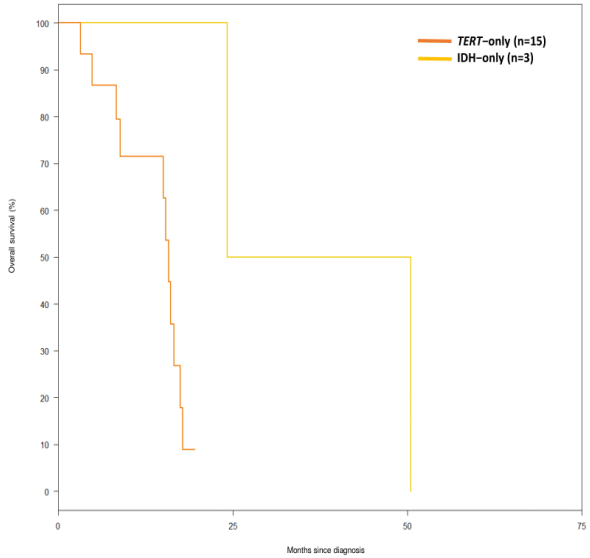
(f)


TERT–only (n=77)
IDH–only (n=9)
TERT–IDH (n=3)
Triple–negative (n=12)
Overall survival (%)
Months since diagnosis

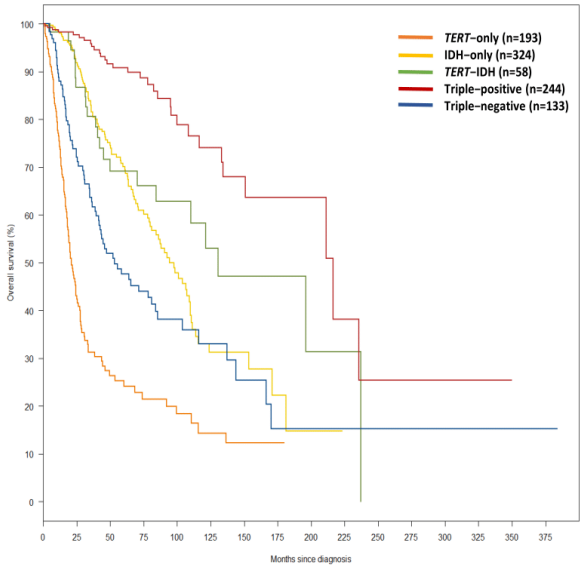
(g)


TERT–only (n=173)
IDH–only (n=17)
TERT–IDH (n=10)
Triple–negative (n=51)
Overall survival (%)
Months since diagnosis

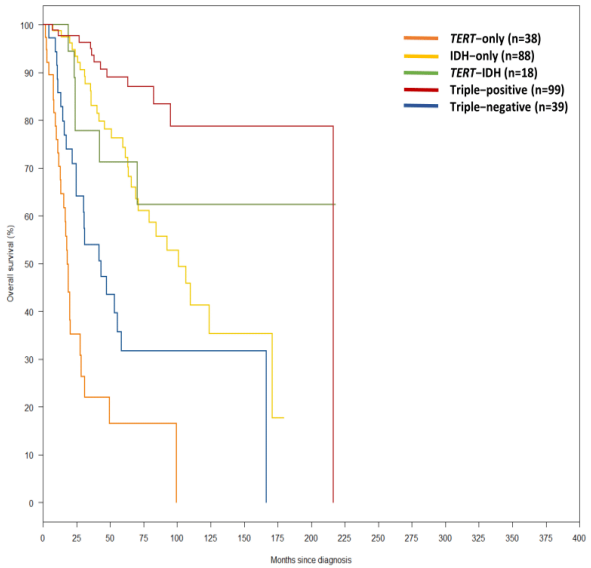
(h)


TERT–only (n=15)
IDH–only (n=3)
Overall survival (%)
Months since diagnosis

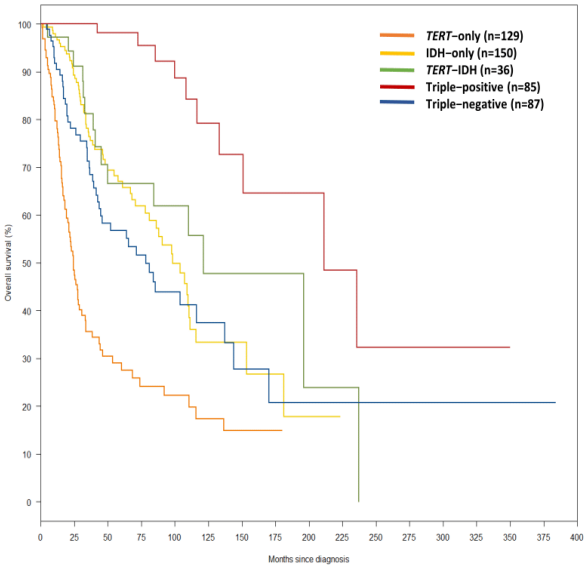
(i)


TERT–only (n=193)
IDH–only (n=324)
TERT–IDH (n=58)
Triple–positive (n=244)
Triple–negative (n=133)
Overall survival (%)
Months since diagnosis

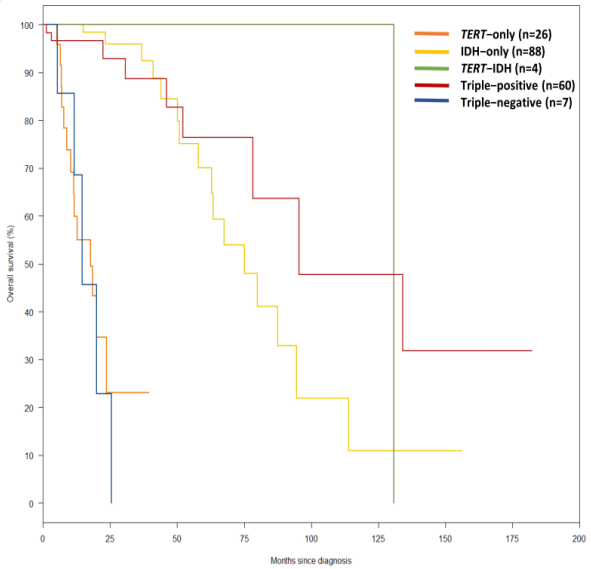
(j)

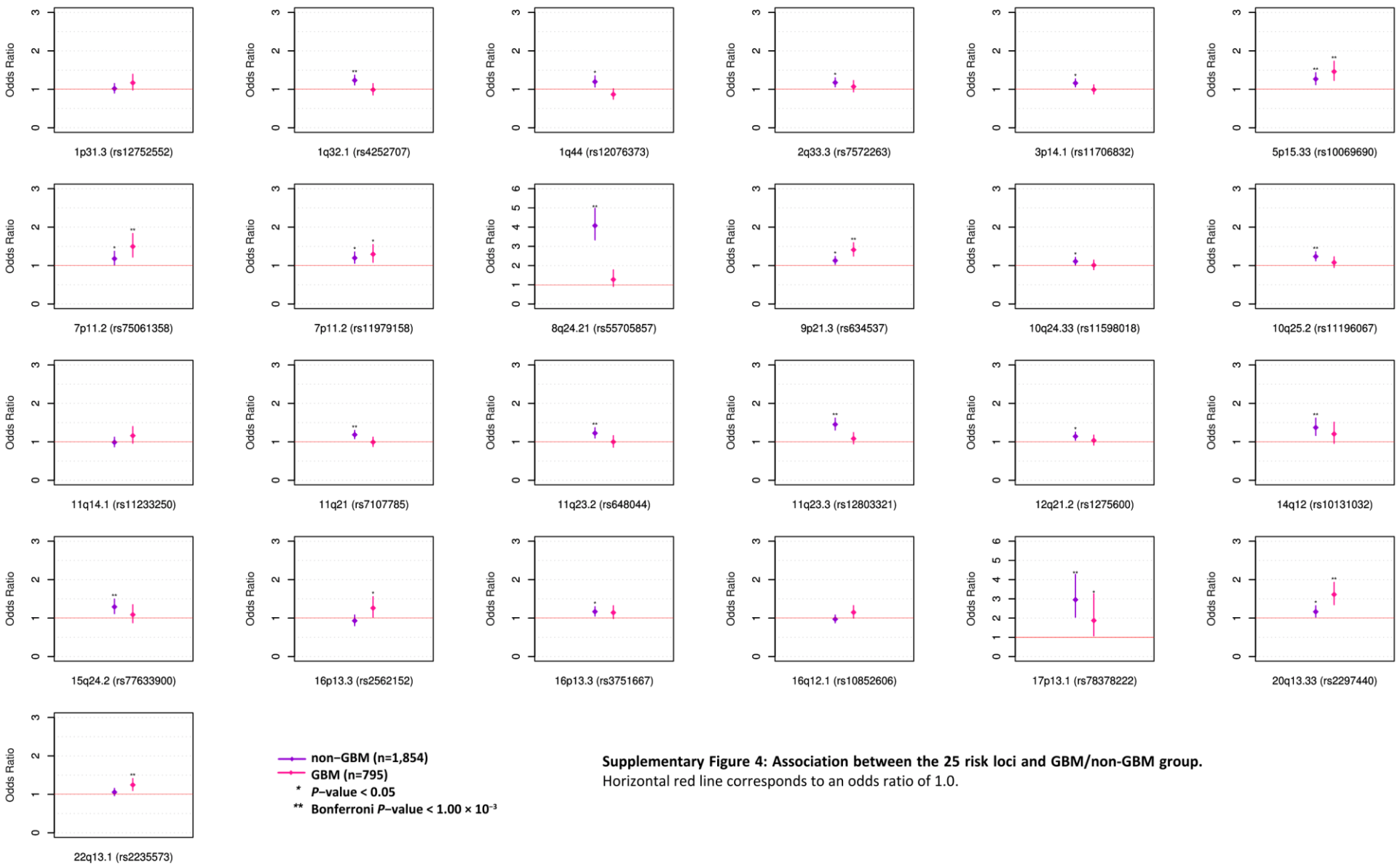

TERT–only (n=38)
IDH–only (n=88)
TERT–IDH (n=18)
Triple–positive (n=99)
Triple–negative (n=39)
Overall survival (%)
Months since diagnosis

(k)

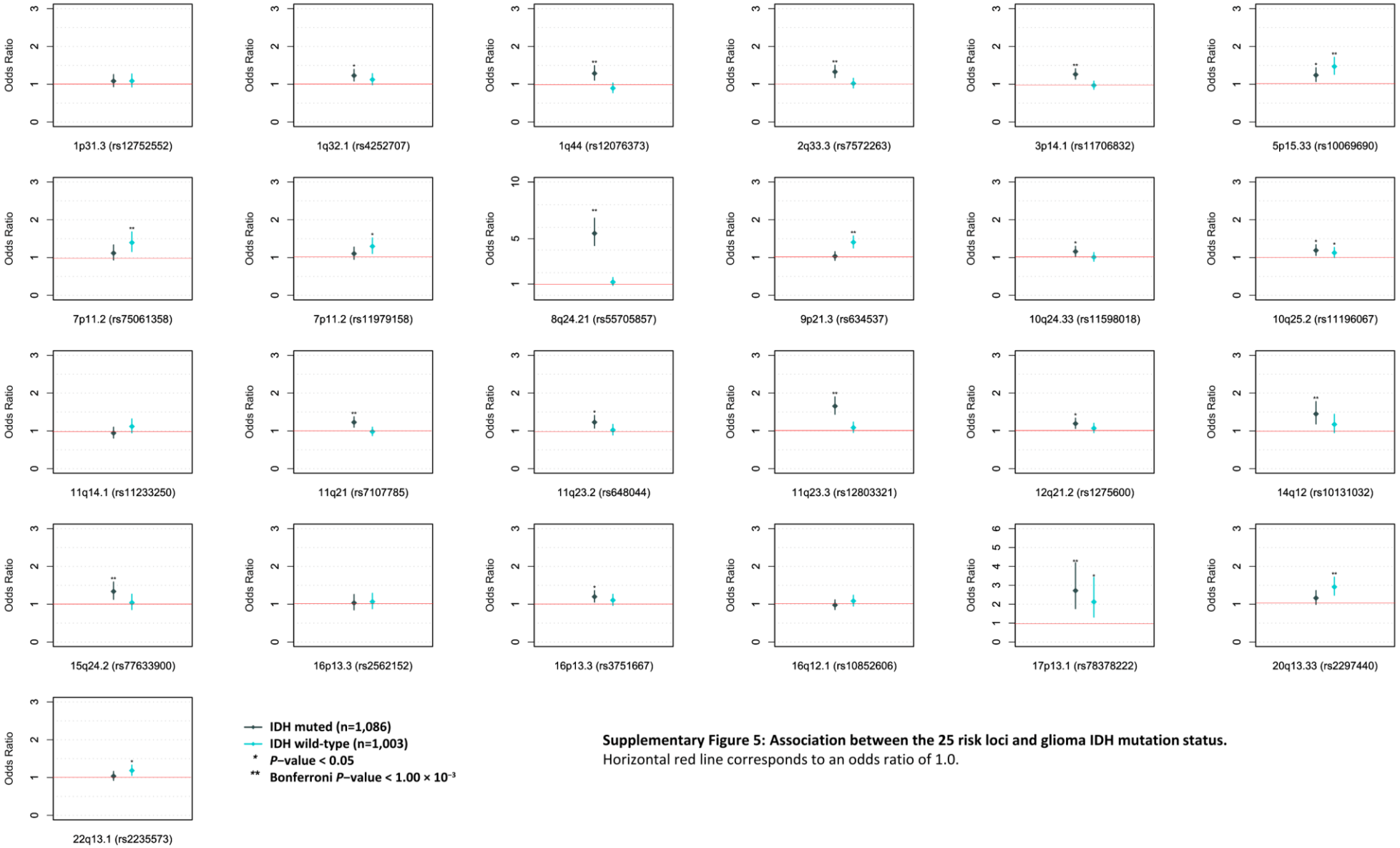

TERT–only (n=129)
IDH–only (n=150)
TERT–IDH (n=36)
Triple–positive (n=85)
Triple–negative (n=87)
Overall survival (%)
Months since diagnosis

(l)

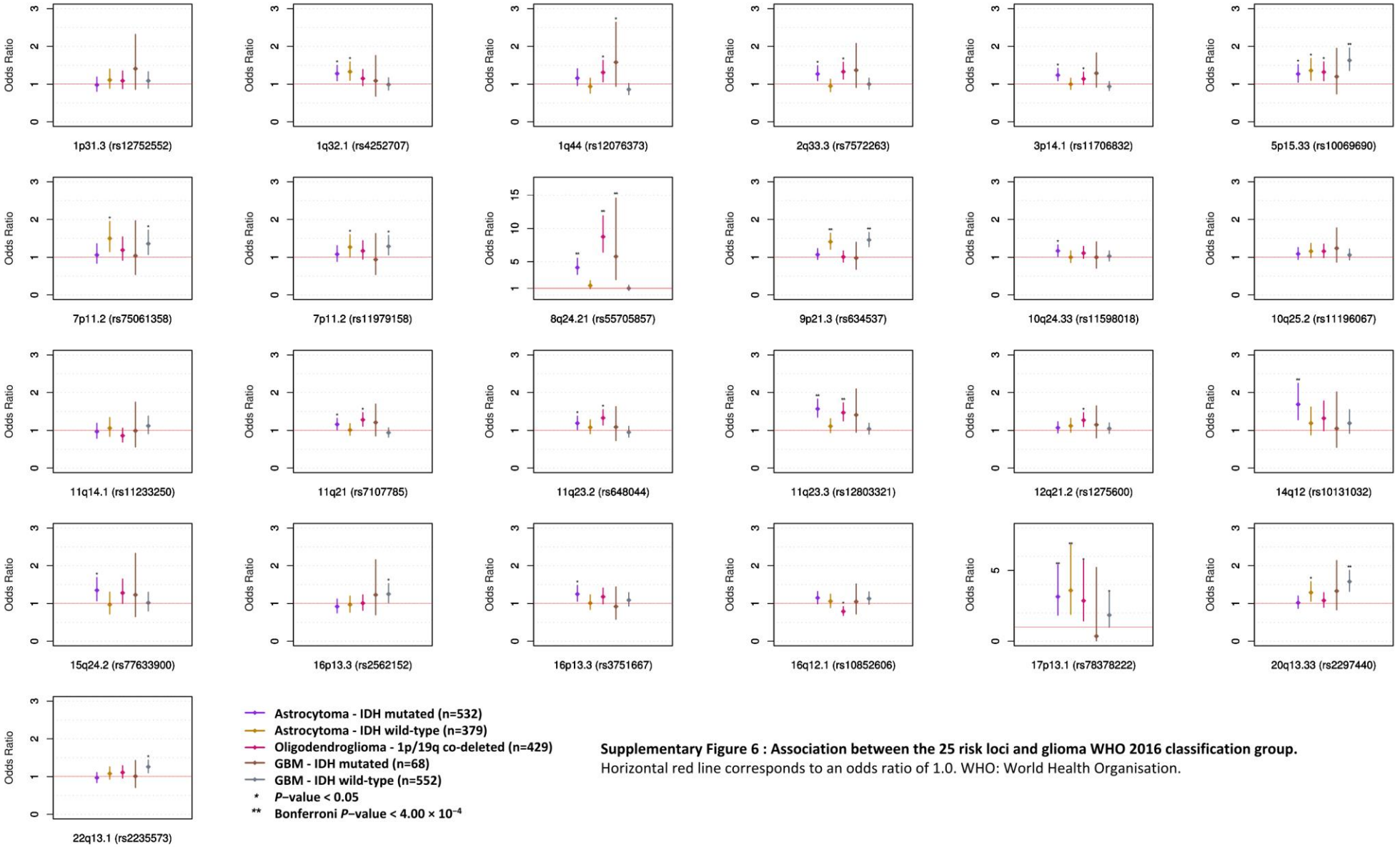

TERT–only (n=26)
IDH–only (n=88)
TERT–IDH (n=4)
Triple–positive (n=60)
Triple–negative (n=7)
Overall survival (%)
Months since diagnosis



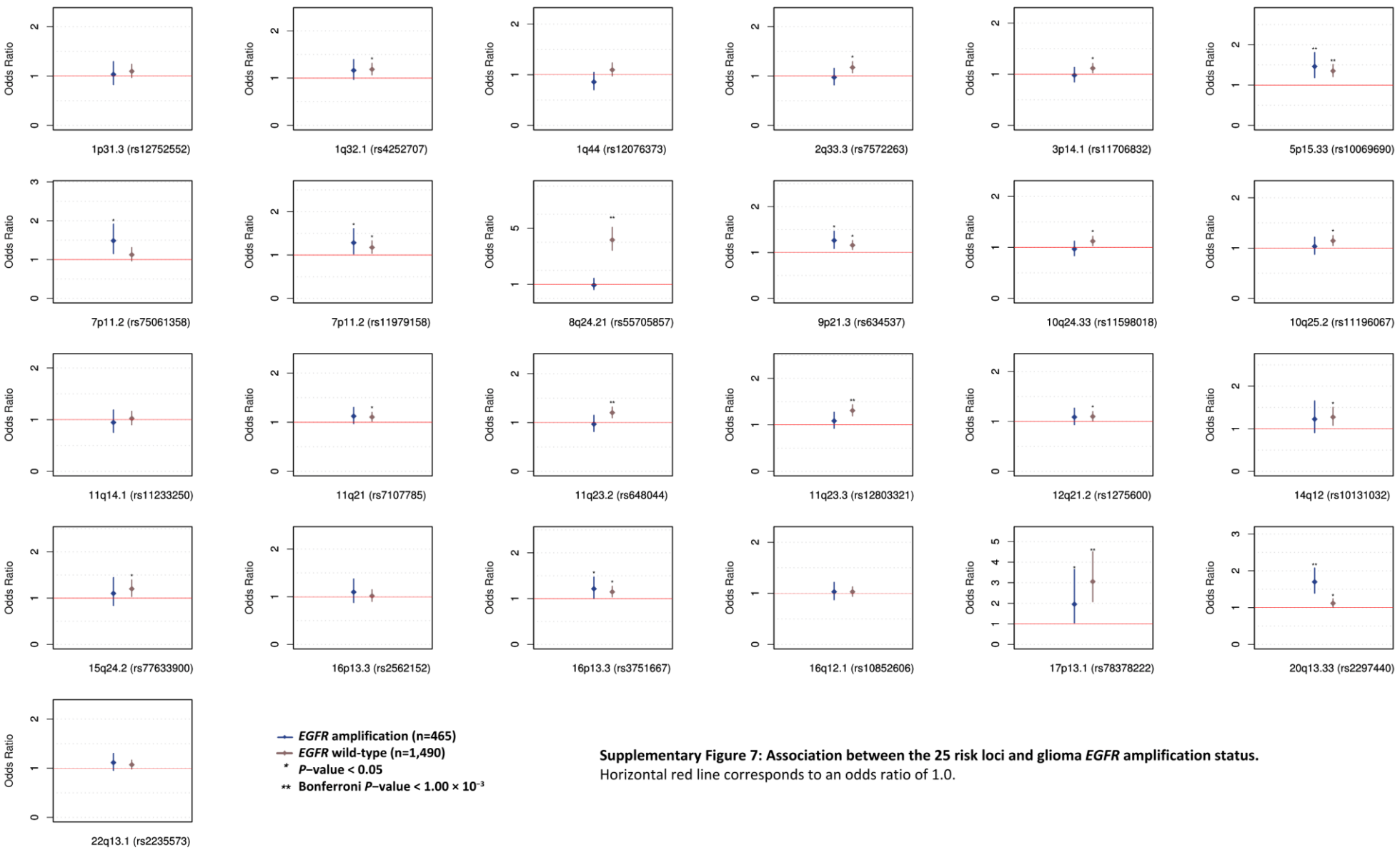
Supplementary Figure 4: Association between the 25 risk loci and GBM/non-GBM group.
Horizontal red line corresponds to an odds ratio of 1.0.



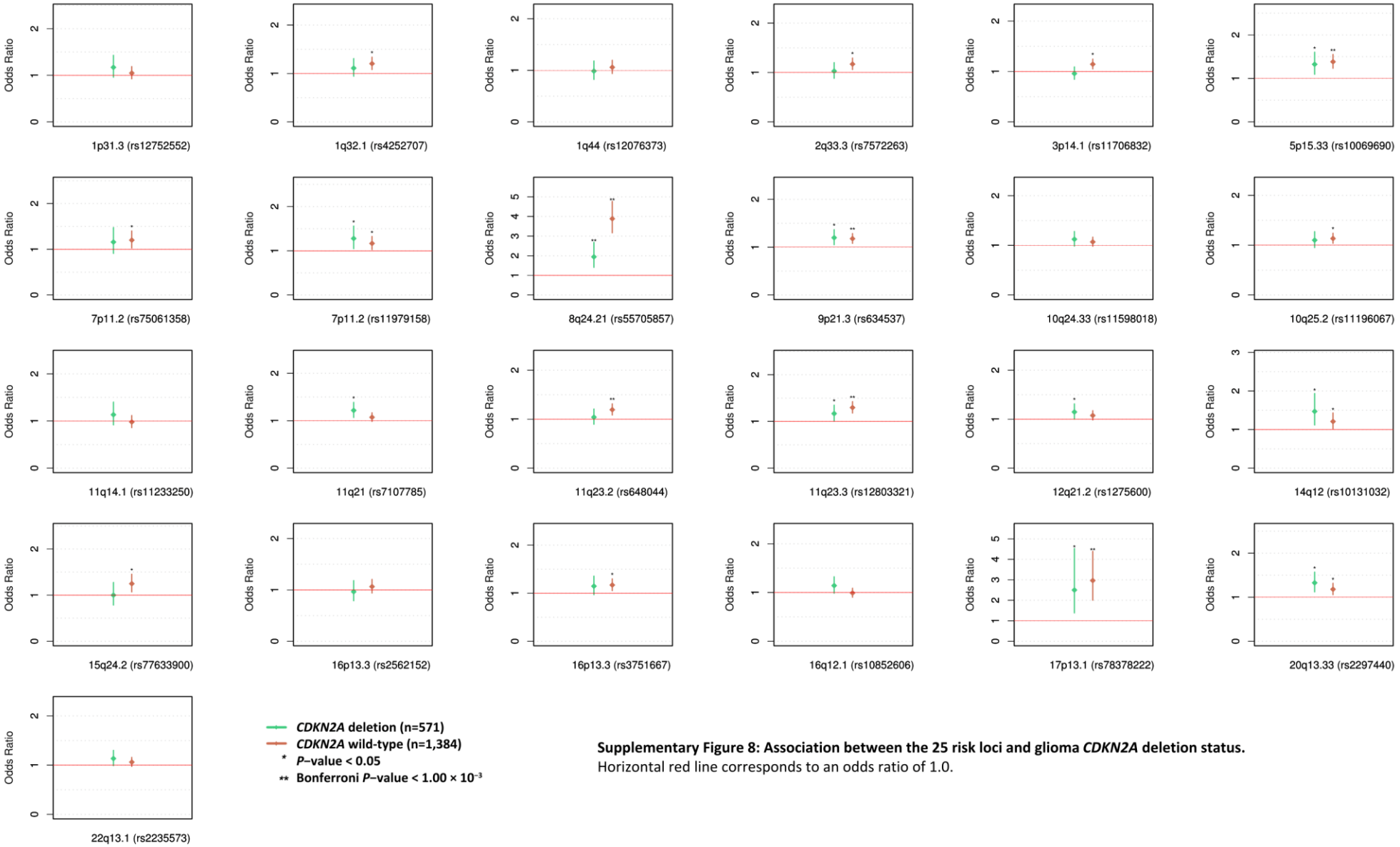
Supplementary Figure 5: Association between the 25 risk loci and glioma IDH mutation status.
Horizontal red line corresponds to an odds ratio of 1.0.



Supplementary Figure 6 : Association between the 25 risk loci and glioma WHO 2016 classification group.
Horizontal red line corresponds to an odds ratio of 1.0. WHO: World Health Organisation.

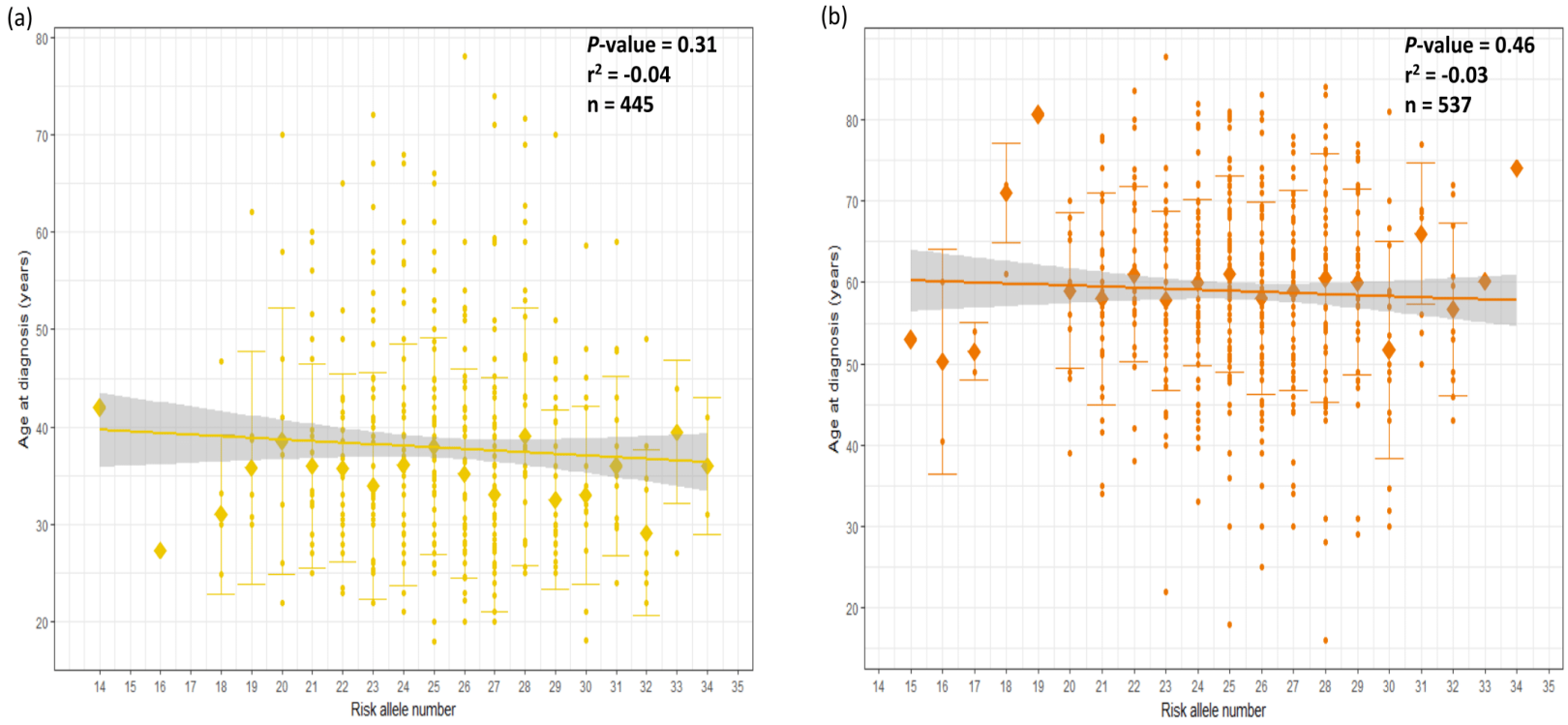


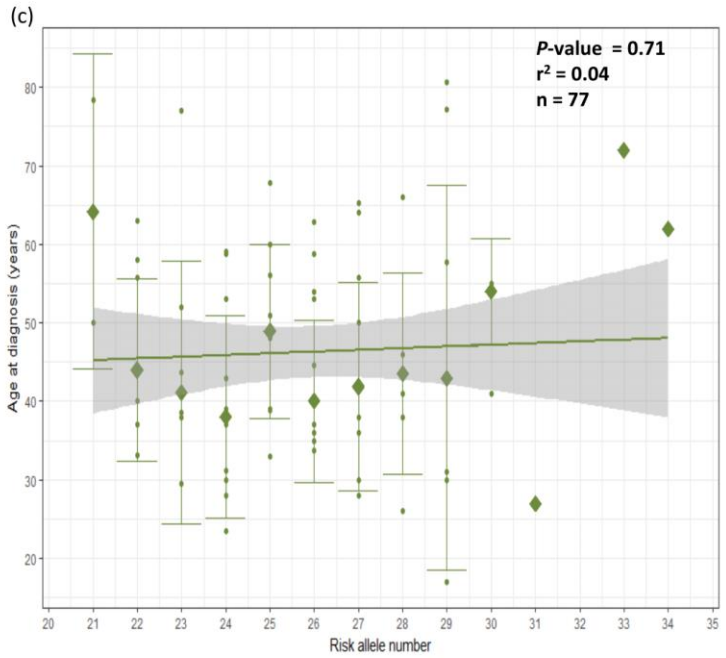
Supplementary Figure 7: Association between the 25 risk loci and glioma *EGFR* amplification status.
Horizontal red line corresponds to an odds ratio of 1.0.

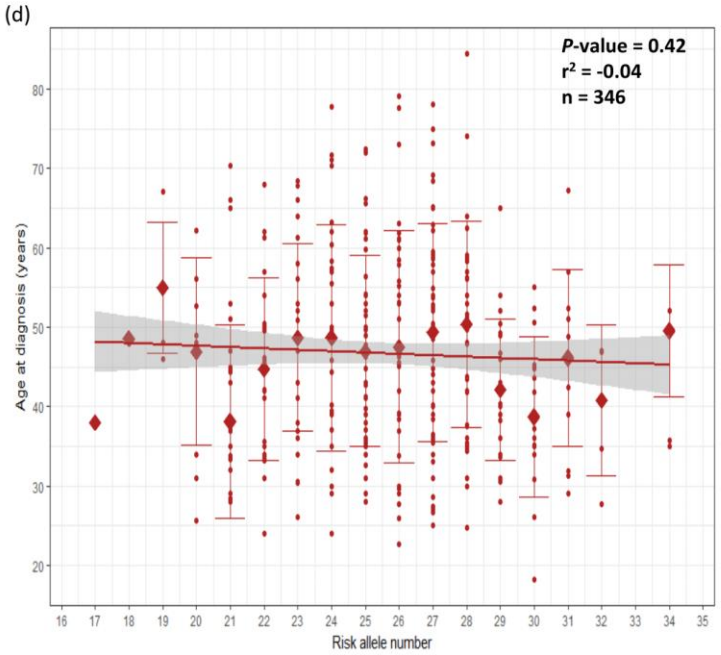


Supplementary Figure 8: Association between the 25 risk loci and glioma *CDKN2A* deletion status.
Horizontal red line corresponds to an odds ratio of 1.0.

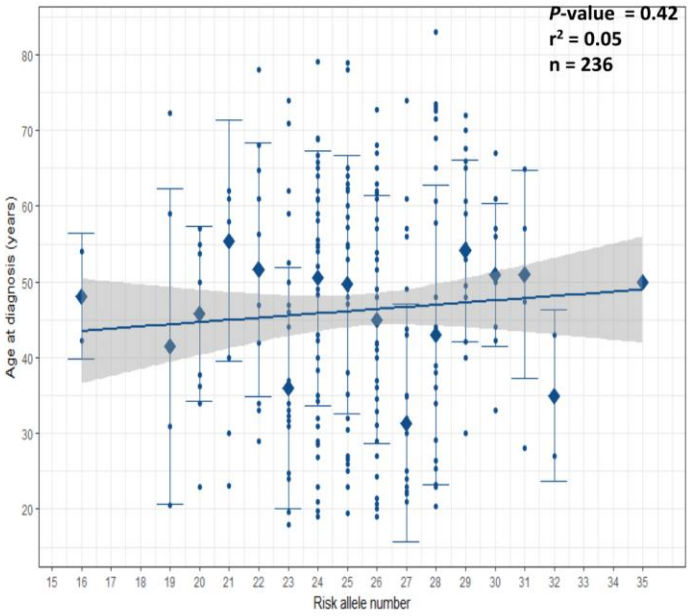
Supplementary Figure 9: Effect of risk allele number on age at diagnosis by glioma molecular subgroup. (a) IDH-only; (b) *TERT*-only; (c) *TERT*-IDH; (d) Triple-positive; (e) Triple-negative; (f) Combined. *P*-values and correlation statistics in the top right corner are Pearson correlation statistics between risk allele number and age at diagnosis.



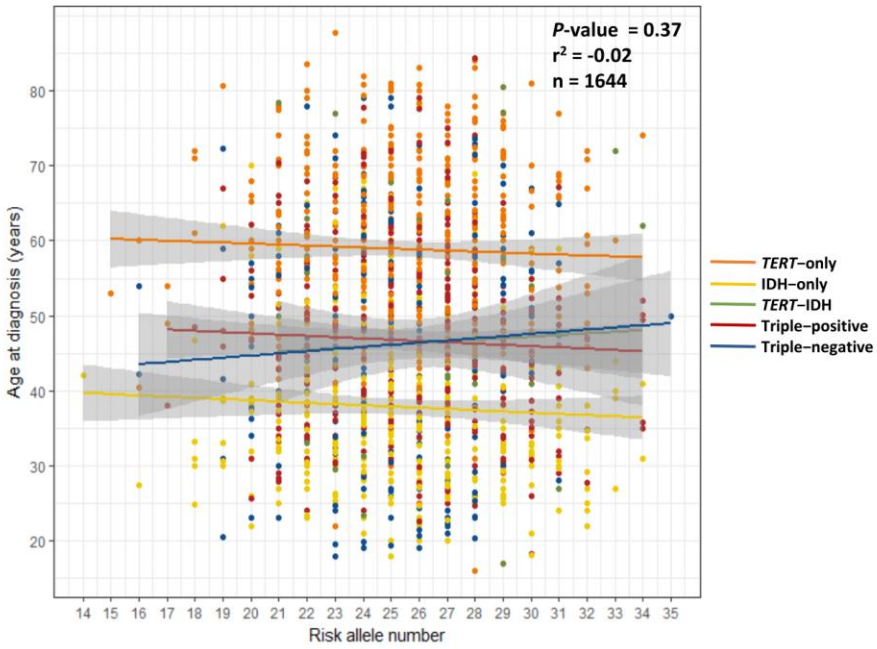

(c)
P-value = 0.71
r² = 0.04
n = 77
Age at diagnosis (years)
Risk allele number


(d)
P-value = 0.42
r² = -0.04
n = 346
Age at diagnosis (years)
Risk allele number

(e)

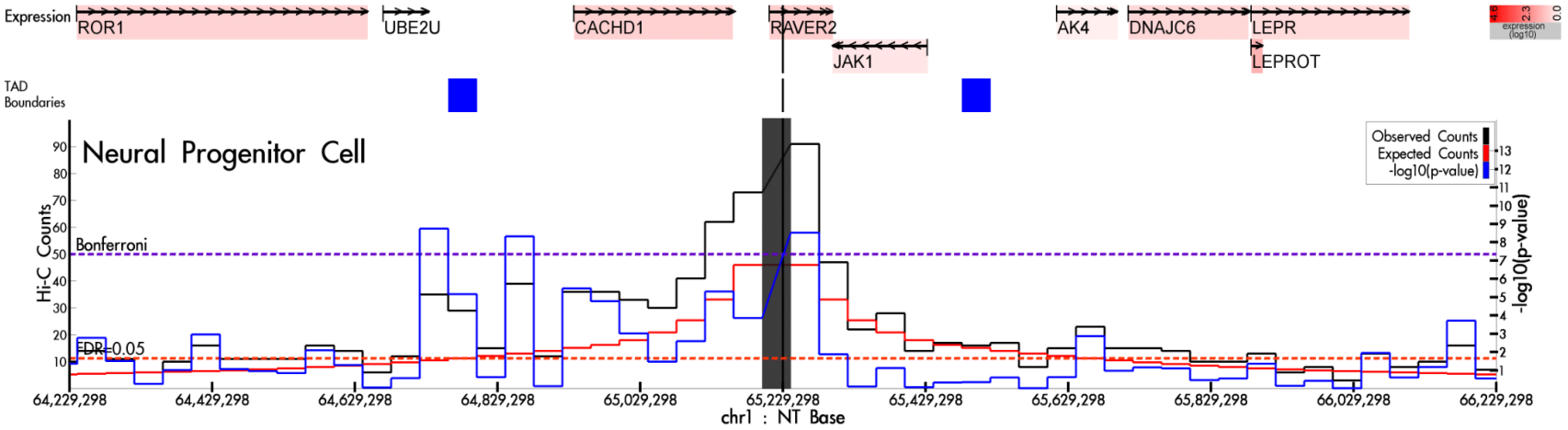

P-value = 0.42
r² = 0.05
n = 236
Age at diagnosis (years)
Risk allele number

(f)

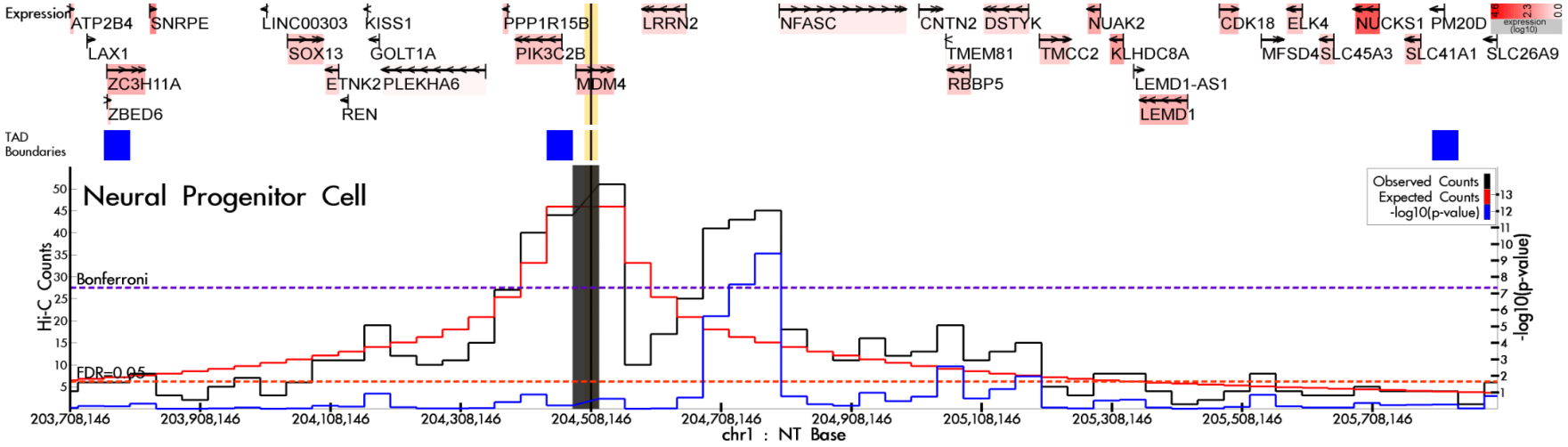

P-value = 0.37
r² = -0.02
n = 1644
Age at diagnosis (years)
Risk allele number
TERT–only
IDH–only
TERT–IDH
Triple–positive
Triple–negative

Supplementary Figure 10: Plots of Hi-C interactions in H1 neuronal progenitor cells at the 25 risk loci. Plots were generated using the HUGIn browser [2]. Each plot shows a "virtual 4C" of all Hi-C interactions with "bait" fragments overlapping the glioma risk SNP of interest (indicated by the shaded rectangle). Topologically associating domain (TAD) boundaries are plotted as filled blue rectangles. The purple dotted line represents the Bonferroni threshold, with interactions exceeding this threshold treated as statistically significant.

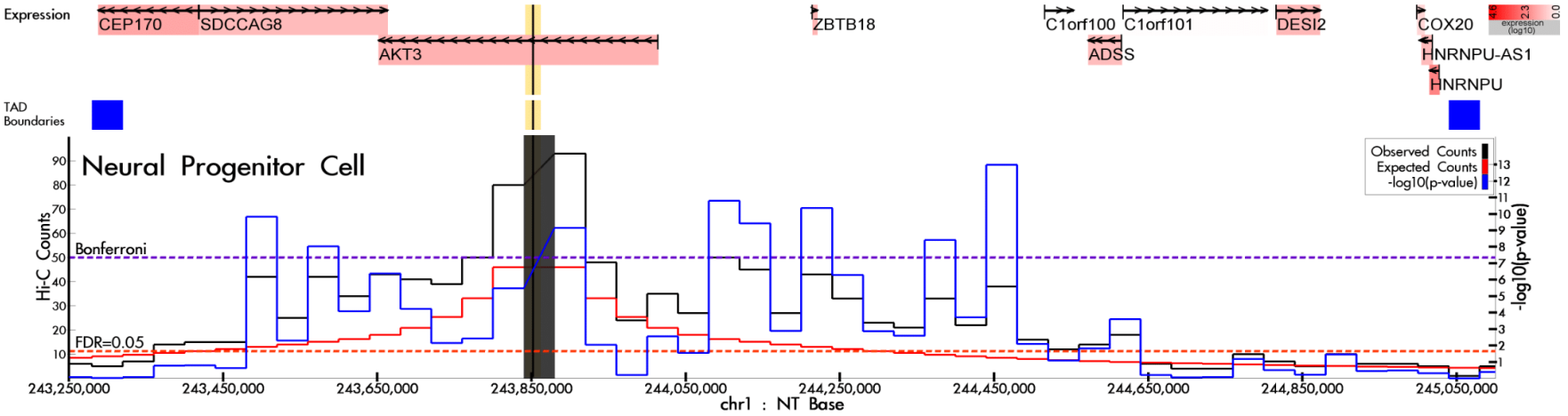
(a) 1p31.3 - rs1272552



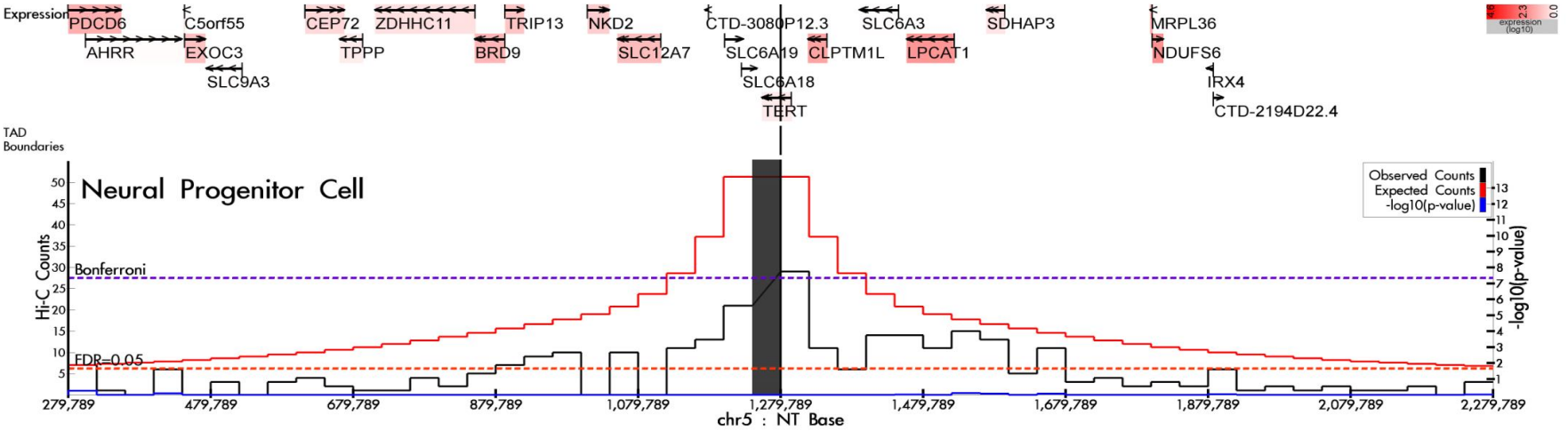
(b) 1q32.1 - rs4252707



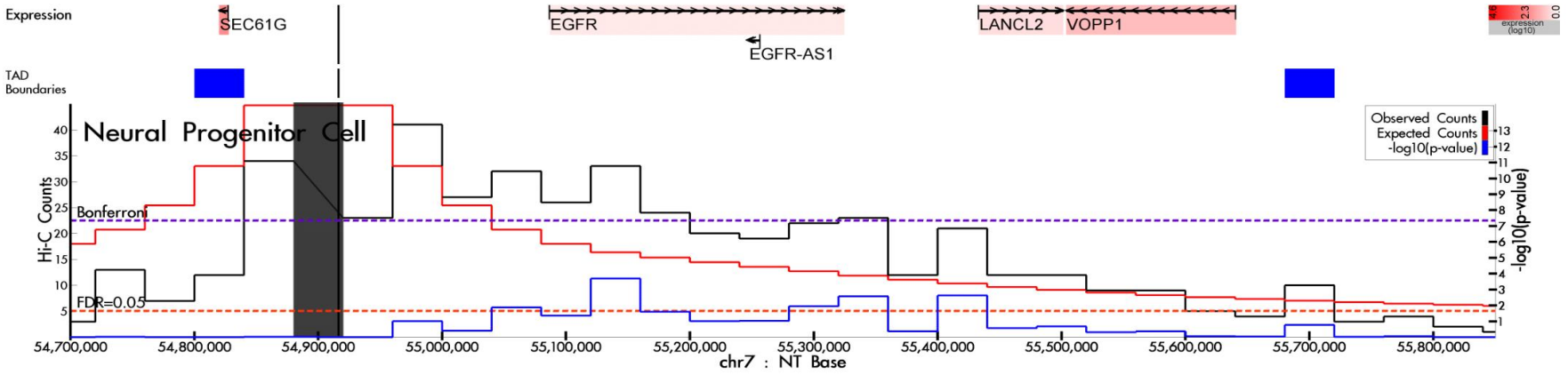
(c) 1q44 - rs12076373



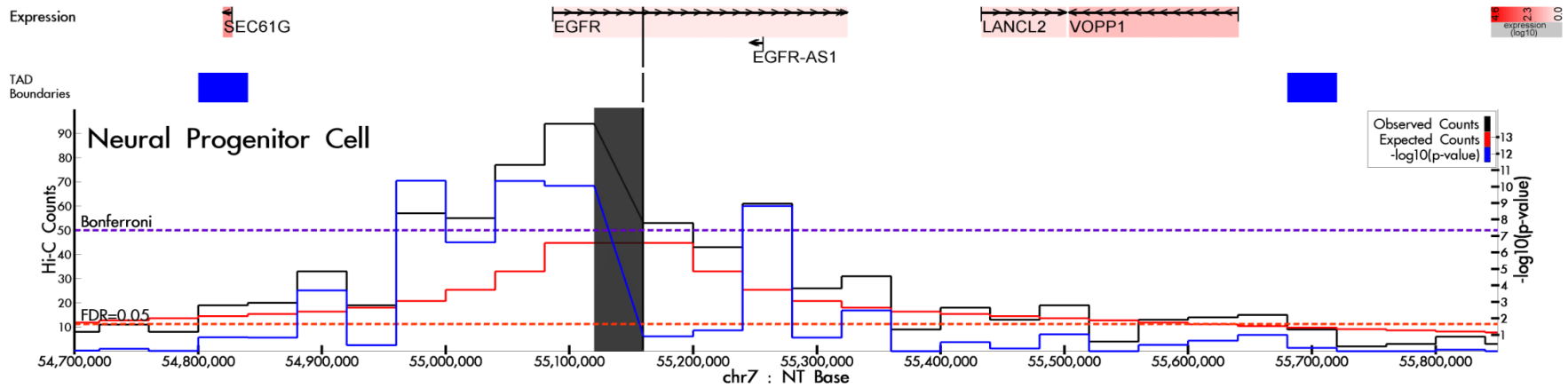
(f) 5p15.33 - rs10069690



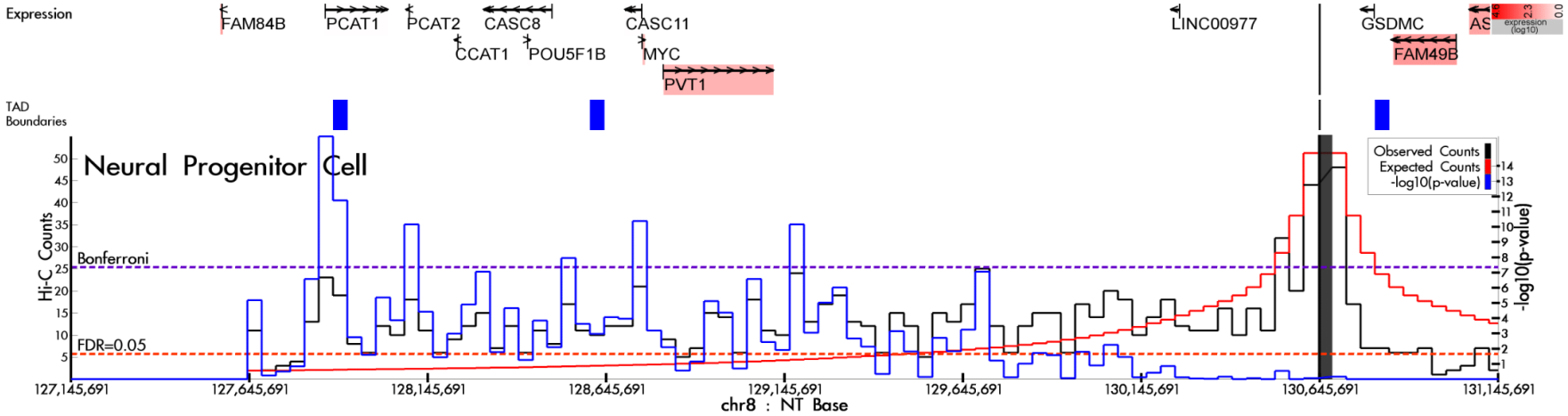
(g) 7p11.2 - rs75061358



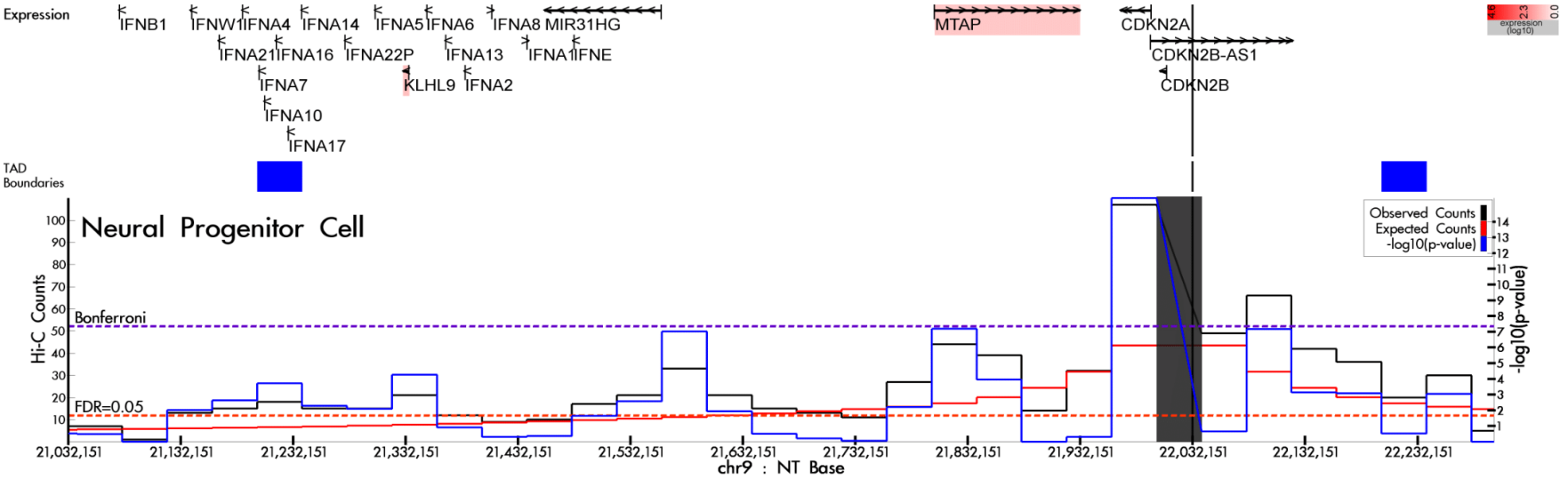
(h) 7p11.2 - rs11979158



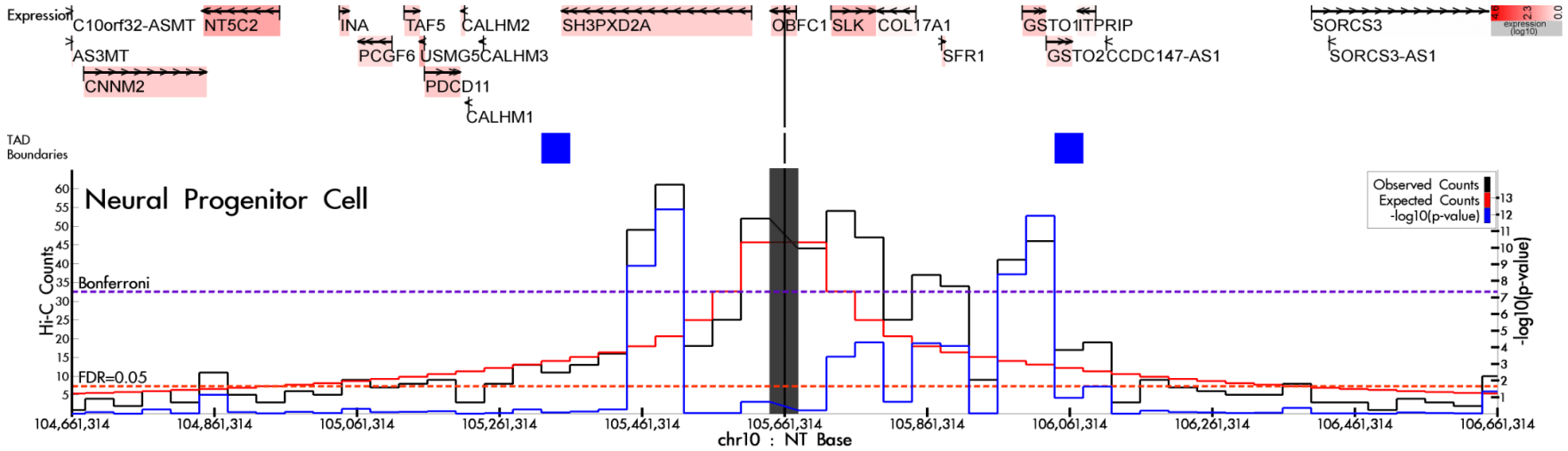
(i) 8q24.21 - rs55705857



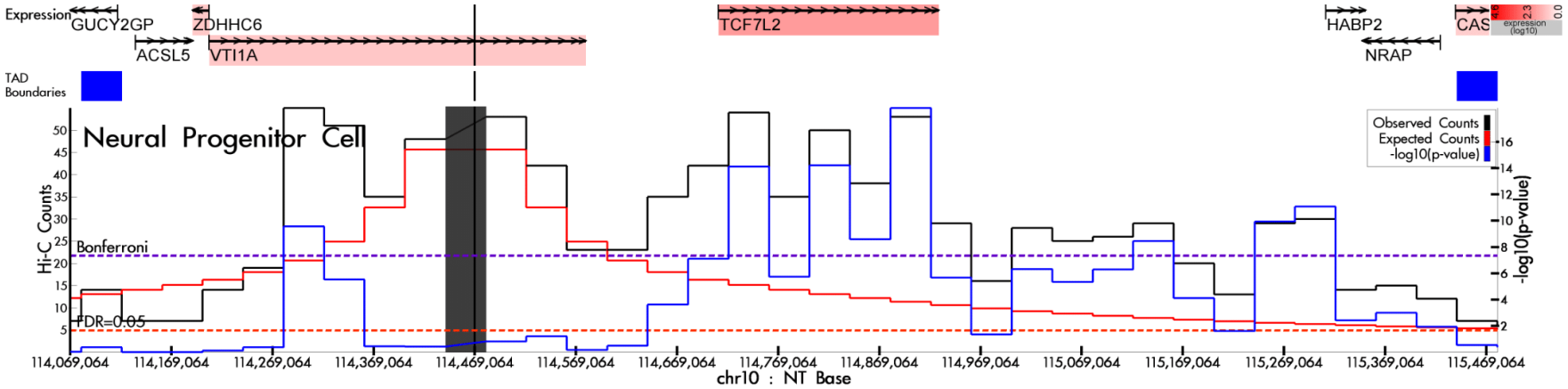
(j) 9p21.3 - rs634537



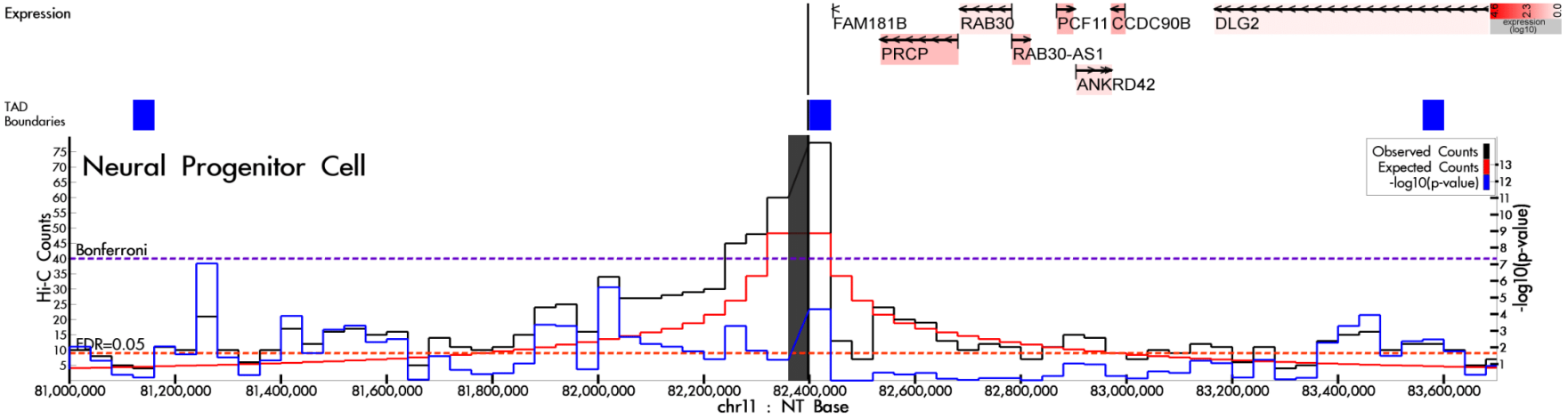
(k) 10q24.33 - rs11598018



(l) 10q25.2 - rs11196067



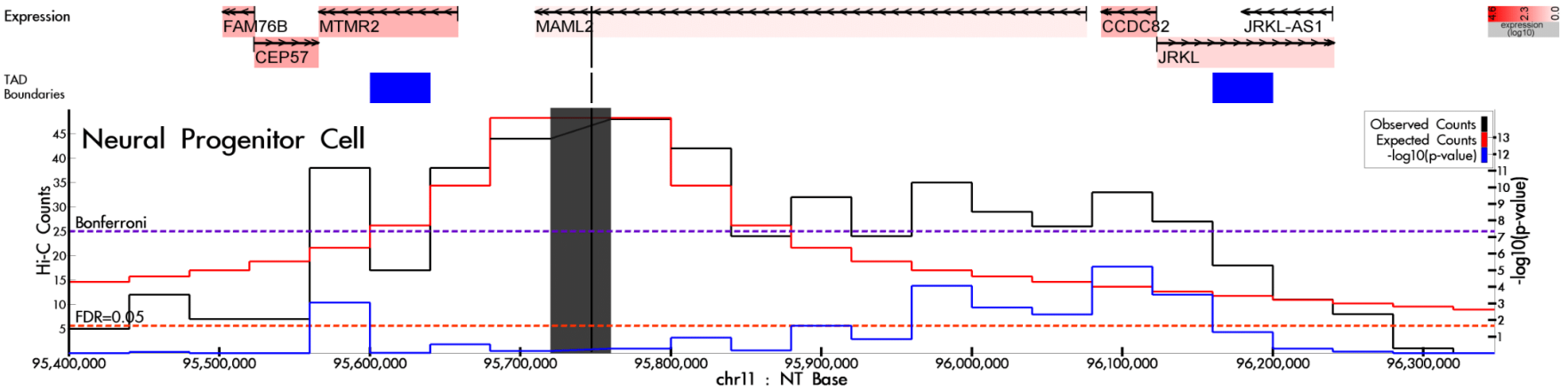
(m) 11q14.1 - rs11233250



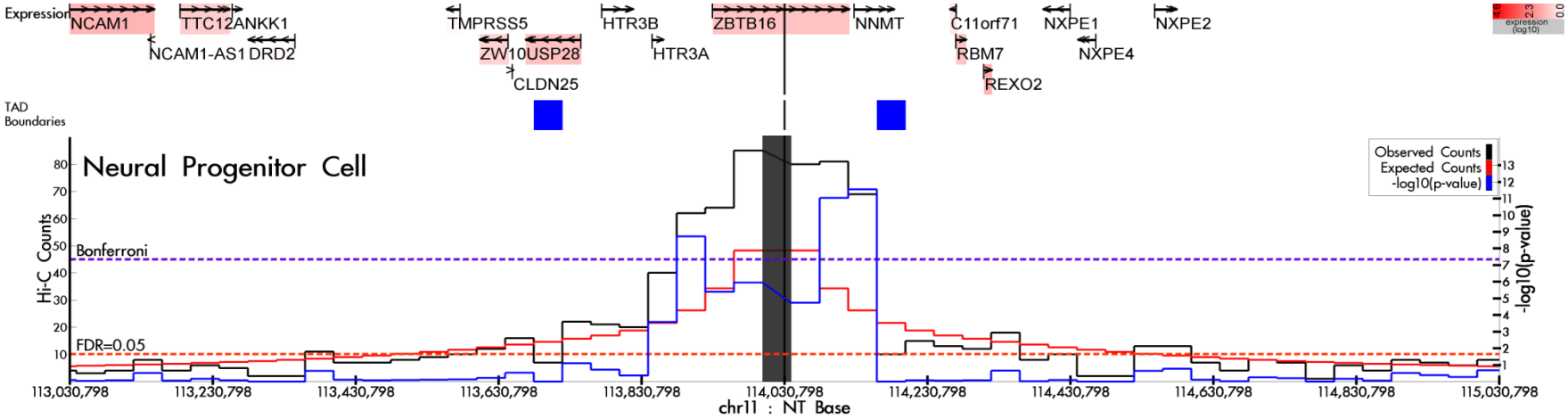
1/1

1/1

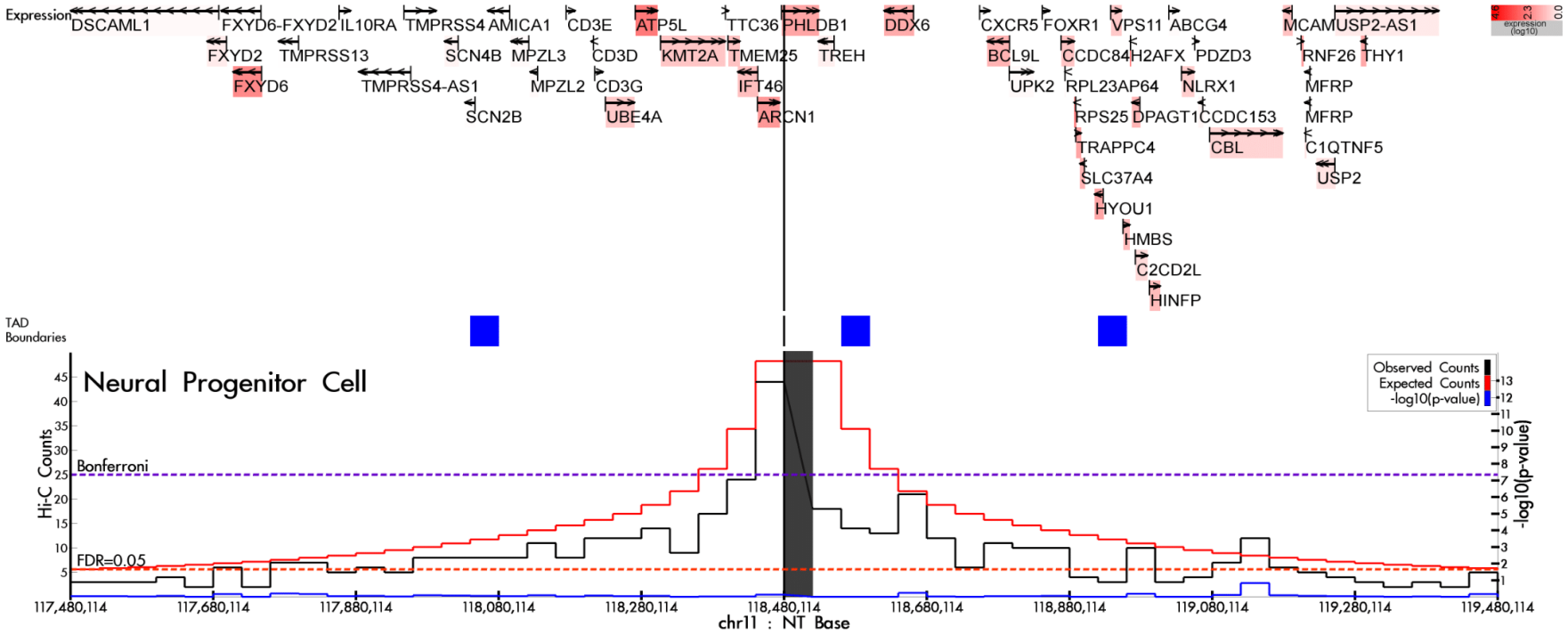
(n) 11q21 - rs7107785



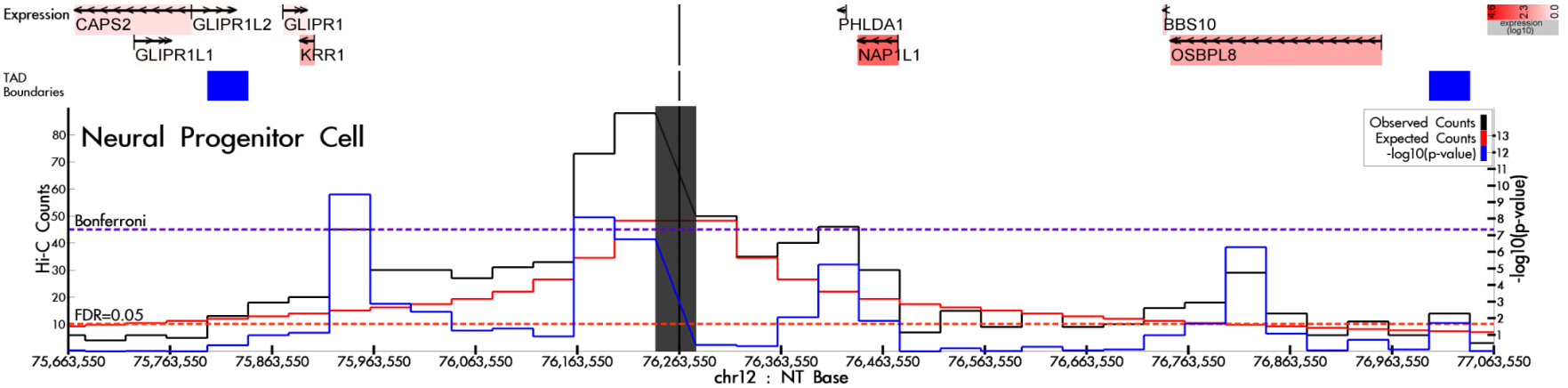
(o) 11q23.2 - rs648044



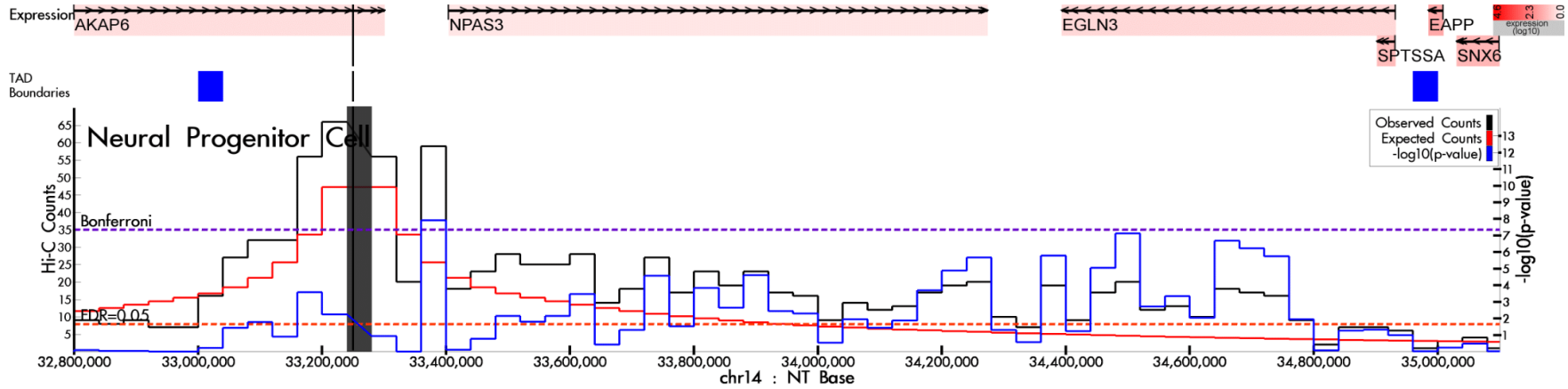
(p) 11q23.3 - rs12803321



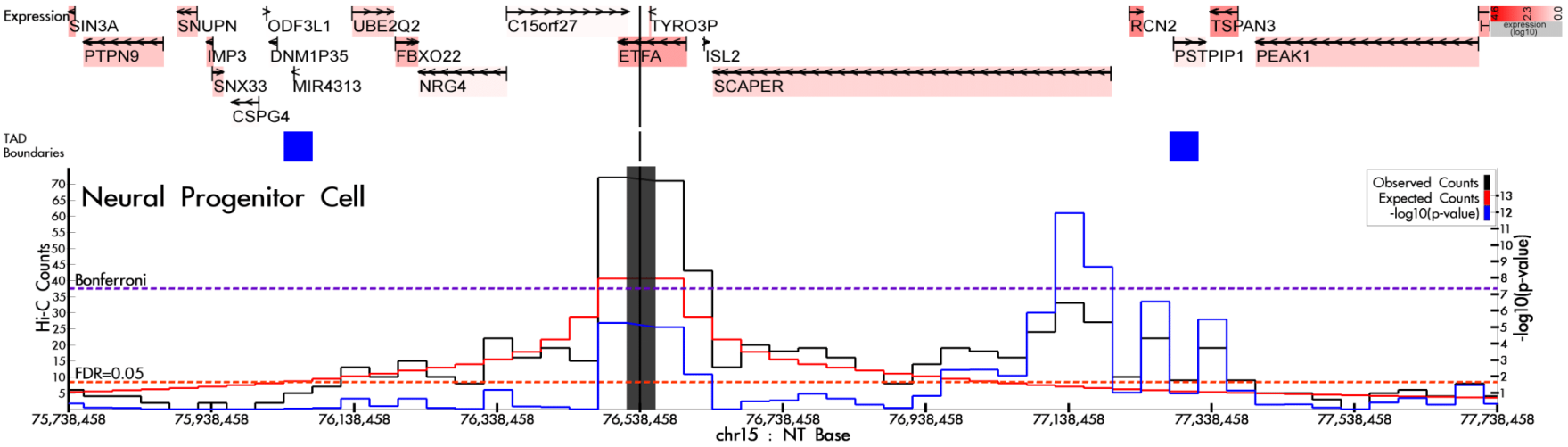
(q) 12q21.2 - rs1275600



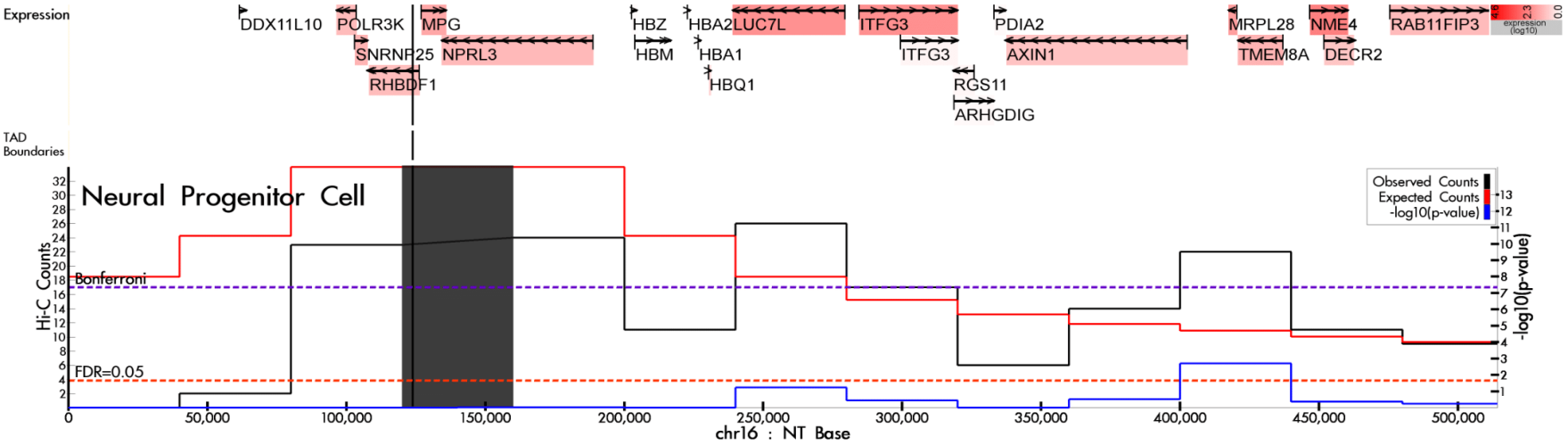
(r) 14q12 - rs10131032



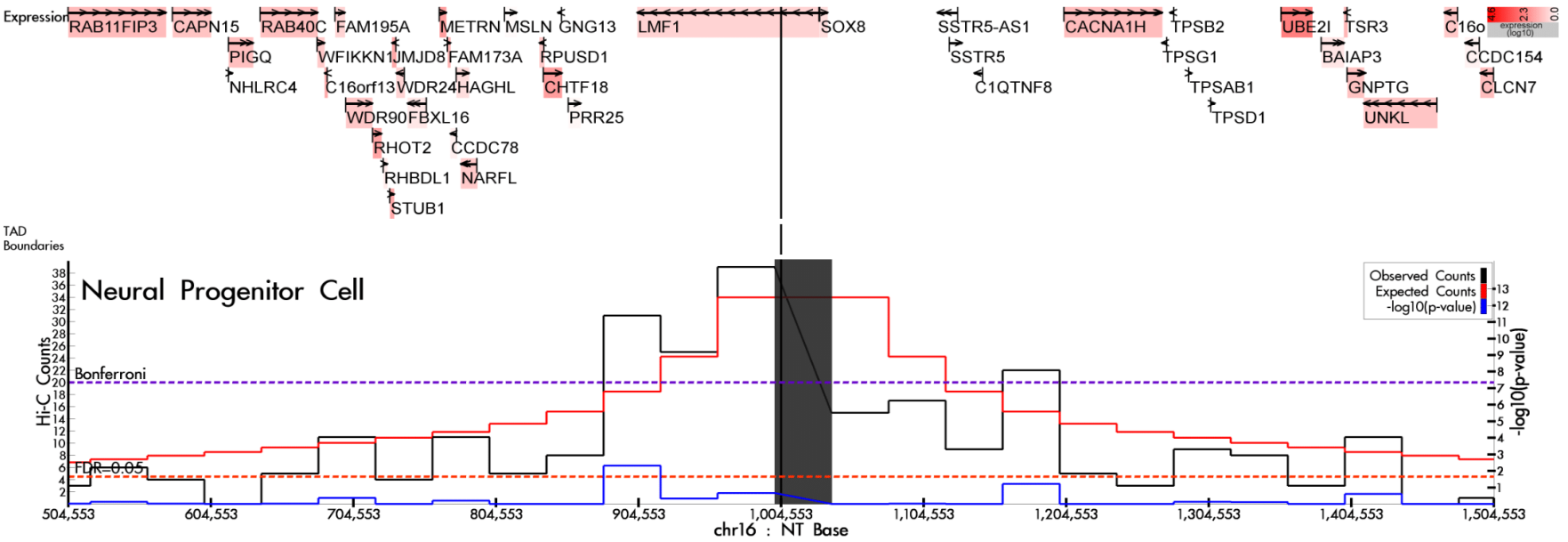
(s) 15q24.2 - rs77633900



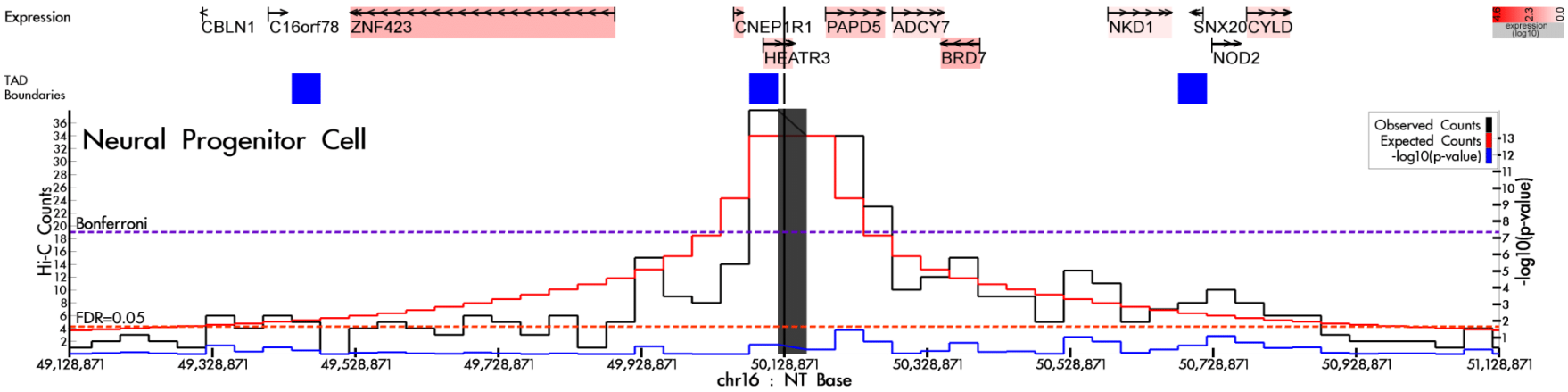
(t) 16p13.3 - rs2562152



(u) 16p13.3 - rs3751667

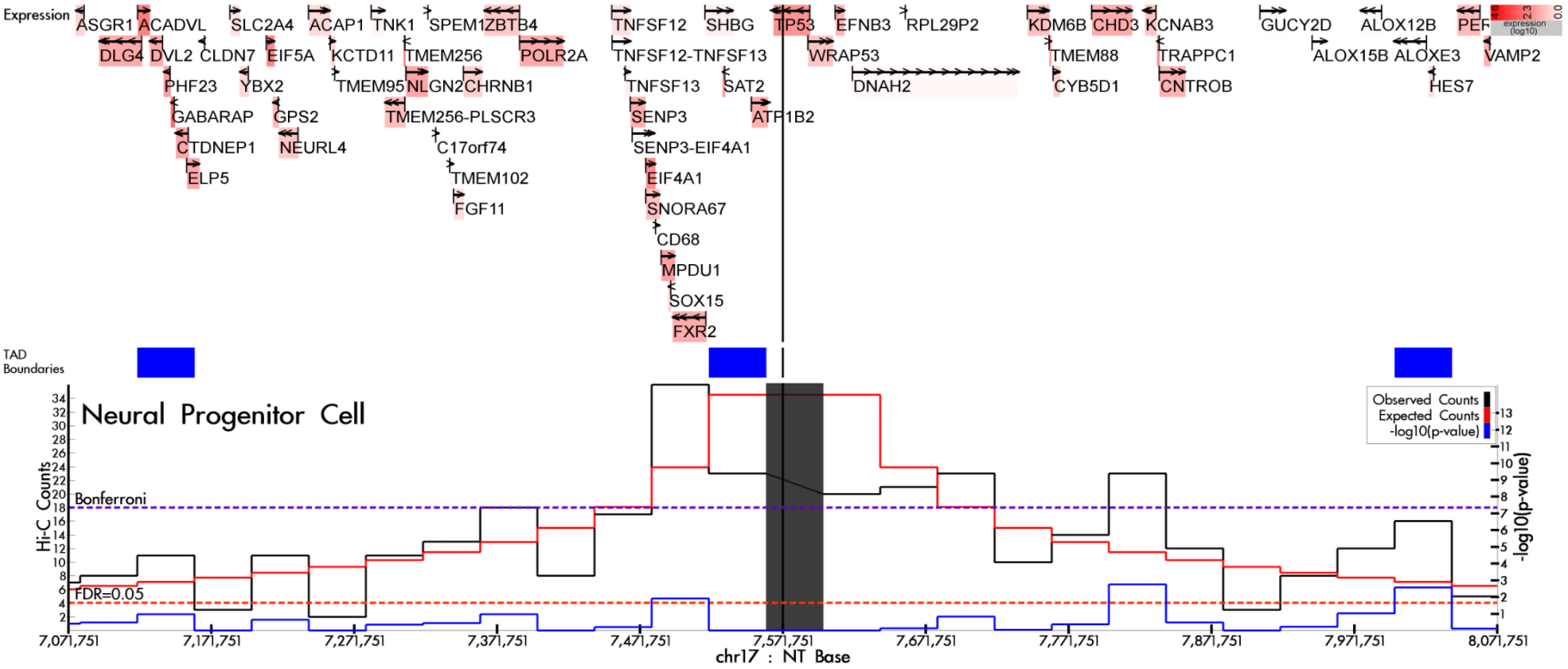


(v) 16q12.1 - rs10852606

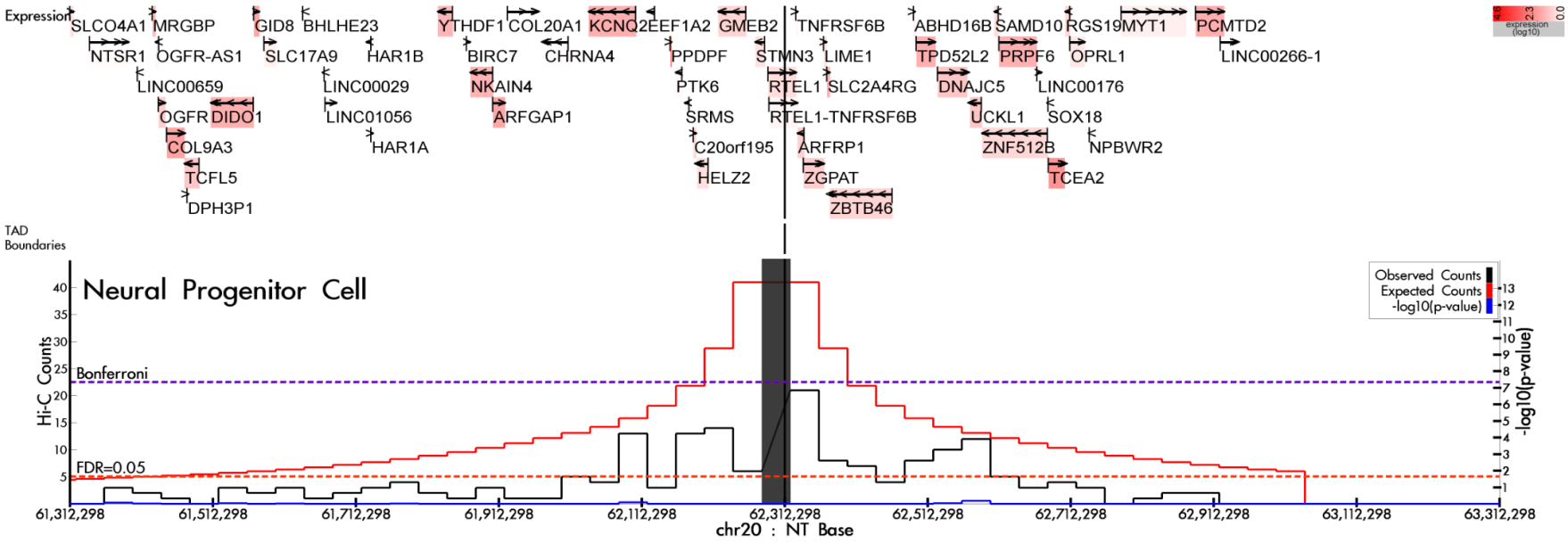


1

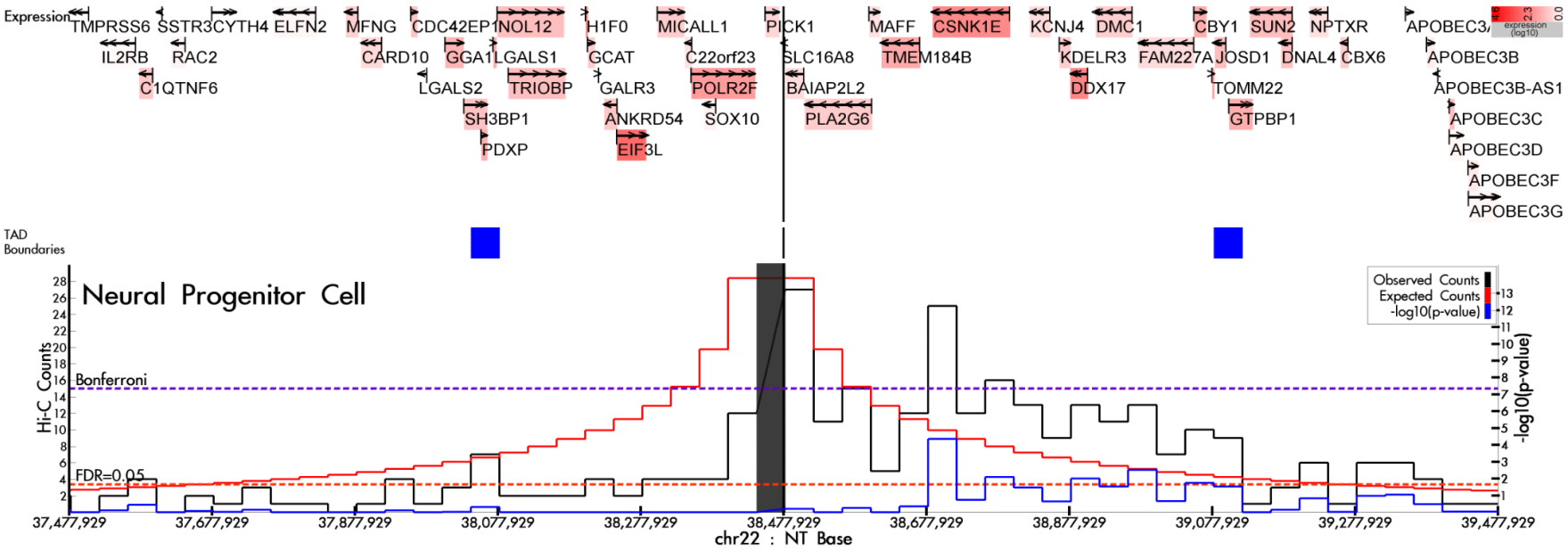
(w) 17p13.1 - rs78378222



(x) 20q13.33 - rs2297440



(y) 22q13.33 - rs2235573



Supplementary Table 2: Case distributions of glioma molecular subgroups by WHO 2016 classifications. Gliomas organised by molecular group were compared with those using the WHO 2016 classification [1]. Sample numbers with missing data are provided to aid comparison with Table 1.

	Molecular group							
WHO 2016 group	**IDH-only**	***TERT*-only**	**IDH-codel**	***TERT*-IDH**	**Triple-positive**	**Triple-negative**	**Missing**	**Total**
Astrocytoma - IDH mutated	404 (90%)	0	0	63 (81%)	0	0	65	532
Astrocytoma - IDH wild-type	0	215 (40%)	0	0	0	152 (63%)	12	379
Oligodendroglioma - 1p/19q co-deleted	0	0	29 (100%)	0	349 (100%)	0	51	429
GBM - IDH mutated	46 (10%)	0	0	15 (19%)	0	0	7	68
GBM - IDH wild-type	0	319 (59%)	0	0	0	78 (32%)	155	552
Missing	0	5 (0.9%)	0	0	0	13 (5%)	0	18
Total	**450**	**539**	**29**	**78**	**349**	**243**	**290**	**1978**

Supplementary Table 4: Linear regression analysis of age at diagnosis according to: (a) molecular group and risk allele number for the whole population; (b) molecular group and each individual SNP; (c) Number of risk alleles for IDH-only glioma; (d) Number of risk alleles for *TERT*-IDH glioma; (e) Number of risk alleles for Triple-positive glioma; (f) Number of risk alleles for Triple-negative glioma; g) number of risk alleles for *TERT*-only glioma. N indicates number of individuals included after exclusion of those with missing data (*i.e.* retaining imputed genotype probabilities > 0.9). SE, standard error.

a) Overall	**N**	**Estimate**	**SE**	**T value**	***P*-value**
Molecular group	1,646	4.87869	0.20131	24.23	$<2x10^{-16}$
Risk allele number		-0.04402	0.10473	-0.42	0.674

b) By SNP, adjusting by molecular group	**N**	**Estimate**	**SE**	**T value**	***P*-value**
1p31.3 – rs12752552	1,601	-0.1225	0.6820	-0.18	0.858
1q32.1 - rs4252707	1,645	-0.1509	0.5548	-0.272	0.786
1q44 – rs12076373	1,630	-0.04314	0.63919	-0.067	0.946
2q33.3 – rs7572263	1,645	-0.1257	0.5189	-0.242	0.809
3p14.1 – rs11706832	1,646	-0.7076	0.4634	-1.527	0.127
5p15.33 – rs10069690	1,246	0.7037	0.5619	1.252	0.211
7p11.2 – rs75061358	1,583	0.3901	0.7875	0.495	0.62
7p11.2 – rs11979158	1,630	-0.2178	0.7032	-0.31	0.757
8q24.21 – rs55705857	1,511	-2.5951	0.9013	-2.879	**0.00404**
9p21.3 – rs634537	1,638	-0.01315	0.44504	-0.03	0.976
10q24.33 – rs11598018	1,644	0.4134	0.4562	0.906	0.365
10q25.2 – rs11196067	1,538	-0.3929	0.4780	-0.822	0.411
11q14.1 – rs11233250	1,640	-0.4332	0.6985	-0.62	0.535
11q21 – rs7107785	1,639	-0.4846	0.4635	-1.046	0.296
11q23.2 – rs648044	1,387	-0.1442	0.5040	-0.286	0.775
11q23.3 – rs12803321	1,567	-0.7047	0.4299	-1.639	0.101
12q21.2 – rs1275600	1,638	-0.4568	0.4694	-0.973	0.331
14q12 – rs10131032	1,641	-0.1583	0.9506	-0.167	0.868
15q24.2 – rs77633900	1,631	0.1261	0.8233	0.153	0.878
16p13.3 – rs2562152	1,566	0.8282	0.6407	1.293	0.196
16p13.3 – rs3751667	1,583	-0.4645	0.5412	-0.858	0.391
16q12.1 – rs10852606	1,638	0.7442	0.5082	1.465	0.143
17p13.1 – rs78378222	1,623	-0.8811	1.7794	-0.495	0.621
20q13.33 – rs2297440	1,622	0.9917	0.5986	1.657	0.0978
22q13.1 – rs2235573	1,596	0.1913	0.4556	0.42	0.675

Molecular group (median age in years)

c) 1 - IDH-only (35.5)	**N**	**Estimate**	**SE**	**T value**	***P*-value**
Risk allele number	445	-0.1387	0.1628	-0.852	0.395

d) 2 - *TERT*-IDH (43.7)	**N**	**Estimate**	**SE**	**T value**	***P*-value**
Risk allele number	77	-0.1066	0.6056	-0.176	0.86077

e) 3 – Triple-positive (46.6)	**N**	**Estimate**	**SE**	**T value**	***P*-value**
Risk allele number	346	-0.04088	0.21088	-0.194	0.846

f) 4 – Triple-negative (47.6)	**N**	**Estimate**	**SE**	**T value**	***P*-value**
Risk allele number	241	0.1128	0.3589	0.314	0.754

f) 5 - *TERT*-only (59.2)	**N**	**Estimate**	**SE**	**T value**	***P*-value**
Risk allele number	537	-0.07282	0.17747	-0.41	0.682

Supplementary Table 5: Cox Proportional-Hazards analysis of overall survival according to: (a) Age, Molecular group, grade and number of risk alleles for the whole population; (b) Age, Molecular group and each individual SNP; (c) Age, grade and number of risk alleles for Triple-positive glioma; (d) Age, grade and number of risk alleles for IDH-only and *TERT*-IDH glioma; (e) Age, grade and number of risk alleles for Triple-negative glioma; (f) Age, grade and number of risk alleles for *TERT*-only glioma. Due to smaller sample number of *TERT*-IDH and similarity in survival with IDH-only tumours, these were combined into one group. N indicates number of individuals included after exclusion of those with missing data (*i.e.* retaining imputed genotype probabilities > 0.9).

a) Overall	N	Coef	Exp(Coef)	Se(Coef)	Z	*P*-value
Molecular group	1,365	0.55223	1.73713	0.04638	11.91	**$<2x10^{-16}$**
Age		0.02725	1.02762	0.00333	8.18	**$3.3x10^{-16}$**
Grade		0.50137	1.65097	0.06152	8.15	**3.3e-16**
Risk allele number		-0.04396	0.95699	0.0135	-3.24	**0.0012**

b) By SNP, adjusting by molecular group	N	Coef	Exp(Coef)	Se(Coef)	Z	*P*-value
1p31.3 – rs12752552	1,324	-0.09958	0.90522	0.08118	-1.23	0.22
1q32.1 - rs4252707	1,364	-0.18006	0.83522	0.07154	-2.52	**0.012**
1q44 – rs12076373	1,349	0.04794	1.04911	0.07913	0.61	0.54
2q33.3 – rs7572263	1,364	-0.01941	0.98078	0.06488	-0.3	0.76
3p14.1 – rs11706832	1,365	-0.06859	0.93371	0.05540	-1.24	0.22
5p15.33 – rs10069690	965	-0.0757	0.9271	0.0750	-1.01	0.31
7p11.2 – rs75061358	1,302	-0.14607	0.86410	0.09925	-1.47	0.14
7p11.2 – rs11979158	1,349	-0.06373	0.93826	0.08612	-0.74	0.46
8q24.21 – rs55705857	1,230	-0.01939	0.98080	0.14485	-0.13	0.89
9p21.3 – rs634537	1,357	-0.04579	0.95524	0.05588	-0.82	0.41
10q24.33 – rs11598018	1,363	-0.01217	0.98790	0.05883	-0.21	0.84
10q25.2 – rs11196067	1,257	-0.07980	0.92331	0.05880	-1.36	0.17
11q14.1 – rs11233250	1,359	0.04696	1.04808	0.08913	0.53	0.6
11q21 – rs7107785	1,358	-0.13690	0.87206	0.05736	-2.39	**0.017**
11q23.2 – rs648044	1,106	-0.02746	0.97291	0.06413	-0.43	0.67
11q23.3 – rs12803321	1,298	-0.13159	0.87670	0.05440	-2.42	**0.016**
12q21.2 – rs1275600	1,357	-0.09904	0.90570	0.05633	-1.76	0.079
14q12 – rs10131032	1,360	0.23582	1.26595	0.12111	1.95	0.052
15q24.2 – rs77633900	1,353	0.04699	1.04811	0.10202	0.46	0.65
16p13.3 – rs2562152	1,285	0.0512	1.0525	0.0793	0.65	0.52
16p13.3 – rs3751667	1,302	-0.00969	0.99036	0.06747	-0.14	0.89
16q12.1 – rs10852606	1,358	0.01386	1.01396	0.06354	0.22	0.83
17p13.1 – rs78378222	1,342	0.11066	1.11702	0.21050	0.53	0.6
20q13.33 – rs2297440	1,342	-0.01284	0.98725	0.07541	-0.17	0.86
22q13.1 – rs2235573	1,316	-0.1039	0.9013	0.0564	-1.84	0.066

By molecular group

c) Triple-positive	N	Coef	Exp(Coef)	Se(Coef)	Z	*P*-value
Age	254	0.0743	1.0770	0.0177	4.20	**$1.7x10^{-5}$**
Grade		-0.0091	0.9909	0.3581	-0.03	0.980
Risk allele number		-0.1103	0.8955	0.0568	-1.94	**0.05**

d) IDH-only and *TERT*-IDH	N	Coef	Exp(Coef)	Se(Coef)	Z	*P*-value
Age	436	0.02682	1.02718	0.00709	43.78	**0.00016**
Grade		0.45560	1.57711	0.11719	3.89	**0.00010**
Risk allele number		-0.01522	0.98489	0.02655	-0.57	0.56640

e) Triple-negative	N	Coef	Exp(Coef)	Se(Coef)	Z	*P*-value
Age	203	0.0254	1.0257	0.0064	3.96	**$7.3x10^{-5}$**
Grade		0.6999	2.0135	0.1266	5.53	**$3.2x10^{-8}$**
Risk allele number		-0.0702	0.9322	0.0317	-2.22	**0.036**

f) *TERT*-only	N	Coef	Exp(Coef)	Se(Coef)	Z	*P*-value
Age	472	0.02981	1.03026	0.00509	5.85	**$4.8x10^{-9}$**
Grade		0.38720	1.47285	0.08816	4.39	**$1.1x10^{-5}$**
Risk allele number		-0.03874	0.96200	0.01929	-2.01	**0.045**

APPENDIX 2

Publications:

1. **Labreche, K*.,** M. Daniau*, A. Sud, P.J. Law, L. Royer-Perron, G. Ahle, P. Soubeyran, L. Taillandier, O.L. Chinot, O. Casasnovas, JO. Bay, F. Jardin, L. Oberic, M. Fabbro, G. Damaj, A. Brion, MP. Moles-Moreau, R. Gressin, V. Delwail, F. Morschhauser, P. Agapé, A. Jaccard, H. Ghesquieres, A. Tempescul, E. Gyan, JP. Marolleau, R. Houot, L. Fornecker, A.L. Di Stefano, I. Detrait, A. Rahimian, M. Lathrop, D. Genet, F. Davi, N. Cassoux, V. Touitou, S. Choquet, A. Vital, M. Polivka, D. Figarella-Branger, K. Mokhtari, C. Philippe, M. Sanson, C. Houillier, C. Soussain, K. Hoang-Xuan, R.S. Houlston, A. Alentorn, LOC Network. "A genome-wide association study identifies susceptibility loci for primary central nervous system lymphoma at 6p25.3 and 3p22.1: a LOC network study group". ***Nature Communication***, (**2018**) in review.
2. Disney-Hogg L., A.J. Cornish, A. Sud, P.J. Law, Kinnersley B, Jacobs D.I., Ostrom QT, **Labreche K,** J.E. Eckel-Passow, G.N Armstrong, E.B. Claus, D. Il'yasova, J. Schildkraut, J.S. Barnholtz-Sloan, S.H. Olson, J.L. Bernstein, R.K. Lai, M.J. Schoemaker, M. Simon, P. Hoffmann, M.M. Nöthen, K.H. Jöckel, S. Chanock, P. Rajaraman, C. Johansen, R.B. Jenkins, B.S. Melin, M.R. Wrensch, M. Sanson, M.L. Bondy, R.S. Houlston. "Impact of atopy on risk of glioma: a Mendelian randomisation study." ***BMC Med*** 16, no. 1 (Mar 15 **2018**):42.
3. Disney-Hogg L., A.J. Cornish, A. Sud, P.J. Law, Kinnersley B, Ostrom QT, **Labreche K,** J.E. Eckel-Passow, G.N Armstrong, E.B. Claus, D. Il'yasova, J. Schildkraut, J.S. Barnholtz-Sloan, S.H. Olson, J.L. Bernstein, R.K. Lai, A.J. Swerdlow, M. Simon, P. Hoffmann, M.M. Nöthen, K.H. Jöckel, S. Chanock, P. Rajaraman, C. Johansen, R.B. Jenkins, B.S. Melin, M.R. Wrensch, M. Sanson, M.L. Bondy, R.S. Houlston. "Influence of obesity-related risk factors in the aetiology of glioma." ***Br J Cancer*** 118, no. 7 (Apr **2018**): 1020–1027.
4. **Labreche, K*.**, B. Kinnersley*, G. Berzero, A. L. Di Stefano, A. Rahimian, I. Detrait, Y. Marie, B. Grenier-Boley, K. Hoang-Xuan, J. Y. Delattre, A. Idbaih, R. S. Houlston, and M. Sanson. "Diffuse Gliomas Classified by 1p/19q Co-Deletion, Tert Promoter and Idh Mutation Status Are Associated with Specific Genetic Risk Loci." ***Acta Neuropathologica*** 135, no. 5 (Feb 19 **2018**) :743-755.
5. Bruno, A.*, **K. Labreche***, M. Daniau*, B. Boisselier, G. Gauchotte, L. Royer-Perron, A. Rahimian, F. Lemoine, P. de la Grange, J. Guegan, F. Bielle, M. Polivka, C. Adam, D. Meyronet, D. Figarella-

Branger, C. Villa, F. Chretien, S. Eimer, F. Davi, A. Rousseau, C. Houillier, C. Soussain, K. Mokhtari, K. Hoang-Xuan, and A. Alentorn. "Identification of Novel Recurrent Etv6-Igh Fusions in Primary Central Nervous System Lymphoma." ***Neuro-Oncology*** (Feb 8 **2018**).

6. Takahashi, H., A. J. Cornish, A. Sud, P. J. Law, B. Kinnersley, Q. T. Ostrom, **K. Labreche**, J. E. Eckel-Passow, G. N. Armstrong, E. B. Claus, D. Ll'yasova, J. Schildkraut, J. S. Barnholtz-Sloan, S. H. Olson, J. L. Bernstein, R. K. Lai, M. J. Schoemaker, M. Simon, P. Hoffmann, M. M. Nothen, K. H. Jockel, S. Chanock, P. Rajaraman, C. Johansen, R. B. Jenkins, B. S. Melin, M. R. Wrensch, M. Sanson, M. L. Bondy, C. Turnbull, and R. S. Houlston. "Mendelian Randomisation Study of the Relationship between Vitamin D and Risk of Glioma." ***Scientific Reports*** 8, no. 1 (Feb 5 **2018**): 2339.
7. Euskirchen, P., F. Bielle, **K. Labreche**, W. P. Kloosterman, S. Rosenberg, M. Daniau, C. Schmitt, J. Masliah-Planchon, F. Bourdeaut, C. Dehais, Y. Marie, J. Y. Delattre, and A. Idbaih. "Same-Day Genomic and Epigenomic Diagnosis of Brain Tumors Using Real-Time Nanopore Sequencing." ***Acta Neuropathologica*** 134, no. 5 (Nov **2017**): 691-703.
8. Melin, B. S*., J. S. Barnholtz-Sloan*, M. R. Wrensch*, C. Johansen, D*. Il'yasova*, B. Kinnersley*, Q. T. Ostrom, **K. Labreche**, Y. Chen, G. Armstrong, Y. Liu, J. E. Eckel-Passow, P. A. Decker, M. Labussiere, A. Idbaih, K. Hoang-Xuan, A. L. Di Stefano, K. Mokhtari, J. Y. Delattre, P. Broderick, P. Galan, K. Gousias, J. Schramm, M. J. Schoemaker, S. J. Fleming, S. Herms, S. Heilmann, M. M. Nothen, H. E. Wichmann, S. Schreiber, A. Swerdlow, M. Lathrop, M. Simon, M. Sanson, U. Andersson, P. Rajaraman, S. Chanock, M. Linet, Z. Wang, M. Yeager, Consortium GliomaScan, J. K. Wiencke, H. Hansen, L. McCoy, T. Rice, M. L. Kosel, H. Sicotte, C. I. Amos, J. L. Bernstein, F. Davis, D. Lachance, C. Lau, R. T. Merrell, J. Shildkraut, F. Ali-Osman, S. Sadetzki, M. Scheurer, S. Shete, R. K. Lai, E. B. Claus, S. H. Olson, R. B. Jenkins, R. S. Houlston, and M. L. Bondy. "Genome-Wide Association Study of Glioma Subtypes Identifies Specific Differences in Genetic Susceptibility to Glioblastoma and Non-Glioblastoma Tumors." ***Nature Genetics*** 49, no. 5 (May **2017**): 789-94.
9. Labussiere, M., A. Rahimian, M. Giry, B. Boisselier, Y. Schmitt, M. Polivka, K. Mokhtari, J. Y. Delattre, A. Idbaih, **K. Labreche**, A. Alentorn, and M. Sanson. "Chromosome 17p Homodisomy Is Associated with Better Outcome in 1p19q Non-Codeleted and Idh-Mutated Gliomas." ***Oncologist*** 21, no. 9 (Sep **2016**): 1131-5.
10. **Labreche, K*.,** I. Simeonova*, A. Kamoun*, V. Gleize*, D. Chubb, E. Letouze, Y. Riazalhosseini, S. E. Dobbins, N. Elarouci, F. Ducray, A. de Reynies, D. Zelenika, C. P. Wardell, M. Frampton, O.

Saulnier, T. Pastinen, S. Hallout, D. Figarella-Branger, C. Dehais, A. Idbaih, K. Mokhtari, J. Y. Delattre, E. Huillard, G. Mark Lathrop, M. Sanson, R. S. Houlston, and Pola Network. "TCF12 Is Mutated in Anaplastic Oligodendroglioma." ***Nature Communication*** 6 (Jun 12 **2015**): 7207.

11. Litchfield, K., B. Summersgill, S. Yost, R. Sultana, **K. Labreche**, D. Dudakia, A. Renwick, S. Seal, R. Al-Saadi, P. Broderick, N. C. Turner, R. S. Houlston, R. Huddart, J. Shipley, and C. Turnbull. "Whole-Exome Sequencing Reveals the Mutational Spectrum of Testicular Germ Cell Tumours." *Nature Communication* 6 (Jan 22 **2015**): 5973.
12. Bruno, A., B. Boisselier, **K. Labreche**, Y. Marie, M. Polivka, A. Jouvet, C. Adam, D. Figarella-Branger, C. Miquel, S. Eimer, C. Houillier, C. Soussain, K. Mokhtari, R. Daveau, and K. Hoang-Xuan. "Mutational Analysis of Primary Central Nervous System Lymphoma." ***Oncotarget*** 5, no. 13 (Jul 15 **2014**): 5065-75.

A genome-wide association study identifies susceptibility loci for primary central nervous system lymphoma at 6p25.3 and 3p22.1: a LOC network study group

Karim Labreche[1,2,*], Maïlys Daniau[2,3,*], Amit Sud[1], Philip J. Law[1], Louis Royer-Perron[2,4], Guido Ahle[5], Pierre Soubeyran[6,7], Luc Taillandier[8], Olivier L. Chinot[9,10], Olivier Casasnovas[11], Jacques-Olivier Bay[12], Fabrice Jardin[13], Lucie Oberic[14], Michel Fabbro[15], Gandhi Damaj[16], Annie Brion[17], Marie-Pierre Moles-Moreau[18], Rémy Gressin[19], Vincent Delwail[20], Franck Morschhauser[21], Philippe Agapé[22], Arnaud Jaccard[23], Hervé Ghesquieres[24], Adrian Tempescul[25], Emmanuel Gyan[26], Jean-Pierre Marolleau[27], Roch Houot[28], Luc Fornecker[29], Anna Luisa Di Stefano[4,30], Inès Detrait,[2,4], Amithys Rahimian[2,4,31], Mark Lathrop[32], Diane Genet[4], Frédéric Davi[33], Nathalie Cassoux[34], Valérie Touitou[35], Sylvain Choquet[36], Anne Vital[37], Marc Polivka[38], Dominique Figarella-Branger[9,10], Karima Mokhtari[2,4,31,39], Cathy Philippe[40], Marc Sanson[2,4,31], Caroline Houillier[2], Carole Soussain[41], Khê Hoang-Xuan[2,4,**], Richard S. Houlston[1,**], Agusti Alentorn [2,4,**,¥], LOC Network[†].

1. Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey SM2 5NG; UK
2. Inserm, U 1127, ICM, F-75013 Paris, France; CNRS, UMR 7225, ICM, F-75013 Paris, France; Institut du Cerveau et de la Moelle épinière ICM, Paris 75013, France; Sorbonne Universités, UPMC Université Paris 06, UMR S 1127, F-75013 Paris, France; AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de neurologie 2-Mazarin, 75013 Paris, France.
3. Institut du Cerveau et de la Moelle épinière, Plateforme iGenSeq, 47 Boulevard de l'Hôpital, 75013 Paris, France.
4. AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de neurologie 2-Mazarin, 75013 Paris, France. Universités, UPMC Université Paris 06, UMR S 1127, F-75013 Paris, France
5. Department of Neurology, Hôpitaux Civils de Colmar, 68024, Colmar Cedex, France.
6. Department of Medical Oncology, Institut Bergnoié, Bordeaux, F-33000, France
7. U1218 INSERM Research Unit, Bordeaux, F-33000, France.
8. Neuro-oncology Department, Nancy University Hospital and CRAN UMR 7039 CNRS, SBS BEAM Department, Nancy University, Vandoeuvre-lès-Nancy, France.
9. Department of pathology and Neuropathology, Hôpital de la Timone, Aix-Marseille Univ, AP-HM, Marseille, 13005, France.

10. AMU, CRO2, 13005, Marseille, 13005, France.
11. Deparment of Hematology, Dijon University Hospital, Dijon, 2100, Dijon, France.
12. Department of Hematology, Clermont-Ferrand University Hospital, Clermont-Ferrand, 63003, France.
13. Department of Hematology, Cancer Center Henri Henri Becquerel Center, 76000 Rouen, France and INSERM U1245, Cancer Center Henri Becquerel, Institute of Research and Innovation in Biomedicine, University of Normandy, Rouen, 76000, France.
14. Department of Hematology, IUCT – Oncopole, 31100 Toulouse – France.
15. Institut du Cancer Val d'Aurelle 34298 Montpellier Cedex 5 – France
16. Department of Hematology, University Hospital of Caen, Caen, 14033, France.
17. Department of Hematology, CHRU Besançon, Besançon, 25030, France
18. Department of Hematology, Angers University Hospital, Angers, 49033, France.
19. Department of Hematology CHU Grenoble Michallon 38043 Grenoble Cedex 02 – France
20. Service d'Oncologie Hématologique et de Thérapie Cellulaire, CHU de Poitiers, INSERM, CIC 1402, Poitiers, Centre d'Investigation Clinique, Université de Poitiers, Poitiers, France
21. Department of Hematology, CHRU Lille, Lille, 59037, France.
22. Institut de Cancérologie – 44800 Saint Herblain – France
23. Department of Hematology CHU Dupuytren 87042 Limoges- France
24. Department of Hematology, University Hospital of Lyon, 69002, Lyon, France
25. Department of Hematology CHU Morvan 29609 Brest Cedex – France
26. Department of Hematology CHU Bretonneau 34044 Tours, France
27. Department of Hematology, University Hospital of Amiens, 80054, Amiens, France.
28. CHU Rennes, Service Hématologie Clinique, F-35033 Rennes, France
29. Department of Hematology CHU Strasbourg 67000 Strasbourg – France
30. Department of Neurology Hôpital Foch 92151 Suresnes – France
31. OncoNeuroTek, Institut du Cerveau et de la Moelle épinière, ICM, Paris, F-75013, France.
32. McGill University and Genome Quebec Innovation Centre, Montreal, Quebec, Canada, H3A 0G1.
33. Department of Biological Hematology, AP-HP, Groupe Hospitalier Pitié-Salpêtrière, 75013, Paris, France.
34. Department of Oncological Ophtalmology, Institut Curie, Paris, France.
35. AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Department of Ophthalmology, Paris, 75013, France.

36. AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Department of Hematology, Paris, 75013, France
37. CNRS, Institut des Maladies Neurodégénératives, UMR 5293, F-33000 Bordeaux, France;Department of Pathology, Bordeaux University Hospital, Bordeaux, France
38. Department of Pathology, CHU Paris-GH St-Louis Lariboisière F.Widal - Hôpital Lariboisière, 75010, Paris, France.
39. AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Department of Neuropathology Raymond Escourolle, Paris, F-75013, France
40. Neurospin Centre CEA, Saclay 91191, Gif sur Yvette, France.
41. Department of Hematology, Hôpital René Huguenin, Institut Curie, 92210, Saint-Cloud, France

* These authors contributed equally to this work

** These authors jointly supervised this work

¥ Corresponding author: Agusti Alentorn; Tel: 00 33 1 42 16 41 60; email: agusti.alentorn@aphp.fr

ABSTRACT

Primary central nervous system lymphoma (PCNSL) is a rare form of extra-nodal non-Hodgkin lymphoma. Here we performed a meta-analysis of two new genome-wide association studies of PCNSL totaling 475 cases and 1,134 controls of European ancestry. We identified independent risk loci at 3p22.1 (rs41289586, *ANO10, P* = 2.17 x 10^{-8}) and 6p25.3 near *EXOC2* (rs116446171, *P*= 1.95 x 10^{-13}). These data provide the first evidence for inherited predisposition to PCNSL.

INTRODUCTION

Primary diffuse large B-cell lymphoma of the central nervous system (PCNSL) is a rare tumor that accounts for ≤1% of all lymphomas, and approximately 2% of all primary CNS tumors[1]. The WHO classification of tumors of hematopoietic and lymphoid tissues recognizes PCNSL as a distinct subtype of non-Hodgkin lymphoma (NHL)[2], with over 95% of tumors belonging to the diffuse large B-cell lymphoma (DLBCL) group[3].

Immunocompromised individuals are considered most at risk of PCNSL, however, the incidence of the disease is increasing in the immunocompetent populations who represent today the vast majority of the patients[4-6]. The disease typically follows an aggressive course and despite advances in the treatment of PCNSL is still associated with very high mortality[3].

Although PCNSL is strongly linked to Epstein-Barr virus (EBV) infection in immunocompromised patients, its detection is virtually absent in PCNSL from immunocompetent patients and little else is known about its etiology and risk factors in the population[7]. To address the possibility that common genetic variants influence the risk of developing PCNSL, we have conducted a genome-wide association study (GWAS) on immunocompetent patients. Specifically, we performed a meta-analysis of two new GWAS of PCNSL and identify independent single nucleotide polymorphisms (SNPs) at 3p22.1 and 6p25.3 associated with risk.

RESULTS

Association analysis

After quality control, the two GWAS provided SNP genotypes on a total of 475 cases and 1,134 controls (**Supplementary Fig. 1** and **2 - Supplementary Tables 1** and **2**). To increase genomic resolution, we imputed >10 million SNPs using the 1000 Genomes Project[8] combined with UK10K[9] as reference. Quantile-Quantile (Q-Q) plots for SNPs with minor allele frequency (MAF) >0.5% post imputation showed only minimal evidence of over-dispersion (λ values for both GWAS = 1.0; **Supplementary Fig. 3**). Meta-analyzing test results from the two GWAS, we derived joint odds ratios (OR) per-allele and 95% confidence intervals (CI) under a fixed-effects model for each SNP and associated *P*-values.

Genome-wide significant associations (*i.e.* $P<5.0 \times 10^{-8}$) were shown for loci at 3p22.1 (rs41289586, $P=2.17 \times 10^{-8}$) and 6p25.3 (rs116446171, $P=1.95 \times 10^{-13}$) (**Fig. 1, Table 1**). Conditional analysis of GWAS data showed no evidence for additional independent signals at either of the two risk loci.

The 6p25.3 risk SNP rs116446171 (**Fig. 2**), which maps intragenic to *EXOC2* (exocyst complex component 2) and *IRF4* (interferon regulatory factor 4), has been previously been shown to influence the risk of DLBCL[10]. EXOC2 is part of the multi-protein exocyst complex essential for polarized vesicle trafficking and the maintenance and intercellular transfer of viral proteins and virions[11]. Thus far there is no evidence to implicate *EXOC2* in lymphoma. In contrast *IRF4* has a well-established role in the development of most B-cell malignancies[12-14]. The 3p22.1 risk SNP rs41289586 (**Fig. 2**) localizes to exon 6 of the anoctamin 10 gene (*ANO10*) and is responsible for the rare missense change (*ANO10*:c.788G>A, p.Arg263His). Defects in *ANO10,* which encodes a calcium-activated chloride channel transmembrane protein are a cause autosomal recessive spinocerebellar ataxia[15]. To date there is no evidence for the role of *ANO10* in any B-cell malignancy.

In addition to the 6p25.3 and 3p22.1 risk loci we identified promising associations ($P<2.0 \times 10^{-7}$), at 6q15 (rs10806425, $P=1.36 \times 10^{-7}$) and 8q24.21 (rs13254990; $P=1.33 \times 10^{-7}$) annotating genes with strong relevance to B-cell tumorigenesis (**Table 1, Supplementary Fig. 4**). rs10806425 localizes to

intron 1 of the gene encoding *BACH2* (basic leucine zipper transcription factor 2). Loss of heterozygosity of *BACH2* has been reported at a frequency of 20% in B-cell lymphoma[16]. In DLBCL patients with higher *BACH2* expression tend to have a better prognosis[17]. *BACH2* is a key regulator of the pre-BCR check point as well as a tumor suppressor in pre-B acute lymphoblastic leukemia[18]. One mechanism of *BACH2* downregulation in leukemias is the loss of the transcription factor *PAX5*, which is intriguingly, commonly mutated in both PCNSL[19] and B-cell ALL[18].

The 8q24 SNP rs13254990 localizes to intron 4 of *PVT1,* a non-coding RNA affecting the activation of MYC. Two independent risk loci at 8q24 defined by SNPs rs13255592 and rs4733601 have previously been shown to influence DLBCL[10]. rs13255592 which also localizes within intron 4 of *PVT1* and is highly correlated with rs13254990 (r^2=0.98, P=3.81 × 10^{-7}). No association between rs4733601, which maps approximately 1.9Mb telomeric to *PVT1,* and PCNSL risk was shown (P=0.99, r^2=4.21 × 10^{-5}; **Supplementary Table 3**). The 8q24.21 128-130Mb genomic interval harbors multiple independent risk loci with different tumor specificities (**Supplementary Table 4**)[10,20-29]. The strongest additional association for PCNSL being shown by the Hodgkin lymphoma risk SNP rs2019960 (P=4.1 × 10^{-5}) raising the possibility of an additional risk locus for the disease at 8q24.21[30].

Following on from this we examined to see if the other reported risk loci for DLBCL influenced PCNSL risk. Respective association *P*-values for the 6p21.22-HLA (rs2523607) and 2p23.3 (rs79480871) risk SNPs were 0.023 and 0.14 (**Supplementary Table 3).**

HLA alleles

Variation at HLA has been linked to risk of DLCBL and a number of other B-cell tumors[10,22,30-32]. The strongest SNP association at 6p21 (HLA) for PCNSL was provided by rs2395192 (P=1.81 × 10^{-7}), which maps between *HLA-DRA* and *HLA-DRB5* (**Supplementary Fig.5, Table 1**). To obtain additional insight into plausible functional variants within the HLA region, we imputed the classical HLA alleles and amino acid residues using SNP2HLA[33]. No imputed HLA alleles or amino acid positions reached genome-wide significance (**Supplementary Fig. 5**). The strongest coding changes within the HLA region were observed for the HLA class II alleles DRB1 Ser11Pro (AA_DRB1_11_32660115_SP, P=3.35 × 10^{-6}) and presence of the haplotype SRG (DRB1_13_32660109_SRG*,* P=3.35 × 10^{-6}) (**Supplementary Table 5**).

Functional annotation of risk loci

To gain insight into the biological basis underlying associations at 6p25.3 and promising risk loci the novel association signals, we first evaluated each of the risk SNPs as well as the correlated variants use of the online resources HaploRegv4[34], RegulomeDB[35] and Fantom5[36] for evidence of functional effects (**Supplementary Data 1**). These data revealed regions of active chromatin state at 6p25.3, 6q15 and 8q24 risk loci in B-cells. To explore whether there was an association between SNP genotype and transcript levels we performed an expression quantitative trait loci (eQTL) analysis using from the Genotype-Tissue Expression (GTEx) project[37], MuTHR[38] and blood eQTL data from Westra *et al*[39]. We used summary-level Mendelian randomization[40] (SMR) analysis to test for a concordance between signals from GWAS and *cis* eQTL for genes within 1 Mb of the sentinel and correlated SNPs (r^2>0.8) at each locus (**Supplementary Data 2**) and derived b_{XY} statistics, which estimate the effect of gene expression on PCNSL risk. After accounting for multiple testing we were unable to demonstrate any consistently significant eQTL for any of the risk loci examined. Chromatin looping interactions formed between enhancer elements and the genes that they regulate map within distinct chromosomal topological associating domains. To identify patterns of local chromatin patterns, we analyzed promoter capture Hi-C data on the LCL cell line GM12878 as a source of B-cell information[41]. Looping chromatin interactions were shown between non-coding regions at 6p25.3 (rs11646171) with the *IRF4* promoter (**Fig. 2**) and at 8q24.21 (rs13254990) with the MYC promoter; both genes with strong relevance to B-cell tumorigenesis.

Using ChIP-seq data on 82 transcription factors (TFs) in GM12878 we examined for an over-representation of the binding of TFs at risk loci. Although not statistically significant the strongest TF bindings were shown for *TBL1XR1* that is mutated in 20% of PCNSL[42] (**Supplementary Fig. 6**).

DISCUSSION

To our knowledge this is the first study providing evidence for a genetic predisposition to PCNSL. While PCNSL is a specific entity it corresponds pathologically to diffuse large B-cell lymphoma. Hence, it is therefore perhaps not surprising that we identified associations at 6q25.3 and 8q24.21 for PCNSL, which were previously reported for DLBCL. However, the absence of associations at the 8q24.21 (rs4733601) and 2p23.3 (rs79480871) risk loci strongly suggests a distinct developmental pathway for PCNSL, presumably reflective of its etiology.

Although in part speculative, the 6q25.3 association implicates *IRF4* in the development of PCNSL. Through interaction with transcription factors including PU.1, IRF4 controls the termination of pre-B-cell receptor signaling and promotes the differentiation of pro-B cells to small B cells[43]. Furthermore, via BLIMP1 and BCL6, IRF4 controls the transition of memory B cells[44]. The observation that *PVT1* rearrangement occurs frequently in highly aggressive B-cell lymphomas harboring an 8q24 abnormality makes it entirely plausible that germline variation in this region influences PCNSL risk[45-47]. The 6q15 association implicates *BACH2* in the development of PCNSL. *BACH2* is an attractive candidate *a priori* for having a role in PCNSL development being regulator of the antibody response mediating effects through *BLIMP1*, *XBP1*, *LRF4*, and *PAX5*[48]. Moreover, it is partly mediates the tumor suppressor activity of c-Rel in lymphoma development[49]. Collectively these data are consistent with aberrant B-cell developmental pathways being central for predisposition to PCNSL.

While not statistically significant the *HLA-DRA* and *HLA-DRB1* associations are intriguing as these alleles have previously been shown to influence the human reaction to viral load and EBV infection respectively[50]. Their link to the development of PCNSL is entirely consistent with an infective basis to this B-cell malignancy even though all of the patients we have analyzed are not immunocompromised.

In summary, our findings represent an important step in defining the contribution of common genetic variation to the risk of developing PCNSL. Our observations are notable since the associations highlighted define regions of the genome harboring plausible candidate genes for further investigation. Given the relatively modest size of our analysis, inevitably constrained by the rarity of PCNSL, it is highly probable that further studies will discover additional common

susceptibility loci. These coupled with functional analyses should provide for an explanation of the biological underpinnings of PCNSL.

METHODS

Subjects and ethics

This study was based on two primary GWAS datasets: (1) GWAS-1 comprised 346 immunocompetent HIV negative patients (184 male; median age 68 years) with PCNSL ascertained through the Service de Neurologie Mazarin, Groupe Hospitalier Pitié-Salpêtrière Paris and the Lymphome oculo-cerebral network (LOC) between 2008-2017. For controls we made use of Illumina HumanHap 660 data 788 individuals from the SU.VI.MAX (SUpplementation en VItamines et MineralsAntioXydants) study healthy subjects (women aged 35–60 years; men aged 45–60 years). (2) GWAS-2 comprised 129 immunocompetent HIV negative patients (76 male; median age 69 years) with primary DLBCL CNS tumors ascertained through the Service de Neurologie Mazarin, Groupe Hospitalier Pitié-Salpêtrière Paris and LOC 2001-2007. For controls, we made use of second series of Illumina HumanHap 660 data generated on 346 individuals from the SU.VI.MAX. Collection of patient samples and associated clinico-pathological information was undertaken with written informed consent and ethical review board approval in accordance with the tenets of the declaration of Helsinki. The diagnosis of PCNSL (ICD-10 C83.3; WHO 9690/3) was established in accordance with WHO guidelines.

Genotyping and quality control

Constitutional DNA was extracted from blood samples using QIAamp DNA Blood Mini Kit (Qiagen) (OncoNeuroTek, Paris). The quality of extracted DNA was analyzed on a Caliper LabchipGX and Nanodrop. DNA samples were prepared according to Qubit quantification. Cases were genotyped using the Infinium OmniExpress-24 v1.2 BeadChip array according to the manufacturer's recommendations (Illumina Inc, San Diego, CA, USA). Standard quality control measures were applied to the GWAS[51]. Specifically, individuals with low call rate (<90%) as well as all individuals with non-European ancestry (using the HapMap version 2 CEU, JPT/CHB and YRI populations as a reference) were excluded. SNPs with a call rate <90% were excluded as were those with a MAF < 0.01 or displaying significant deviation from Hardy-Weinberg equilibrium (*i.e.* $P<10^{-6}$). GWAS data were imputed to >10 million SNPs with IMPUTE2 v2.3[52] software using a merged reference panel consisting of data from 1000 Genomes Project (phase 1 integrated release 3, March 2012)[8] and UK10K[9]. Genotypes were aligned to the positive strand in both imputation and genotyping. Imputation was conducted separately for each GWAS, and in each, the data were pruned to a common set of SNPs between cases and controls before imputation. Poorly imputed SNPs defined

by an information measure <0.80 were excluded. Tests of association between imputed SNPs and *P*-values were calculated using logistic regression under an additive genetic model in SNPTESTv2.5[53]. The adequacy of the case-control matching and possibility of differential genotyping of cases and controls were evaluated using Q-Q plots of test statistics (**Supplementary Fig. 1**). The fidelity of rs41289586 imputation was confirmed by the finding of 99% concordance between imputed and directly sequenced genotypes in a subset of 345 samples (31 heterozygous) (Pearson correlation coefficient, *r*=0.99).

HLA imputation and analysis

To examine if specific coding variants within HLA genes contributed to the association signals, we imputed the classical HLA alleles (A, B, C, DQA1, DQB1, DRB1) and coding variants across the HLA region (chr6:29–34 Mb) using SNP2HLA[33]- http://www.broadinstitute.org/mpg/snp2hla/. Imputation was based on a reference panel from the Type 1 Diabetes Genetics Consortium (T1DGC) which comprises genotype data from 5,225 individuals of European descent typed for HLA-A, B, C, DRB1, DQA1, DQB1, DPB1, DPA1 4-digit alleles. A total of 8,961 classical HLA alleles (two- and four-digit resolution) and 1,873 AA markers including 580 AA positions that were 'multi-allelic', were successfully imputed (info score >0.8 for variant). Multi-allelic markers were analyzed as binary markers and a meta-analysis was conducted where we tested SNPs, HLA alleles and AAs across the HLA region for association with PCNSL using SNPTEST.

Meta-analysis

Meta-analyses were performed using the fixed-effects inverse-variance method based on the β estimates and standard errors from each study using META v1.6[54]. Cochran's Q-statistic to test for heterogeneity, and the I^2 statistic to quantify the proportion of the total variation due to heterogeneity were calculated[55].

eQTL analysis

To examine the relationship between SNP genotype and gene expression we carried out Summary-data-based Mendelian Randomization (SMR) analysis as per Zhu *et al*., 2016 (http://cnsgenomics.com/software/smr/index.html)[40]. We used publicly available lymphoblastoid cell line data from the GTEx[37] (http://www.gtexportal.org) v6p release and MuTHR[38]. Briefly, GWAS summary statistics files were generated from the meta-analysis. Reference files were generated from merging 1000 genomes phase 3 and UK10K (ALSPAC and TwinsUK) vcfs. Results

from the SMR test for each of the five risk loci are reported in **Supplementary Data 2**. As previously advocated only probes with at least one eQTL P-value of $<5.0 \times 10^{-8}$ were considered for SMR analysis. We set a threshold for the SMR test of PSMR$<7.57 \times 10^{-4}$ and PSMR$<2.5 \times 10^{-3}$ corresponding to a Bonferroni correction for 66 tests (66 probes with a top eQTL $P<5.0 \times 10^{-8}$ across the 5 loci and two LCL eQTL dataset) and 20 tests (20 probes with a top eQTL $P<5.0 \times 10^{-8}$ across the 5 loci and Muther eQTL dataset) respectively.

Functional annotation

Novel risk SNPs and their proxies (*i.e.* $r^2>0.2$ in the 1000 Genomes EUR reference panel) were annotated for putative functional effect based upon histone mark ChIP-seq/ChIPmentation data for H3K27ac, H3K4Me1 and H3K27Me3 from GM12878 (LCL)[56] and primary B-cells[57]. We searched for overlap with "super-enhancer" regions as defined by Hnisz *et al*[58], restricting the analysis to the GM12878 cell line and $CD19^+$ B-cells. The novel risk SNPs and their proxies ($r^2>0.2$ as above) were intersected with regions of accessible chromatin in CLL cells, as defined by Rendeiro *et al*[57], which were used as a surrogate for likely sites of TF binding. SNPs falling within accessible sites (n=47) were taken forward to TF binding motif analysis and were also annotated for genomic evolutionary rate profiling (GERP) score[59] as well as bound TFs based on ENCODE project[56] ChIP-seq data.

Transcription factor binding disruption analysis

To examine enrichment in specific TF binding across risk loci, we adapted the variant set enrichment method of Cowper-Sal lari et al[60]. Briefly, for each risk locus, a region of strong LD (defined as $r^2>0.8$ and D'>0.8) was determined, and these SNPs were termed the associated variant set (AVS). TF ChIP-seq uniform peak data were obtained from ENCODE for the GM12878 cell line, which included data for 82 TF. For each of these marks, the overlap of the SNPs in the AVS and the binding sites was determined to produce a mapping tally. A null distribution was produced by randomly selecting SNPs with the same characteristics as the risk-associated SNPs, and the null mapping tally calculated. This process was repeated 10,000 times, and approximate *P*-values were calculated as the proportion of permutations where the null mapping tally was greater or equal to the AVS mapping tally. An enrichment score was calculated by normalizing the tallies to the median of the null distribution. Thus, the enrichment score is the number of s.d.'s of the AVS mapping tally from the mean of the null distribution tallies.

DATA AVAILABILITY

Genotype data that support the findings of this study have been deposited in the database of the European Genome-phenome Archive (EGA) with accessions codes PRJEB21814. The remaining data are contained within the paper and Supplementary files are available from the authors upon request.

ACKNOWLEDGEMENTS

The primary source of funding was provided by la Ligue Nationale Contre le Cancer-RE 2015 (K.H.X), French National Institute of Cancer (InCa) LOC Network (K.H.X and C.S). A.A has also been granted with a “poste d’accueil AP-HP-CEA”. K.L is supported by l’Association pour la Recherche sur les Tumeurs Cérébrales (ARTC) and Institute CARNOT – Institut du Cerveau et de la Moelle Epinière (ICM). Finally, also acknowledge support from Cancer Research UK (C1298/A8362). A.S. is supported by a clinical fellowship from Cancer Research UK. We are grateful to all investigators and all the patients and individuals for their participation. Samples from AP-HM were retrieved from AP-HM tumor bank, authorization number 2013-1786. We also thank the clinicians, other hospital staff and study staff that contributed to the blood sample and data collection for this study and OncoNeuroTek that provided and prepared DNA samples.

AUTHOR CONTRIBUTIONS

A.A. and K.H.X., developed the project and provided overall project management; K.L., M.D., R.S.H., K.H.X. and A.A. drafted the manuscript. K.L., A.S., performed bioinformatic and statistical analyses; Patient samples and phenotype data were provided by C.D., D.G., K.H.X, C.S. and other members of the LOC Network. M.D., I.D., L.R.P., A.R., D.G. performed project management and supervised genotyping; M.D., I.D. and A.R. performed sequencing and genotyping. A.A., K.H.X., M.D., A.R., L.R.P. supervised laboratory management and oversaw genotyping of cases; D.G., M.D., I.D., L.R.P. performed sample management of cases. All authors reviewed and approved the manuscript prior to submission.

COMPETING INTERESTS STATEMENT

The remaining authors declare no competing financial interests.

FIGURE AND TABLE LEGENDS

Figure 1: Manhattan plot of association *P*-values. Shown are the genome-wide –log10P-values (two-sided) of >10 million successfully imputed autosomal SNPs in 475 cases and 1,134 controls. The red horizontal line represents the genome-wide significance threshold of $P=5.0 \times 10-8$.

Figure 2: Regional plots of association results and recombination rates for new risk loci for primary cerebral nervous system lymphoma. Results shown for (a) 6p25 and (b) 3q21. Plots (drawn using visPig[61]) show association results of both genotyped (triangles) and imputed (circles) SNPs in the GWAS samples and recombination rates. $-\log_{10}P$ values (*y* axes) of the SNPs are shown according to their chromosomal positions (*x* axes). The sentinel SNP in each combined analysis is shown as a large circle or triangle and is labelled by its rsID. The color intensity of each symbol reflects the extent of LD with the top genotyped SNP, white ($r^2=0$) through to dark red ($r^2=1.0$). Genetic recombination rates, estimated using 1000 Genomes Project samples, are shown with a light blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the chromatin-state segmentation track (ChromHMM) for lymphoblastoid cells using data from the HapMap ENCODE Project, and the positions of genes and transcripts mapping to the region of association.

Table 1: Summary results for risk SNPs

Table 1: Summary results for SNPs associated with CNS Lymphoma risk

Locus	Nearest gene(s)	SNP	Position (bp, hg19)	Risk allele	Dataset	RAF (case;control)	OR	95% CI	*P*-value
6p25.3	*EXOC2*	rs116446171	484,453	G	GWAS-1	(0.066; 0.022)	4.11	(2.47- 6.85)	$5.13x10^{-8}$
					GWAS-2	(0.088;0.019)	7.87	(3.59 - 17.21)	$2.36x10^{-7}$
					Combined		4.99	(3.26 - 7.65)	$1.53x10^{-13}$
								I^2=46%	P_{het}=0.17
3p22.1	*ANO10*	rs41289586	43,618,558	T	GWAS-1	(0.048;0.017)	3.42	(1.94 - 6.02)	$1.90x10^{-5}$
					GWAS-2	(0.065;0.019)	4.84	(2.10 - 11.13)	$2.05x10^{-4}$
					Combined		3.82	(2.39 - 6.09)	$1.87x10^{-8}$
								I^2=0%	P_{het}=0.50
8q24.21	*PTV1*	rs13254990	129,076,451	T	GWAS-1	(0.43;0.33)	1.58	(1.31 - 1.91)	$2.21x10^{-6}$
					GWAS-2	(0.40;0.32)	1.44	(1.05 - 1.96)	0.021
					Combined		1.54	(1.31 - 1.81)	$1.33x10^{-7}$
								I^2=0%	P_{het}=0.60
6q15	*BACH2*	rs10806425	90,926,612	C	GWAS-1	(0.68;0.58)	1.50	(1.25 - 1.80)	$8.93x10^{-6}$
					GWAS-2	(0.69;0.59)	1.53	(1.14 - 2.05)	0.0045
					Combined		1.51	(1.30 - 1.77)	$1.36x10^{-7}$
								I^2=0%	P_{het}=0.93
6p21.32	*HLA-DRA*	rs2395192	32,447,644	C	GWAS-1	(0.48;0.59)	1.56	(1.30 - 1.88)	$1.65x10^{-6}$
					GWAS-2	(0.52;0.60)	1.38	(1.03 - 1.84)	0.029
					Combined		1.51	(1.29 - 1.76)	$1.81x10^{-7}$
								I^2=0%	P_{het}=0.47

bp, base pair position; OR, odds ratio; 95% CI, 95% confidence interval; P_{het}, *P*-value for heterogeneity; I^2, proportion of the total variation due to heterogeneity.
RAF is risk allele frequency across all of the GWAS-1 and GWAS-2 datasets, respectively. Odds ratios are derived with respect to the risk allele.

REFERENCES

1. PM Kluin, M.D., JA Ferry. Primary diffuse large B-cell lymphoma of the CNS. *IARC Press, Lyon*, 240-241 (2008).
2. Swerdlow, S.H. *et al.* The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood* **127**, 2375-90 (2016).
3. Hoang-Xuan, K. *et al.* Diagnosis and treatment of primary CNS lymphoma in immunocompetent patients: guidelines from the European Association for Neuro-Oncology. *Lancet Oncol* **16**, e322-32 (2015).
4. Bessell, E.M., Dickinson, P., Dickinson, S. & Salmon, J. Increasing age at diagnosis and worsening renal function in patients with primary central nervous system lymphoma. *J Neurooncol* **104**, 191-3 (2011).
5. Villano, J.L., Koshy, M., Shaikh, H., Dolecek, T.A. & McCarthy, B.J. Age, gender, and racial differences in incidence and survival in primary CNS lymphoma. *Br J Cancer* **105**, 1414-8 (2011).
6. O'Neill, B.P., Decker, P.A., Tieu, C. & Cerhan, J.R. The changing incidence of primary central nervous system lymphoma is driven primarily by the changing incidence in young and middle-aged men and differs from time trends in systemic diffuse large B-cell non-Hodgkin's lymphoma. *Am J Hematol* **88**, 997-1000 (2013).
7. Bashir, R., Luka, J., Cheloha, K., Chamberlain, M. & Hochberg, F. Expression of Epstein-Barr virus proteins in primary CNS lymphoma in AIDS patients. *Neurology* **43**, 2358-62 (1993).
8. Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
9. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* **6**, 8111 (2015).
10. Cerhan, J.R. *et al.* Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. *Nat Genet* **46**, 1233-8 (2014).
11. Mukerji, J., Olivieri, K.C., Misra, V., Agopian, K.A. & Gabuzda, D. Proteomic analysis of HIV-1 Nef cellular binding partners reveals a role for exocyst complex proteins in mediating enhancement of intercellular nanotube formation. *Retrovirology* **9**, 33 (2012).
12. Acquaviva, J., Chen, X. & Ren, R. IRF-4 functions as a tumor suppressor in early B-cell development. *Blood* **112**, 3798-806 (2008).
13. Pathak, S. *et al.* IRF4 is a suppressor of c-Myc induced B cell leukemia. *PLoS One* **6**, e22628 (2011).
14. Kreher, S. *et al.* Prognostic impact of B-cell lymphoma 6 in primary CNS lymphoma. *Neuro Oncol* **17**, 1016-21 (2015).
15. Renaud, M. *et al.* Autosomal recessive cerebellar ataxia type 3 due to ANO10 mutations: delineation and genotype-phenotype correlation study. *JAMA Neurol* **71**, 1305-10 (2014).
16. Sasaki, S. *et al.* Cloning and expression of human B cell-specific transcription factor BACH2 mapped to chromosome 6q15. *Oncogene* **19**, 3739-49 (2000).
17. Sakane-Ishikawa, E. *et al.* Prognostic significance of BACH2 expression in diffuse large B-cell lymphoma: a study of the Osaka Lymphoma Study Group. *J Clin Oncol* **23**, 8012-7 (2005).
18. Swaminathan, S. *et al.* BACH2 mediates negative selection and p53-dependent tumor suppression at the pre-B cell receptor checkpoint. *Nat Med* **19**, 1014-22 (2013).
19. Montesinos-Rongen, M., Van Roost, D., Schaller, C., Wiestler, O.D. & Deckert, M. Primary diffuse large B-cell lymphomas of the central nervous system are targeted by aberrant somatic hypermutation. *Blood* **103**, 1869-75 (2004).
20. Al Olama, A.A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat Genet* **46**, 1103-9 (2014).

21. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet* **47**, 373-80 (2015).
22. Law, P.J. *et al.* Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. *Nat Commun* **8**, 14175 (2017).
23. Mitchell, J.S. *et al.* Genome-wide association study identifies multiple susceptibility loci for multiple myeloma. *Nat Commun* **7**, 12050 (2016).
24. Rothman, N. *et al.* A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat Genet* **42**, 978-84 (2010).
25. Tenesa, A. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* **40**, 631-7 (2008).
26. Goode, E.L. *et al.* A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat Genet* **42**, 874-9 (2010).
27. Gudmundsson, J. *et al.* A common variant at 8q24.21 is associated with renal cell cancer. *Nat Commun* **4**, 2776 (2013).
28. Wolpin, B.M. *et al.* Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nat Genet* **46**, 994-1000 (2014).
29. Melin, B.S. *et al.* Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nat Genet* **49**, 789-794 (2017).
30. Sud, A. *et al.* Genome -wide association study of classical Hodgkin lymphoma identifies key regulators of disease susceptibility. *Nat Commun* **In press**(2017).
31. Conde, L. *et al.* Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. *Nat Genet* **42**, 661-4 (2010).
32. Wang, S.S. *et al.* Human leukocyte antigen class I and II alleles in non-Hodgkin lymphoma etiology. *Blood* **115**, 4820-3 (2010).
33. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).
34. Ward, L.D. & Kellis, M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res* **44**, D877-81 (2016).
35. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-7 (2012).
36. Kawaji, H. *et al.* The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Genome Biol* **10**, R40 (2009).
37. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
38. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-9 (2012).
39. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-43 (2013).
40. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).
41. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47**, 598-606 (2015).
42. Bruno, A. *et al.* Mutational analysis of primary central nervous system lymphoma. *Oncotarget* **5**, 5065-75 (2014).
43. Rickert, R.C. New insights into pre-BCR and BCR signalling with relevance to B cell malignancies. *Nat Rev Immunol* **13**, 578-91 (2013).

44. Schmidlin, H., Diehl, S.A. & Blom, B. New insights into the regulation of human B-cell differentiation. *Trends Immunol* **30**, 277-85 (2009).
45. McDonnell, T.J. & Korsmeyer, S.J. Progression from lymphoid hyperplasia to high-grade malignant lymphoma in mice transgenic for the t(14; 18). *Nature* **349**, 254-6 (1991).
46. Taub, R. *et al.* Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells. *Proc Natl Acad Sci U S A* **79**, 7837-41 (1982).
47. Ladanyi, M., Offit, K., Jhanwar, S.C., Filippa, D.A. & Chaganti, R.S. MYC rearrangement and translocations involving band 8q24 in diffuse large cell lymphomas. *Blood* **77**, 1057-63 (1991).
48. Muto, A. *et al.* The transcriptional programme of antibody class switching involves the repressor Bach2. *Nature* **429**, 566-71 (2004).
49. Hunter, J.E. *et al.* The NF-kappaB subunit c-Rel regulates Bach2 tumour suppressor expression in B-cell lymphoma. *Oncogene* **35**, 3476-84 (2016).
50. Hammer, C. *et al.* Amino Acid Variation in HLA Class II Proteins Is a Major Determinant of Humoral Response to Common Viruses. *Am J Hum Genet* **97**, 738-43 (2015).
51. Anderson, C.A. *et al.* Data quality control in genetic case-control association studies. *Nat Protoc* **5**, 1564-73 (2010).
52. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
53. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
54. Liu, J.Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* **42**, 436-40 (2010).
55. Higgins, J.P. & Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat Med* **21**, 1539-58 (2002).
56. de Souza, N. The ENCODE project. *Nat Methods* **9**, 1046 (2012).
57. Rendeiro, A.F. *et al.* Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat Commun* **7**, 11938 (2016).
58. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).
59. Davydov, E.V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).
60. Cowper-Sal lari, R. *et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat Genet* **44**, 1191-8 (2012).
61. Scales, M., Jager, R., Migliorini, G., Houlston, R.S. & Henrion, M.Y. visPIG--a web tool for producing multi-region, multi-track, multi-scale plots of genetic data. *PLoS One* **9**, e107497 (2014).

† **LOC Network**: Alexandra Benouaich-Amiel[1], Chantal Campello[2] , Frédéric Charlotte[3] , Nadine Martin-Duverneuil[4] , Loïc Feuvret[5] , Aurélie Kas[6] , Soledad Navarro[7], Chiara Villa[8] , Franck Bielle[9] , Fabrice Chretien[10] , Marie-Claire Tortel[11] , Guillaume Gauchotte[12] , Emmanuelle Uro-Coste[13] , Catherine Godfrain[14] ,Valérie Rigau[15] , Myrto Costopoulos[16] , Magalie Le Garff-Tavernier[17], David Meyronnet[18] , Audrey Rousseau[19] , Clovis Adam[20] , Thierry Lamy[21] , Cécile Chabrot[22] , Eileen Boyle[23], Marie Blonski[24], Anna Schmitt[25].

1. Department of Neurology, CHU Toulouse, 31059 Toulouse, France.
2. Service de Neuro-Oncologie, CHU Timone,13005 Marseille – France.
3. Department of Pathology, CHU Pitié-Salpêtrière, 75013 Paris, France.
4. Department of Neuroradiology, CHU Pitié-Salpêtrière, 75013 Paris, France.
5. Department of Radiotherapy, CHU Pitié-Salpêtrière, 75013 Paris, France.
6. Department of Nuclear Medicine, CHU Pitié-Salpêtrière, 75013 Paris, France.
7. Department of neurosurgery, CHU Pitié-Salpêtrière, 75013 Paris, France.
8. Department of Pathology, Hôpital Foch, 92151 Suresnes – France.
9. Inserm, U 1127, ICM, F-75013 Paris, France ; CNRS, UMR 7225, ICM, F-75013 Paris, France ; Institut du Cerveau et de la Moelle épinière ICM, Paris 75013, France ; Sorbonne Universités, UPMC Université Paris 06, UMR S 1127, F-75013 Paris, France ; AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de Neuropathologie, 75013 Paris, France.
10. Department of Neuropathology, Centre Hospitalier Sainte Anne, Paris, France.
11. Department of Pathology, Hôpitaux Civils de Colmar, 68024, Colmar Cedex, France.
12. Department of Pathology, Nancy University Hospital, Vandoeuvre-lès-Nancy, France
13. Department of Pathology, Toulouse University Hospital and INSERM UMR1037 , Institut Universitaire du Cancer-Oncopole, 1 avenue Irène Joliot-Curie, 31059 Toulouse cedex 9, France.
14. Department of Pathology, CHU Clermont-Ferrand, 63000 Clermont-Ferrand, France.
15. Department of Pathology, CHU Montpellier, 34000 Montpellier, France.
16. Department of Biological Hematology, AP-HP, Groupe Hospitalier Pitié-Salpêtrière, 75013, Paris, France.
17. Department of Biological Hematology, AP-HP, Groupe Hospitalier Pitié-Salpêtrière, 75013, Paris, France.
18. Department of Pathology, University Hospital of Lyon, 69002, Lyon, France.
19. Department of Pathology, CHU Angers, 49933 Angers – France.

20. Department of Pathology, CHU Kremlin Bicêtre, 94270 Le Kremlin Bicêtre.
21. CHU Rennes, Rennes, 35033, France.
22. CHU Clermont Ferrand, Clermont Ferrand, 63000, France.
23. Department of Haematology, Lille University Hospital, Lille, 59037, France.
24. CHU Nancy. Nancy, 54500, France.
25. CHU Bordeaux. Bordeaux ,33000, France.

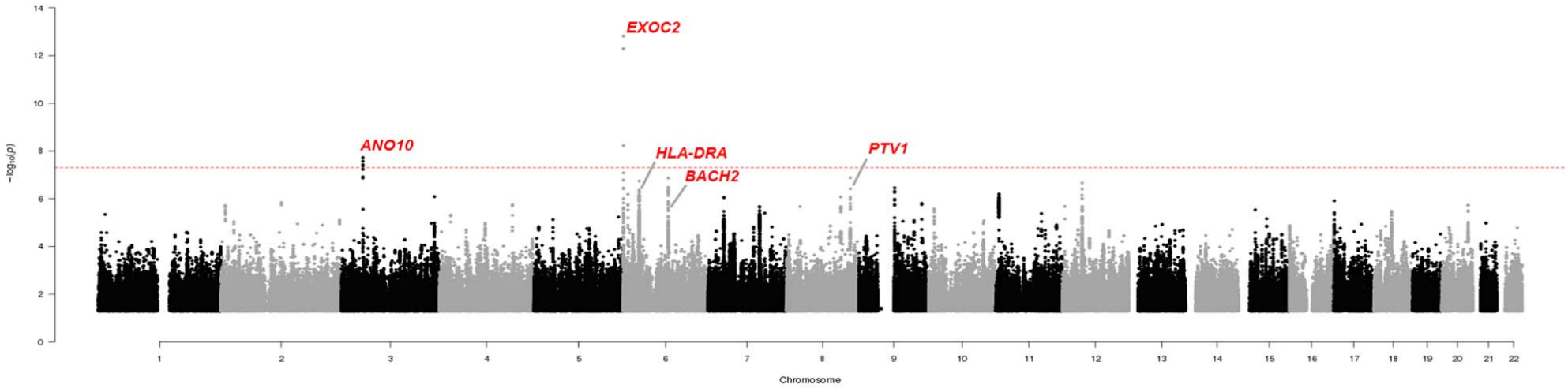


Figure 1: Manhattan plot of association *P*-values for primary cerebral nervous system lymphoma. Shown are the genome-wide $-\log_{10}P$-values (two-sided) of >10 million successfully imputed autosomal SNPs in 475 cases and 1,134 controls. The red horizontal line represents the genome-wide significance threshold of $P=5.0 \times 10^{-8}$.

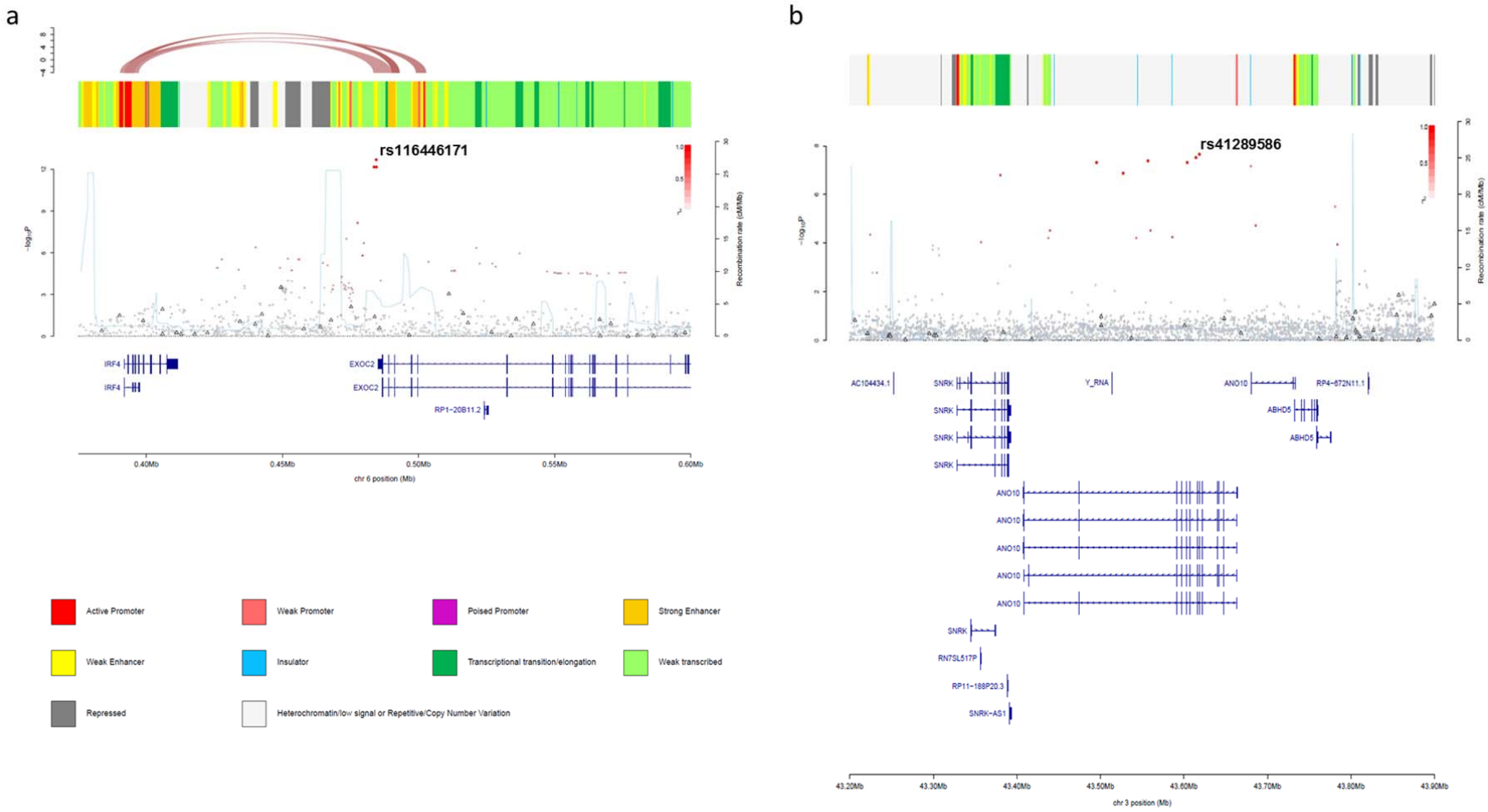


Figure 2: Regional plots of association results and recombination rates for new risk loci for primary cerebral nervous system lymphoma. Results shown for (a) 6p25 and (b) 3q21. Plots (drawn using visPig64) show association results of both genotyped (triangles) and imputed (circles) SNPs in the GWAS samples and recombination rates. −log10P values (y axes) of the SNPs are shown according to their chromosomal positions (x axes). The sentinel SNP in each combined analysis is shown as a large circle or triangle and is labelled by its rsID. The color intensity of each symbol reflects the extent of LD with the top genotyped SNP, white (r^2=0) through to dark red (r^2=1.0). Genetic recombination rates, estimated using 1000 Genomes Project samples, are shown with a light blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the chromatin-state segmentation track (ChromHMM) for lymphoblastoid cells using data from the HapMap ENCODE Project, and the positions of genes and transcripts mapping to the region of association.

Disney-Hogg et al. BMC Medicine (2018) 16:42
https://doi.org/10.1186/s12916-018-1027-5

BMC Medicine

RESEARCH ARTICLE Open Access

Impact of atopy on risk of glioma: a Mendelian randomisation study

Linden Disney-Hogg[1†], Alex J. Cornish[1†], Amit Sud[1], Philip J. Law[1], Ben Kinnersley[1], Daniel I. Jacobs[2], Quinn T. Ostrom[3], Karim Labreche[1], Jeanette E. Eckel-Passow[4], Georgina N. Armstrong[2], Elizabeth B. Claus[5,6], Dora Il'yasova[7,8,9], Joellen Schildkraut[8,9], Jill S. Barnholtz-Sloan[3], Sara H. Olson[10], Jonine L. Bernstein[10], Rose K. Lai[11], Minouk J. Schoemaker[1], Matthias Simon[12], Per Hoffmann[13,14], Markus M. Nöthen[14,15], Karl-Heinz Jöckel[16], Stephen Chanock[17], Preetha Rajaraman[17], Christoffer Johansen[18,19], Robert B. Jenkins[20], Beatrice S. Melin[21], Margaret R. Wrensch[22,23], Marc Sanson[24,25], Melissa L. Bondy[2] and Richard S. Houlston[1,26*]

Abstract

Background: An inverse relationship between allergies with glioma risk has been reported in several but not all epidemiological observational studies. We performed an analysis of genetic variants associated with atopy to assess the relationship with glioma risk using Mendelian randomisation (MR), an approach unaffected by biases from temporal variability and reverse causation that might have affected earlier investigations.

Methods: Two-sample MR was undertaken using genome-wide association study data. We used single nucleotide polymorphisms (SNPs) associated with atopic dermatitis, asthma and hay fever, IgE levels, and self-reported allergy as instrumental variables. We calculated MR estimates for the odds ratio (OR) for each risk factor with glioma using SNP-glioma estimates from 12,488 cases and 18,169 controls, using inverse-variance weighting (IVW), maximum likelihood estimation (MLE), weighted median estimate (WME) and mode-based estimate (MBE) methods. Violation of MR assumptions due to directional pleiotropy were sought using MR-Egger regression and HEIDI-outlier analysis.

Results: Under IVW, MLE, WME and MBE methods, associations between glioma risk with asthma and hay fever, self-reported allergy and IgE levels were non-significant. An inverse relationship between atopic dermatitis and glioma risk was found by IVW (OR 0.96, 95% confidence interval (CI) 0.93–1.00, $P = 0.041$) and MLE (OR 0.96, 95% CI 0.94–0.99, $P = 0.003$), but not by WME (OR 0.96, 95% CI 0.91–1.01, $P = 0.114$) or MBE (OR 0.97, 95% CI 0.92–1.02, $P = 0.194$).

Conclusions: Our investigation does not provide strong evidence for relationship between atopy and the risk of developing glioma, but findings do not preclude a small effect in relation to atopic dermatitis. Our analysis also serves to illustrate the value of using several MR methods to derive robust conclusions.

Keywords: Mendelian randomisation, Allergy, Cancer, Glioma, Risk

* Correspondence: richard.houlston@icr.ac.uk
†Equal contributors
[1]Division of Genetics and Epidemiology, The Institute of Cancer Research, 15 Cotswold Road, London SM2 5NG, UK
[26]Division of Molecular Pathology, The Institute of Cancer Research, London, UK
Full list of author information is available at the end of the article

BioMed Central

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Although glioma accounts for approximately 80% of malignant primary brain tumours [1], to date, few aetiological risk factors are well established for the disease [2]. Over the past three decades the search for an immune-mediated risk factor that might influence risk has led to studies of a possible relationship between multiple allergic conditions and autoimmune disorders with glioma [3].

Several case-control studies have shown that self-reported allergic conditions may protect against glioma [4]. For example, in the International Adult Brain Tumour Study, based on 1178 glioma patients, an odds ratio (OR) of 0.59 was found for any self-reported allergy [5]. Other case-control studies have reported similar ORs, however, most have been reliant on substantial numbers of proxy informants (up to 44%) [4, 6] and have potential bias as a consequence of how controls were ascertained, thereby casting doubt on findings. In contrast to case-control studies, evidence for an association between glioma and allergy from cohort-based analyses has been less forthcoming [7], although such studies have been poorly powered to demonstrate a relationship.

Assaying IgE potentially reduces bias stemming from self-reporting despite levels not necessarily corresponding to specific allergies or equating to a single allergic response. Nevertheless, measurement of IgE has been explored by a number of researchers seeking to identify risk factors for glioma [8–10]. In a case-control study of 228 cases and 289 controls performed in 2004 [8], self-reported allergies and IgE levels were both inversely associated with glioma, but concordance between the two outcomes was poor. In a larger study of 535 cases and 532 controls [11], both self-reported allergies and IgE levels were inversely related to glioma risk; however, IgE levels in patients were affected by temozolomide treatment. A case-control study nested within the European Prospective Investigation into Cancer and Nutrition cohort based on prospectively collected serum IgE levels reported a non-significant OR of 0.73 [9]. A similar nested case-control study performed in the USA based on 181 cases reported a non-significant OR of 0.72 for high serum IgE [10].

Several mechanisms have been proposed to explain a possible association between atopic disease and glioma [12]. The findings could reflect a true causal effect of the heightened immune function reported for atopy on tumour development. Alternatively, the associations observed might be non-causal, arising as a consequence of methodological biases inherent in the study design. Imprecisely defined exposures, such as allergic disease, are likely to have affected the validity of the findings of both case-control and cohort studies. The heterogeneous description of allergy in studies and different levels of detail in self-reporting on individual allergies complicate the interpretation of results. Additional biases include possible selection bias in controls, recall bias from self-reported allergy assessment and reverse causation or confounding from unmeasured effects. Finally, the high frequency of exposure ascertainment by proxy for cases is also likely to have systematically biased findings.

Mendelian randomisation (MR) analysis can be used to minimise potential biases in conventional observational studies and to determine the causal association of an exposure with an outcome such as disease risk [13]. The causal association can also be manifested by common genetic and biological pathways that determine two sequentially developed phenotypes such as an atopic trait and glioma risk. Atopy has a strong heritable basis [14, 15] and, thus far, genome-wide association studies (GWAS) have identified over 50 loci associated with different atopy-related traits [16]. The alleles associated with atopy should be randomly assigned to offspring from parents during mitosis, a process analogous to the random assignment of subjects to an exposure of interest in randomised clinical trials. Thus, genetic scores summarising the effects of single nucleotide polymorphisms (SNPs) associated with atopy-related traits can serve as instrumental variables (IVs) in a MR analysis of atopy and glioma risk.

To examine the nature of the association between atopy and glioma, we implemented two-sample MR [17] to estimate associations between atopy-associated SNPs and glioma risk using summary data from the recent GWAS meta-analysis performed by the Glioma International Case-Control Consortium study [18].

Methods

Two-sample MR was undertaken using GWAS data. Ethical approval was not sought for this specific project because all data came from the summary statistics of published GWAS, and no individual-level data were used.

Glioma genotyping data

Glioma genotyping data were derived from the most recent meta-analysis of GWAS in glioma, which related > 10 million genetic variants (after imputation) to glioma, in 12,488 glioma patients and 18,169 controls from eight independent GWAS datasets of individuals of European descent [18] (Additional file 1: Table S1). Comprehensive details of the genotyping and quality control of the seven GWAS have been previously reported [18].

Genetic variant instruments for atopic traits

SNPs associated with each of the atopy-related traits investigated, namely atopic dermatitis (eczema), asthma and hay fever, IgE level, and self-reported allergy, by the

NHGRI-EBI GWAS Catalog [19–26] at genome-wide significance (*i.e.* $P \leq 5.0 \times 10^{-8}$) in individuals with European ancestry were used as IVs. To avoid co-linearity between SNPs for each trait, we excluded SNPs that were correlated (*i.e.* r^2 value of ≥ 0.001) within each trait, and only considered the SNPs with the strongest effect on the trait for use as IVs (Additional file 2: Table S2). For each SNP, we recovered the chromosome position, risk allele, association estimates (*per*-allele log-OR) and standard errors (Table 1). The allele that was associated with increased risk of the exposure was considered the effect allele. For IgE level, the allele associated with an increase in serum IgE was considered the effect allele. Allele frequencies for these SNPs were compared between the atopy-related trait and glioma datasets to ensure that the effect estimates were recorded with respect to the same allele. Gliomas are heterogeneous and different tumour subtypes, defined in part by malignancy grade (e.g. pilocytic astrocytoma World Health Organization (WHO) grade I, diffuse 'low-grade' glioma WHO grade II, anaplastic glioma WHO grade III and glioblastoma (GBM) WHO grade IV) can be distinguished [27]. For the sake of brevity we considered gliomas as being either GBM or non-GBM.

Two-sample MR method

The association between each atopy-related trait and glioma was examined using MR on summary statistics using the inverse-variance weighting (IVW) method and maximum likelihood estimation (MLE) as *per* Burgess *et al.* [28]. The IVW ratio estimate ($\hat{\beta}$) of all SNPs associated with each atopy-related trait on glioma risk was calculated as follows:

$$\hat{\beta} = \frac{\sum_k X_k Y_k \sigma_{Y_k}^{-2}}{\sum_k X_k^2 \sigma_{Y_k}^{-2}}$$

Where X_k corresponds to the association of SNP k (as log of the OR per risk allele) with the atopy-related trait Y_k is the association between SNP k and glioma risk (as log OR) with standard error σ_{Y_k}. The estimate for ($\hat{\beta}$) represents the causal increase in the log odds of glioma for each trait. The standard error of the combined ratio estimate is given by:

$$se\left(\hat{\beta}\right) = \sqrt{\frac{1}{\sum_k X_k^2 \sigma_{Y_k}^{-2}}}$$

For the MLE, a bivariate normal distribution for the genetic associations was assumed, and the R function *optim* was used to estimate β. $se(\hat{\beta})$ was calculated using observed information. The correlation between the errors of Y_k and X_k was taken to be 0 as they were derived from independent studies.

A central tenet in MR is the absence of pleiotropy (*i.e.* a gene influencing multiple traits) between the SNPs influencing the exposure and outcome disease risk [13]. This would be revealed as deviation from a linear relationship between SNPs and their effect size for atopy and glioma risk. To examine for violation of the standard IV assumptions in our analysis we first performed MR-Egger regression, as well as HEIDI-outlier analysis, as per Zhu *et al.* [29], imposing the advocated threshold of $P \leq 0.01$. Additionally, we derived weighted median estimates (WME) [30] and mode-based estimates (MBE) [31] to establish the robustness of findings.

Atopic dermatitis, asthma and hay fever, and self-reported allergy as well as all of the disease outcomes (all glioma, GBM and non-GBM glioma) are binary. The causal effect estimates therefore represent the odds for outcome disease risk per unit increase in the log OR of the exposure disease [32]. These ORs were converted to represent the OR for the outcome disease per doubling in odds of the exposure disease to aid interpretation [32].

For each statistical test we considered a global significance level of $P < 0.05$ as being satisfactory to derive conclusions. To assess the robustness of our conclusions, we initially imposed a conservative Bonferroni-corrected significance threshold of 0.0125 (*i.e.* 0.05/4 atopy-related traits). We considered a P value ≥ 0.05 as non-significant (*i.e.* no association), a $P < 0.05$ as evidence for a potential causal association, and a $P < 0.0125$ as significant evidence for an association. All statistical analyses were undertaken using R software (Version 3.1.2). The meta and gsmr packages were used to generate forest plots and perform HEIDI-outlier analysis [29].

The power of a MR investigation depends greatly on the proportion of variance in the risk factor that is explained by the IV. We estimated study power *a priori* using the methodology of Burgess *et al.* [33], making use of published estimates of the heritability of trait associated IV SNPs [34–36], as well as estimates found by direct calculation (Additional file 3: Table S3), and the reported effect of each trait on glioma risk reported in a meta-analysis of epidemiological studies [18]. Additional file 4: Table S4 shows the range of ORs for which we had less than 80% power to detect for each of the four atopy-related traits.

Simulation model

Through simulation we evaluated the suitability of using each employed MR method in a two-sample setting with binary-exposure and binary-outcome data. Let i index genetic variants, N be the total number of genetic variants, and j index individuals. Genetic variants g_{ij} were generated independently by sampling from a Binomial(2,p_j) distribution with probability p_j drawn from a Uniform(0.1,0.9) distribution, to mimic bi-allelic SNPs in Hardy–Weinberg

Table 1 Variant and effect allele with frequencies and magnitude of effect on each atopy-related trait and strength of association with glioma

Region	SNP	Position (bp)[a]	Alleles[b]	MAF	Hay fever and asthma	Glioma
					OR (95% CI)	OR (95% CI)
2q12.1	rs10197862	102,966,549	G/A	G = 0.161	1.24 (1.16–1.32)	0.98 (0.93–1.03)
4p14	rs4833095	38,799,710	C/T	T = 0.425	1.20 (1.14–1.26)	1.03 (0.99–1.08)
5q22.1	rs1837253	110,401,872	T/C	T = 0.382	1.17 (1.11–1.23)	0.96 (0.93–1.00)
8q21.13	rs7009110	81,291,879	C/T	C = 0.467	1.14 (1.09–1.19)	0.98 (0.94–1.01)
9p24.1	rs72699186	6,175,855	A/T	T = 0.110	1.26 (1.17–1.36)	0.97 (0.93–1.02)
11q13.5	rs2155219	76,299,194	G/T	G = 0.468	1.17 (1.13–1.21)	1.01 (0.97–1.05)
15q22.33	rs17294280	67,468,285	A/G	G = 0.120	1.18 (1.12–1.25)	0.98 (0.94–1.03)
16p13.13	rs62026376	11,228,712	T/C	T = 0.144	1.17 (1.11–1.23)	0.97 (0.93–1.01)
17q21.1	rs7212938	38,122,680	T/G	G = 0.473	1.16 (1.11–1.22)	1.00 (0.97–1.04)
Region	SNP	Position[a]	Alleles[b]	MAF	Atopic dermatitis	Glioma
					OR (95% CI)	OR (95% CI)
1q21.3	rs11205006	152,440,176	T/A	A = 0.265	1.62 (1.48–1.77)	0.96 (0.91–1.02)
1q21.3	rs2228145	154,426,970	A/C	C = 0.293	1.15 (1.10–1.20)	0.99 (0.96–1.03)
2p25.1	rs10199605	8,495,097	A/G	A = 0.244	1.04 (1.03–1.06)	1.01 (0.97–1.05)
2p13.3	rs112111458	71,100,105	G/A	G = 0.224	1.08 (1.05–1.10)	0.98 (0.92–1.03)
2q24.3	rs6720763	167,992,286	T/C	C = 0.320	1.29 (1.18–1.41)	1.02 (0.97–1.06)
5p13.2	rs10214237	35,883,734	C/T	C = 0.176	1.06 (1.05–1.08)	0.98 (0.94–1.02)
5q31.1	rs1295686	131,995,843	C/T	T = 0.422	1.35 (1.22–1.49)	0.99 (0.95–1.03)
6p21.32	rs12153855	32,074,804	T/C	C = 0.125	1.58 (1.40–1.78)	0.97 (0.92–1.03)
8q21.13	rs6473227	81,285,892	A/C	A = 0.473	1.06 (1.05–1.08)	0.98 (0.94–1.02)
9p21.3	rs10738626	22,373,457	C/T	C = 0.397	1.23 (1.15–1.32)	0.96 (0.93–1.00)
10p15.1	rs6602364	6,038,853	G/C	G = 0.492	1.05 (1.03–1.07)	1.03 (0.99–1.07)
11q13.1	rs10791824	65,559,266	A/G	G = 0.490	1.15 (1.12–1.19)	0.99 (0.95–1.02)
11q24.3	rs7127307	128,187,383	C/T	C = 0.488	1.09 (1.07–1.11)	0.99 (0.95–1.03)
11q13.5	rs7130588	76,270,683	G/A	G = 0.216	1.29 (1.20–1.38)	1.02 (0.98–1.06)
14q13.2	rs2143950	35,572,357	C/T	T = 0.215	1.08 (1.06–1.10)	1.01 (0.97–1.06)
16p13.13	rs2041733	11,229,589	C/T	T = 0.496	1.09 (1.06–1.11)	0.97 (0.94–1.01)
19p13.2	rs2164983	8,789,381	C/A	A = 0.169	1.16 (1.10–1.22)	0.95 (0.90–1.00)
20q13.33	rs909341	62,328,742	T/C	T = 0.262	1.32 (1.21–1.44)	1.32 (1.26–1.37)
Region	SNP	Position[a]	Alleles[b]	MAF	IgE level[c]	Glioma
					OR (95% CI)	OR (95% CI)
1q23.2	rs2251746	159,272,060	C/T	C = 0.015	1.09 (1.08–1.11)	0.98 (0.95–1.02)
5q31.1	rs20541	131,995,964	A/G	A = 0.270	1.08 (1.06–1.10)	1.01 (0.97–1.06)
6p22.1	rs2571391	29,923,838	C/A	C = 0.303	1.06 (1.05–1.08)	0.97 (0.94–1.01)
6p21.32	rs2858331	32,681,277	A/G	G = 0.490	1.04 (1.03–1.06)	1.02 (0.98–1.06)
12q13.3	rs1059513	57,489,709	C/T	C = 0.070	1.13 (1.09–1.17)	0.97 (0.92–1.03)
Region	SNP	Position[a]	Alleles[b]	MAF	Self–reported allergy	Glioma
					OR (95% CI)	OR (95% CI)
2q12.1	rs10189699	102,879,464	A/C	A = 0.143	1.16 (1.12–1.20)	0.99 (0.94–1.04)
2q33.1	rs10497813	198,914,072	T/G	T = 0.401	1.08 (1.05–1.11)	0.99 (0.96–1.03)
3q28	rs9860547	188,128,979	G/A	A = 0.272	1.08 (1.05–1.11)	1.02 (0.98–1.06)
4p14	rs2101521	38,811,551	A/G	A = 0.475	1.15 (1.12–1.18)	1.02 (0.98–1.07)

Table 1 Variant and effect allele with frequencies and magnitude of effect on each atopy-related trait and strength of association with glioma *(Continued)*

4q27	rs17388568	123,329,369	G/A	A = 0.141	1.08 (1.05–1.11)	1.01 (0.97–1.05)
5p13.1	rs7720838	40,486,896	G/T	T = 0.362	1.08 (1.06–1.11)	1.02 (0.99–1.06)
5q22.1	rs1438673	110,467,499	T/C	C = 0.296	1.12 (1.09–1.15)	0.97 (0.94–1.01)
6p21.33	rs9266772	31,352,113	T/C	C = 0.175	1.11 (1.08–1.14)	1.03 (0.98–1.08)
9p24.1	rs7032572	6,172,380	A/G	G = 0.114	1.12 (1.08–1.16)	0.97 (0.93–1.02)
10p14	rs962993	9,053,132	T/C	T = 0.106	1.07 (1.05–1.10)	1.02 (0.98–1.06)
11q13.5	rs2155219	76,999,194	G/T	G = 0.468	1.11 (1.09–1.14)	1.01 (0.97–1.05)
15q22.33	rs17228058	67,450,305	A/G	G = 0.100	1.08 (1.05–1.11)	1.00 (0.96–1.04)
17q21.1	rs9303280	38,074,031	T/C	T = 0.346	1.07 (1.05–1.09)	0.98 (0.94–1.02)
20q13.2	rs6021270	50,141,264	C/T	T = 0.346	1.16 (1.10–1.22)	1.02 (0.94–1.10)

[a]NCBI build 37
[b]Reference allele/effect allele
[c]Per standard deviation
MAF minor allele frequency, *OR* odds ratio, *SNP* single nucleotide polymorphism

equilibrium. Let w_j correspond to the per-allele OR for the exposure disease, sampled from ORs reported for genome-wide significant SNPs reported in the GWAS Catalog [37], and v be the OR for the outcome disease per doubling in odds of the exposure disease. For each individual, exposure disease odds x_j, outcome disease odds y_j, exposure disease status a_j, and outcome disease status b_j were determined as follows:

$$x_j = x_0 \prod_{i=1}^{N} w_i^{g_{ij}}$$

$$y_j = y_0 \times 2^{\log_2 x_j \times \log_2 v}$$

$$a_j \sim \text{Binomial}\left(1, \frac{x_j}{1+x_j}\right)$$

$$b_j \sim \text{Binomial}\left(1, \frac{y_j}{1+y_j}\right)$$

Data for 1,000,000 individuals were simulated and partitioned at random to reflect the two-sample setting. Cases and controls for the exposure and outcome GWAS were sampled from each half of the dataset using the exposure and outcome disease statuses of each individual, and association statistics computed under an additive logistic regression model. To ensure the simulated data closely resembled the atopy-related trait and glioma data, the simulation analysis was repeated for each binary atopy-related trait using the same number of genetic variants as IVs and the same numbers of case and control individuals as used to estimate the atopy-related trait and glioma association statistics (Additional file 5: Table S5). Parameters x_0 = 0.0005 and y_0 = 0.01 were chosen to ensure the prevalence of the simulated exposure and outcome diseases were similar to that of the atopy-related traits and glioma, respectively (Additional file 5: Table S5). To determine the suitability of each MR method we considered two scenarios: (1) no causal relationship between exposure and outcome (v = 1.00) and (2) a causal relationship between exposure and outcome (v = 1.33). We performed 100 simulations for each scenario for each binary atopy-related trait.

Results

The atopic dermatitis risk SNP rs909341, which is highly correlated with the chromosome 20q13.33 glioma risk SNP rs2297440 (D' = 0.89, r^2 = 0.77), was strongly associated with risk of glioma ($P = 2.10 \times 10^{-34}$). Testing for pleiotropy using HEIDI-outlier analysis formally identified rs909341 as violating the assumption of the instrument on the outcome. Henceforth, we confined our analysis of the relationship between atopic dermatitis and glioma to a dataset excluding this SNP.

Figure 1 shows forest plots of ORs for glioma generated from the SNPs. There was minimal evidence of heterogeneity between variants for asthma and hay fever, atopic dermatitis, IgE levels and self-reported allergy (respective I^2 and P_{het} values being 28% and 0.192, 8% and 0.377, 0% and 0.444, and 0% and 0.707). Including rs909341 in the analysis for atopic dermatitis, the I^2 value was 90% and $P_{het} < 10^{-4}$ (Additional file 6: Figure S1), providing further evidence that inclusion of this SNP would invalidate the MR analysis.

The results of the IVW, MLE, WME, MBE and MR-Egger methods are summarised in Table 2. Using the IVW method to pool results from individual SNPs, no associations (*i.e.* $P \geq 0.05$) were identified between genetically conferred risk of raised IgE level (OR 0.88, 95% CI 0.69–1.13, P = 0.319), asthma and hay fever (OR 0.96, 95% CI 0.90–1.03, P = 0.248), or self-reported allergy (OR 1.03, 95% CI 0.95–1.11, P = 0.534) with risk of all glioma. There was some support for an inverse relationship

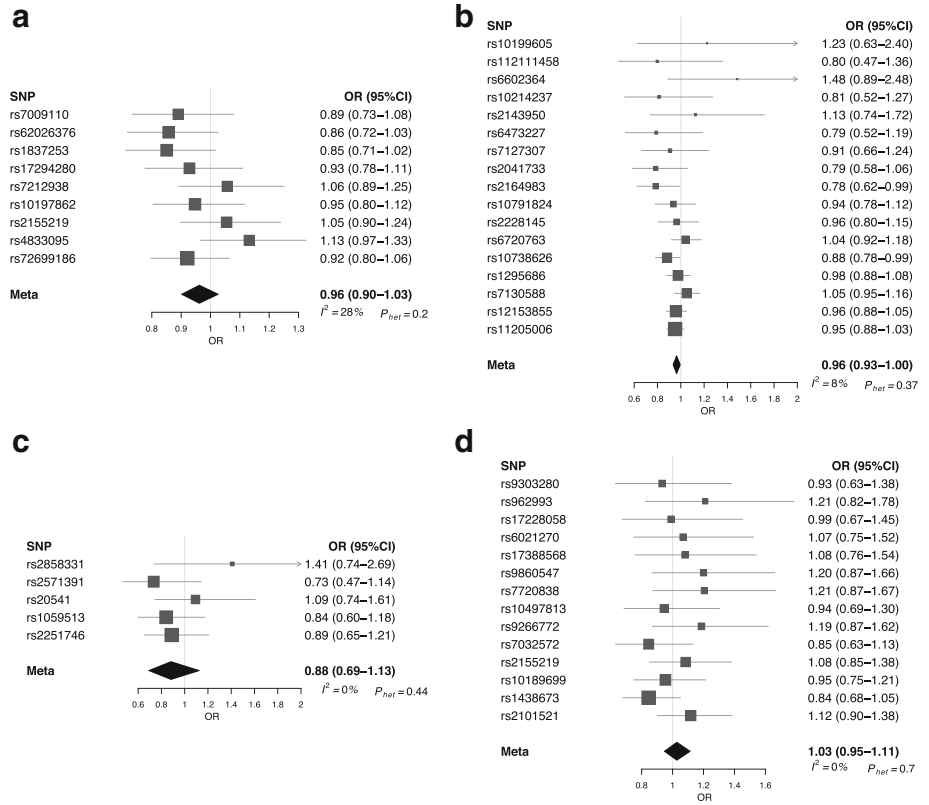


Fig. 1 Forest plot of Wald odds ratios (ORs) and 95% confidence intervals generated from single nucleotide polymorphisms (SNPs) associated with atopy-related traits. ORs for individual SNPs are listed according to magnitude of effect in the instrumental variable analysis and are presented with pooled effects using the inverse-variance weighting method. Squares represent the point estimate, and the bars are the 95% confidence intervals. **a** Asthma and hay fever, **b** atopic dermatitis, **c** IgE level, **d** self-reported allergy

between atopic dermatitis and glioma risk (OR 0.96, 95% CI 0.93–1.00, P = 0.041), albeit not significant after adjustment for multiple testing.

Using MLE, no associations were identified between asthma and hay fever (OR 0.96, 95% CI 0.93–1.00, P = 0.066), IgE levels (OR 0.88, 95% CI 0.74–1.05, P = 0.157) or self-reported allergy (OR 1.02, 95% CI 0.97–1.08, P = 0.429) with risk of all glioma. For atopic dermatitis, an OR of 0.96 (95% CI 0.94–0.99, P = 0.003) was shown, which remained significant after adjusting for multiple testing. Figure 2 shows relaxation of the assumption that the correlation between the errors in X_k and Y_k is zero for each of the atopy-related traits demonstrating the consistency of findings. Specifically, for a correlation in the range –0.15 to 0.15, the association between atopic dermatitis and glioma risk remained significant.

In contrast to findings from IVW and MLE, no significant support was provided by either the WME or MBE for an association between any of the atopy-related traits and glioma risk, including atopic dermatitis (WME: OR 0.96, 95% CI 0.91–1.01, P = 0.114; MBE: OR 0.97, 95% CI 0.92–1.02, P = 0.194; Table 2).

The respective effect estimated from MR-Egger regression (Fig. 3) were 0.97 for atopic dermatitis (95% CI 0.92–1.03; P = 0.375), 0.63 for IgE levels (95% CI 0.32–1.25; P = 0.184), 0.99 for asthma and hay fever (95% CI 0.72–1.36, P = 0.951) and 0.92 for self-reported allergy (95% CI 0.69–1.22; P = 0.540), with intercepts of –0.004 (95% CI –0.014 to 0.006, P = 0.396), 0.027 (95% CI 0.001 to 0.053, P = 0.042), –0.007 (95% CI –0.030 to 0.016, P = 0.542) and 0.017 (95% CI 0.003–0.031, P = 0.018). Collectively, these findings provide possible evidence of systematic bias in the IVW estimate for IgE level and self-reported allergy, which might have arisen through overall unbalanced horizontal pleiotropy. There was no such evidence for such pleiotropy in respect of atopic dermatitis.

We explored the possibility that a relationship between atopy and glioma might be subtype specific, considering GBM and non-GBM separately. Imposing a stronger significance threshold of P = 0.00625 (0.05/8, to correct for testing four traits over two outcomes), no histology-specific associations were shown by the IVW method between asthma and hay fever, IgE levels and self-reported allergy and glioma risk, with the respective ORs for the IVW method being 0.97, 0.92 and 1.04 for

Table 2 Inverse-variance weighting, maximum likelihood estimation, weighted median estimate, mode-based estimate and Mendelian randomisation-Egger test results for combined atopy-related instrumental variables

Trait	IVW		MLE		WME		MBE		MR-Egger slope		MR-Egger intercept	
	OR (95% CI)	*P*	OR (95% CI)	*P*	OR (95% CI)	*P*	OR (95% CI)	*P*	OR (95% CI)	*P*	Estimate (95% CI)	*P*
Asthma and hay fever	0.96 (0.90–1.03)	0.248	0.96 (0.93–1.00)	0.066	0.93 (0.86–1.01)	0.087	0.91 (0.80–1.04)	0.191	0.99 (0.72–1.36)	0.951	−0.007 (−0.030 to 0.016)	0.542
Atopic dermatitis	0.96 (0.93–1.00)	0.041	0.96 (0.94–0.99)	0.003	0.96 (0.91–1.01)	0.114	0.97 (0.92–1.02)	0.194	0.97 (0.92–1.03)	0.375	0.004 (−0.014 to 0.006)	0.396
IgE level	0.88 (0.69–1.13)	0.319	0.88 (0.74–1.05)	0.157	0.83 (0.61–1.12)	0.218	0.82 (0.57–1.19)	0.355	0.63 (0.32–1.25)	0.184	0.027 (0.001 to 0.053)	0.042
Self-reported allergy	1.03 (0.95–1.11)	0.534	1.02 (0.97–1.08)	0.429	1.08 (0.97–1.20)	0.184	1.12 (0.92–1.36)	0.275	0.92 (0.69–1.22)	0.540	0.017 (0.003 to 0.031)	0.018

CI confidence interval, *IVW* inverse-variance weighting, *MBE* mode-based estimate, *MLE* maximum likelihood estimation, *MR* Mendelian randomisation, *OR* odds ratio, *WME* weighted median estimate

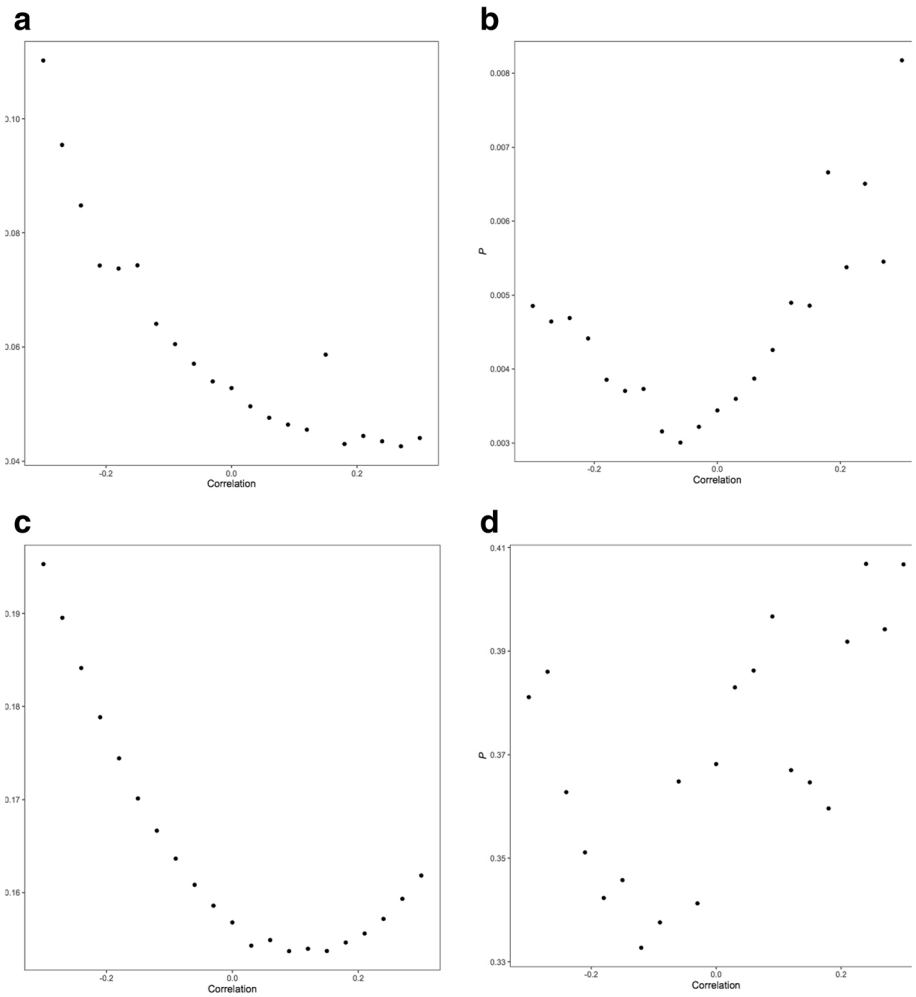


Fig. 2 Plot of *P* value of maximum likelihood estimation associations with glioma against correlation between errors in X_k and Y_k. **a** Asthma and hay fever, **b** atopic dermatitis, **c** IgE level, **d** self-reported allergy

GBM tumours, and 0.96, 0.97 and 1.04 for non-GBM tumours (Additional file 7: Table S6). For atopic dermatitis, a significant OR of 0.94 (95% CI 0.90–0.98, P = 0.004) was shown for GBM but not for non-GBM (OR 0.98, 95% CI 0.93–1.03, P = 0.421). The association between atopic dermatitis and risk of GBM was also apparent in the MLE analysis, which provided an OR of 0.94 (95% CI 0.91–0.97, $P = 2.17 \times 10^{-4}$). MR-Egger regression provided for an intercept of −0.007 (95% CI −0.019 to 0.005, P = 0.247). As with the analysis of all glioma, the association between atopic dermatitis and GBM was weaker under the WME (OR 0.96, 95% CI 0.91–1.02, P = 0.172) and MBE (OR 0.95, 95% CI 0.90–1.01, P = 0.096) frameworks.

Although previously implemented in other studies [32, 38], ratio estimators may not fully recapitulate an estimate of the causal OR in the case of binary exposures, such as atopic dermatitis, and binary outcomes such as glioma [39]. We therefore evaluated, through simulation, whether the IVW, MLE, WME, MBE and MR-Egger methods provide reliable estimates of causal ORs. When no causal relationship between exposure and outcome was simulated, each MR method provided accurate estimates of the null relationship (Additional file 5: Table S5). Conversely, when a causal relationship was simulated, the magnitudes of the relationship estimates were weakly inflated in some instances (Additional file 5: Table S5), indicating the importance of considering additional evidence when evaluating causal relationships between binary exposures and binary outcomes.

Discussion

To our knowledge, this is the first MR study evaluating a range of atopy-related traits with glioma risk. Overall, our results provide evidence for a causal protective effect of atopic dermatitis with GBM tumours, but do not provide evidence that asthma and hay fever, raised IgE

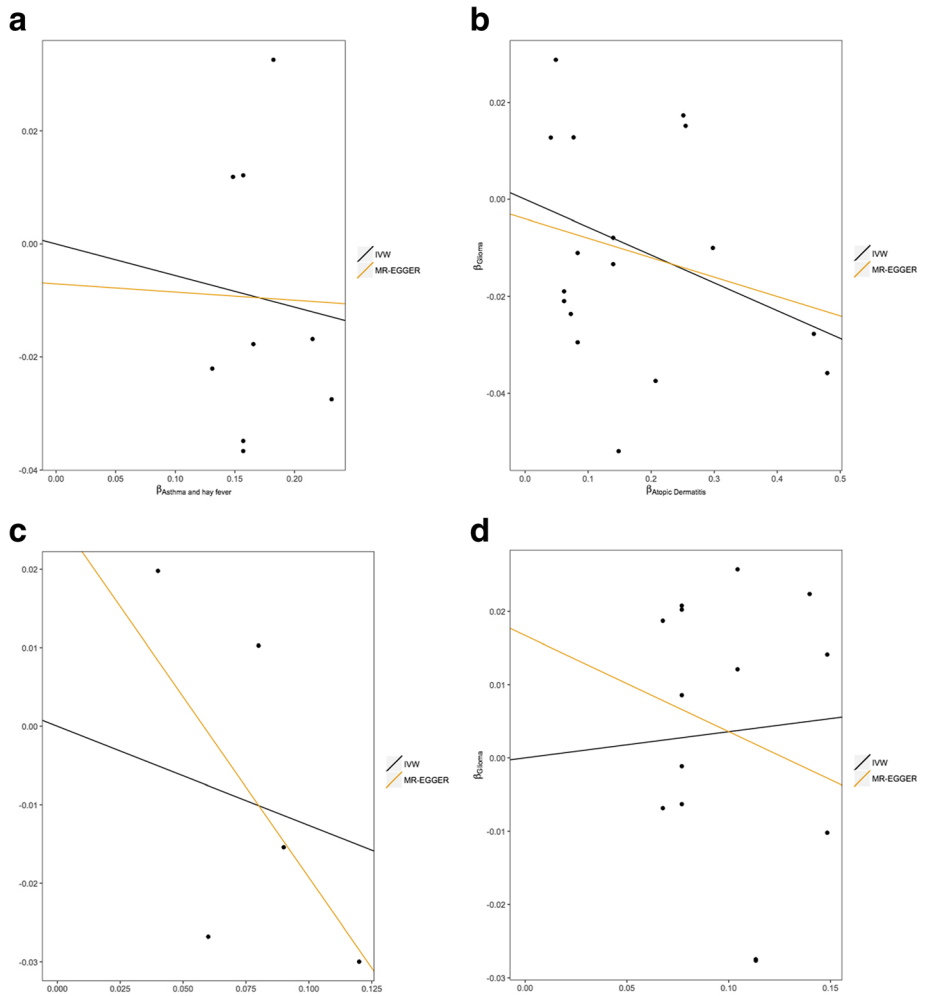


Fig. 3 Scatter plots of genetic associations with glioma against genetic associations with the exposure. **a** Asthma and hay fever, **b** atopic dermatitis, **c** IgE level, **d** self-reported allergy

levels, or self-reported allergy is protective against the risk of developing glioma.

Possible mechanisms explaining an observed inverse relation between the risk of atopic dermatitis and the risk of glioma have been suggested in previous papers [12], postulated to be the consequence of immune system hyperactivity. The question thus arises as to how such divergent findings for other atopic traits can be explained or reconciled, when they have been previously reported in high numbers.

A key assumption in MR is that the instrument affects glioma risk through its effect on a specific phenotype/exposure (*i.e.* atopic traits), and does not have a direct effect on glioma risk. We tested this assumption using MR-Egger regression and HEIDI-outlier analysis and found possible evidence of violation of this assumption for IgE and self-reported allergy. It is notable that self-reported allergy does not show an approximately quadratic response to correlation, in contrast to asthma and hay fever, atopic dermatitis and IgE level. This is likely to be a consequence of imprecise estimates of the association between SNPs and allergy, illustrating the inherent issue in attempting to make use of self-reported allergy data as an atopy-related trait.

The meta-analyses of published epidemiological observational studies has indeed provided strong evidence for an inverse relationship between atopy and glioma risk [40]. However, most of the support for such a relationship came from case-control studies [4]. A common limitation in retrospective studies of glioma has been the use of proxy respondents for patients with cognitive impairment, who may not remember past exposures accurately due to cognitive deficits [4]. Such issues are compounded by the fact that, across studies, multiple atopic traits have been assessed. The strength of support for a relationship seen across case-control studies contrasts markedly with the limited evidence for a relationship from prospective cohort-based analyses [7].

By inference, a relationship between long-term antihistamine use could theoretically provide supporting evidence, albeit indirect, that atopic-mediated mechanisms influence glioma risk. However, the impact of antihistamine use is difficult to disentangle from that of allergies, as these factors are highly correlated and few individuals without allergies use antihistamines regularly. Paradoxically, an increased risk for glioma associated with antihistamines, particularly among individuals with allergic conditions, has been found in some studies [41, 42].

Raised IgE levels and self-reported allergy suffer limitations as traits used to assess the effect of atopy on glioma risk as they are both variable over short time scales in their level of expression (in contrast to clinical diagnosis of atopic dermatitis). Further, allergies may develop later in life, and patients may not necessarily exhibit symptoms. This introduces the possibility of bias and error due to the time varying association of SNPs with the exposure. However, it has been suggested that seasonality does not have a significant effect [11].

An additional possible explanation for the lack of causal association between IgE levels and glioma risk seen in this study is that the causality is in fact reversed, which could result in epidemiological observational studies reporting inverse relationships [8, 9], but would not affect an MR analysis. Immunosuppression caused by glioblastoma is well documented [43, 44] and may lead to reduced expression of atopy. Furthermore, in addition to steroids, temozolomide therapy, routinely used to treat GBM nowadays, leads to reduced blood IgE levels [11].

Using data from large genetic consortia for multiple atopy-related traits and glioma risk has enabled us to more precisely test our study hypotheses than if we had used individual-level data from a smaller study. Through simulation scenarios, the IVW, MLE, WME, MBE and MR-Egger methods have been demonstrated to accurately estimate causal effects using summary-level data [28, 30, 31, 45]. However, using summary-level data instead of individual-level data limits the approaches that can be used to test the validity of genetic variants as IVs, as adjusting for measured covariates and assessing gene-environment interactions is generally not possible using summary-level data [46]. The first-stage F statistic was large (> 25 for all traits), and therefore weak instrument bias is unlikely.

Epidemiological observational studies have reported inverse relationships between atopy-related traits and glioma risk, with ORs in the range 0.43–0.96 for asthma [6, 47], 0.42–0.90 for atopic dermatitis [6, 47], 0.37–0.73 for IgE levels [8–10] and 0.47–0.69 for self-reported allergies [4, 5, 8]. Odds ratios for binary exposures estimated in this MR study represent the OR for the outcome disease per doubling in odds of the exposure disease, and the magnitudes of these causal effect estimates are therefore not directly comparable to those reported in observational studies.

Our MR analysis has several strengths. Firstly, by utilising the random allocation of genetic variants, we were able to overcome potential confounding and reverse causation that may bias estimates from observational studies. Secondly, given that a poor outcome from glioma is almost universal, it is unlikely that survival bias will have influenced study findings. Lastly, the findings from this study represent the association of a lifelong atopy with glioma in the general European population.

Nevertheless, our study does have limitations. Firstly, while it is entirely appropriate to implement different MR methods to assess the robustness of findings, they have a differing power to demonstrate associations, with the WME, MBE and MR-Egger methods having less power than IVW and MLE. Irrespective of such factors, our study only had 80% power to detect ORs of 1.16, 1.09, 1.16 and 1.22 for asthma and hay fever, atopic dermatitis, IgE level and self-reported allergy, respectively (Additional file 4: Table S4), due to the very low proportion of variability in the atopy-related traits explained by the SNPs used. Hence, we cannot exclude the possibility that these traits influence glioma risk, albeit modestly. To explore this possibility, will require additional IVs and larger sample sizes affording increased power. Furthermore, it is possible that an effect of atopy on glioma risk might be mediated through mechanisms associated with a trait that we have not captured by using MR to assess asthma and hay fever and self-reported allergy. Secondly, a weakness of the two-sample MR strategy is that it does not allow examination of non-linear relationships between exposures and outcomes. Finally, we have sought to examine whether bias could be introduced when considering a binary exposure for a binary outcome. Although in our simulation study we found no evidence of bias when estimating non-causal relationships, we did not extend our analysis to consider the potential impact of invalid SNPs.

Conclusions

In conclusion, our investigation does not provide strong evidence for a relationship between atopy-related diseases and risk of developing glioma, but findings do not preclude a small effect for atopic dermatitis. Our analysis also serves to illustrate the value of using several MR methods to derive robust conclusions.

Additional files

Additional file 1: Figure S1. Forest plot of Wald odds ratios (ORs) and 95% confidence intervals generated from single nucleotide polymorphisms (SNPs) associated with atopic dermatitis, including

rs909341. ORs for individual SNPs are listed according to magnitude of effect in the instrumental variable analysis and are presented with pooled effects using the inverse-variance weighting method. Squares represent the point estimate, and the bars are the 95% confidence intervals. (DOCX 89 kb)

Additional file 2: Table S1. Summary of the eight glioma genome-wide association studies. (XLSX 29 kb)

Additional file 3: Table S2. Table of single nucleotide polymorphisms (SNPs) reported in the NHGRI-EBI Genome-wide Association Studies Catalog for each trait, with correlations between SNPs. (XLSX 48 kb)

Additional file 4: Table S3. Percentage of variance explained by the combined sets of single nucleotide polymorphisms used as instrumental variables. (XLSX 33 kb)

Additional file 5: Table S4. Range of odds ratios for which study had < 80% power, for each atopy-related trait ($P = 0.05$, two-sided). (XLSX 9 kb)

Additional file 6: Table S5. Simulation analyses. (XLSX 28 kb)

Additional file 7: Table S6. Inverse-variance weighting, maximum likelihood estimation, weighted median estimate, mode-based estimate and Mendelian randomisation-Egger test results for combined atopy-related instrumental variables and glioma subtypes. (XLSX 39 kb)

Abbreviations

CI: confidence interval; GBM: glioblastoma; GWAS: genome-wide association study; IV: instrumental variable; IVW: inverse-variance weighting; MBE: mode-based estimate; MLE: maximum likelihood estimation; MR: Mendelian randomisation; OR: odds ratio; SNP: single nucleotide polymorphism; WHO: World Health Organization; WME: weighted median estimate

Acknowledgements

Not applicable.

Funding

LD-H was supported by a Wellcome Trust Summer Student bursary. AS is supported by a Cancer Research UK clinical Fellowship. In the UK, funding was provided by Cancer Research UK (C1298/A8362) supported by the Bobby Moore Fund. The Glioma International Case-Control Consortium Study was supported by grants from the National Institutes of Health, Bethesda, Maryland (R01CA139020, R01CA52689, P50097257, P30CA125123). The UK Interphone Study was supported by the European Commission Fifth Framework Program "Quality of Life and Management of Living Resources" and the UK Mobile Telecommunications and Health Programme. The Mobile Manufacturers Forum and the GSM Association provided funding for the study through the scientifically independent International Union against Cancer (UICC).

Availability of data and materials

Genotype data from the Glioma International Case-Control Consortium Study GWAS are available from the database of Genotypes and Phenotypes (dbGaP) under accession phs001319.v1.p1. Additionally, genotypes from the GliomaScan GWAS can be accessed through dbGaP accession phs000652.v1.p1.

Authors' contributions

RSH and AJC managed the project. LD-H, AJC, AS, PJL and RSH drafted the manuscript. LD-H and AJC performed statistical analyses. BK, KL, MJS and RSH acquired and analysed the UK data. MSi, PH, MMN and K-HJ acquired and analysed the German data. DIJ, QTO, JEE-P, GNA, EBC, DI, JS, JSB-S, SHO, JLB, RKL, CJ, RBJ, BSM, MRW, MLB and RSH acquired and analysed the Glioma International Case-Control Consortium Study data. SC and PR acquired and analysed the National Cancer Institute data. MSa acquired and analysed the French data. All authors reviewed the final manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Two-sample Mendelian randomisation was undertaken using GWAS data. Ethical approval was not sought for this specific project because all data came from the summary statistics of published GWAS, and no individual-level data were used.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Division of Genetics and Epidemiology, The Institute of Cancer Research, 15 Cotswold Road, London SM2 5NG, UK. [2]Department of Medicine, Section of Epidemiology and Population Sciences, Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA. [3]Case Comprehensive Cancer Center, School of Medicine, Case Western Reserve University, Cleveland, OH, USA. [4]Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, MN, USA. [5]School of Public Health, Yale University, New Haven, CT, USA. [6]Department of Neurosurgery, Brigham and Women's Hospital, Boston, MA, USA. [7]Department of Epidemiology and Biostatistics, School of Public Health, Georgia State University, Atlanta, GA, USA. [8]Duke Cancer Institute, Duke University Medical Center, Durham, NC, USA. [9]Cancer Control and Prevention Program, Department of Community and Family Medicine, Duke University Medical Center, Durham, NC, USA. [10]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [11]Departments of Neurology and Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [12]Department of Neurosurgery, University of Bonn Medical Center, Sigmund-Freud Str. 25, 53105 Bonn, Germany. [13]Human Genomics Research Group, Department of Biomedicine, University of Basel, Basel, Switzerland. [14]Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany. [15]Institute of Human Genetics, University of Bonn School of Medicine & University Hospital Bonn, Bonn, Germany. [16]Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen, University of Duisburg-Essen, Essen, Germany. [17]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, USA. [18]Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen, Denmark. [19]Rigshospitalet, University of Copenhagen, Copenhagen, Denmark. [20]Department of Laboratory Medicine and Pathology, Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, MN, USA. [21]Department of Radiation Sciences, Umeå University, Umeå, Sweden. [22]Department of Neurological Surgery, School of Medicine, University of California, San Francisco, San Francisco, CA, USA. [23]Institute of Human Genetics, University of California, San Francisco, CA, USA. [24]Sorbonne Universités UPMC Univ Paris 06, INSERM CNRS, U1127, UMR 7225, ICM, F-75013 Paris, France. [25]AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de neurologie 2-Mazarin, Paris, France. [26]Division of Molecular Pathology, The Institute of Cancer Research, London, UK.

Received: 24 October 2017 Accepted: 16 February 2018
Published online: 15 March 2018

References

1. Ostrom QT, Gittleman H, Xu J, Kromer C, Wolinsky Y, Kruchko C, Barnholtz-Sloan JS. CBTRUS Statistical Report: primary brain and other central nervous system tumors diagnosed in the United States in 2009-2013. Neuro Oncol. 2016;18(Suppl 5):v1–v75.
2. Ostrom QT, Bauchet L, Davis FG, Deltour I, Fisher JL, Langer CE, Pekmezci M, Schwartzbaum JA, Turner MC, Walsh KM, et al. The epidemiology of glioma in adults: a "state of the science" review. Neuro Oncol. 2014;16(7):896–913.
3. Wiemels JL, Wiencke JK, Sison JD, Miike R, McMillan A, Wrensch M. History of allergies among adults with glioma and controls. Int J Cancer. 2002;98(4):609–15.
4. Johansen C, Schüz J, Andreasen A-MS, Dalton SO. Study designs may influence results: the problems with questionnaire-based case–control studies on the epidemiology of glioma. Br J Cancer. 2017;116(7):841–8.
5. Schlehofer B, Blettner M, Preston-Martin S, Niehoff D, Wahrendorf J, Arslan A, Ahlbom A, Choi WN, Giles GG, Howe GR, et al. Role of

medical history in brain tumour development. Results from the international adult brain tumour study. Int J Cancer. 1999;82(2):155–60.
6. Cicuttini FM, Hurley SF, Forbes A, Donnan GA, Salzberg M, Giles GG, McNeil JJ. Association of adult glioma with medical conditions, family and reproductive history. Int J Cancer. 1997;71(2):203–7.
7. Schwartzbaum J, Jonsson F, Ahlbom A, Preston-Martin S, Lonn S, Soderberg KC, Feychting M. Cohort studies of association between self-reported allergic conditions, immune-related diagnoses and glioma and meningioma risk. Int J Cancer. 2003;106(3):423–8.
8. Wiemels JL, Wiencke JK, Patoka J, Moghadassi M, Chew T, McMillan A, Miike R, Barger G, Wrensch M. Reduced immunoglobulin E and allergy among adults with glioma compared with controls. Cancer Res. 2004; 64(22):8468–73.
9. Schlehofer B, Siegmund B, Linseisen J, Schuz J, Rohrmann S, Becker S, Michaud D, Melin B, Bas Bueno-de-Mesquita H, Peeters PH, et al. Primary brain tumours and specific serum immunoglobulin E: a case-control study nested in the European Prospective Investigation into Cancer and Nutrition cohort. Allergy. 2011;66(11):1434–41.
10. Calboli FC, Cox DG, Buring JE, Gaziano JM, Ma J, Stampfer M, Willett WC, Tworoger SS, Hunter DJ, Camargo CA Jr, et al. Prediagnostic plasma IgE levels and risk of adult glioma in four prospective cohort studies. J Natl Cancer Inst. 2011;103(21):1588–95.
11. Wiemels JL, Wilson D, Patel C, Patoka J, McCoy L, Rice T, Schwartzbaum J, Heimberger A, Sampson JH, Chang S, et al. IgE, allergy, and risk of glioma: update from the San Francisco Bay Area Adult Glioma Study in the temozolomide era. Int J Cancer. 2009;125(3):680–7.
12. Linos E, Raine T, Alonso A, Michaud D. Atopy and risk of brain tumors: a meta-analysis. J Natl Cancer Inst. 2007;99(20):1544–50.
13. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014; 23(R1):R89–98.
14. Meyers DA, Xu J, Postma DS, Levitt RC, Bleecker ER. Two locus segregation and linkage analysis for total serum IgE levels. Clin Exp Allergy. 1995;25:113–5.
15. Wilkinson J, Grimley S, Collins A, Simon Thomas N, Holgate ST, Morton N. Linkage of asthma to markers on chromosome 12 in a sample of 240 families using quantitative phenotype scores. Genomics. 1998;53(3):251–9.
16. Portelli MA, Hodge E, Sayers I. Genetic risk factors for the development of allergic disease identified by genome-wide association. Clin Exp Allergy. 2015;45(1):21–31.
17. Pierce BL, Burgess S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. Am J Epidemiol. 2013;178(7):1177–84.
18. Melin BS, Barnholtz-Sloan JS, Wrensch MR, Johansen C, Il'yasova D, Kinnersley B, Ostrom QT, Labreche K, Chen Y, Armstrong G, et al. Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. Nat Genet. 2017;49(5):789–94.
19. Granada M, Wilk JB, Tuzova M, Strachan DP, Weidinger S, Albrecht E, Gieger C, Heinrich J, Himes BE, Hunninghake GM, et al. A genome wide association study of plasma total IgE concentration in the Framingham Heart Study. J Allergy Clin Immunol. 2012;129(3):840–5.
20. Baurecht H, Hotze M, Brand S, Büning C, Cormican P, Corvin A, Ellinghaus D, Ellinghaus E, Esparza-Gordillo J, Fölster-Holst R, et al. Genome-wide comparative analysis of atopic dermatitis and psoriasis gives insight into opposing genetic mechanisms. Am J Hum Genet. 2015;96(1):104–20.
21. Paternoster L, Standl M, Waage J, Baurecht H, Hotze M, Strachan DP, Curtin JA, Bønnelykke K, Tian C, Takahashi A, et al. Multi-ethnic genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nat Genet. 2015;47(12):1449–56.
22. Schaarschmidt H, Ellinghaus D, Rodríguez E, Kretschmer A, Baurecht H, Lipinski S, Meyer-Hoffert U, Harder J, Lieb W, Novak N, et al. A genome-wide association study reveals 2 new susceptibility loci for atopic dermatitis. J Allergy Clin Immunol. 2015;136(3):802–6.
23. Weidinger S, Willis-Owen SAG, Kamatani Y, Baurecht H, Morar N, Liang L, Edser P, Street T, Rodriguez E, O'Regan GM, et al. A genome-wide association study of atopic dermatitis identifies loci with overlapping effects on asthma and psoriasis. Hum Mol Genet. 2013; 22(23):4841–56.
24. Ferreira MA, Matheson MC, Tang CS, Granell R, Ang W, Hui J, Kiefer AK, Duffy DL, Baltic S, Danoy P, et al. Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. J Allergy Clin Immunol. 2014;133(6):1564–71.
25. Ramasamy A, Curjuric I, Coin LJ, Kumar A, McArdle WL, Imboden M, Leynaert B, Kogevinas M, Schmid-Grendelmeier P, Pekkanen J, et al. A genome-wide meta-analysis of genetic variants associated with allergic rhinitis and grass sensitization and their interaction with birth order. J Allergy Clin Immunol. 2011;128(5):996–1005.
26. Hinds DA, McMahon G, Kiefer AK, Do CB, Eriksson N, Evans DM, St Pourcain B, Ring SM, Mountain JL, Francke U, et al. A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. Nat Genet. 2013;45(8):907–11.
27. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, Ohgaki H, Wiestler OD, Kleihues P, Ellison DW. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. 2016;131(6):803–20.
28. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol. 2013;37(7):658–65.
29. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, Robinson M, McGrath J, Visscher P, Wray N, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat Commun. 2018;9(1):224.
30. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epidemiol. 2016;40(4):304–14.
31. Hartwig FP, Smith GD, Bowden J. Robust inference in summary data Mendelian randomization via the zero model pleiotropy assumption. Int J Epidemiol. 2017;46(6):1985–98.
32. Gage SH, Jones HJ, Burgess S, Bowden J, Davey Smith G, Zammit S, Munafo MR. Assessing causality in associations between cannabis use and schizophrenia risk: a two-sample Mendelian randomization study. Psychol Med. 2017;47(5):971–80.
33. Burgess S. Sample size and power calculations in Mendelian randomization with a single instrumental variable and a binary outcome. Int J Epidemiol. 2014;43(3):922–9.
34. Paternoster L, Standl M, Johannes W, Baurecht H, Hotze M, Strachan DP, Curtin JA. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nat Genet. 2015;47(12):1449–56.
35. Weidinger S, Gieger C, Rodriguez E, Baurecht H, Mempel M, Klopp N, Gohlke H, Wagenpfeil S, Ollert M, Ring J, et al. Genome-wide scan on total serum IgE levels identifies FCER1A as novel susceptibility locus. PLoS Genet. 2008; 4(8):e1000166.
36. Ramasamy A, Kuokkanen M, Vedantam S, Gajdos ZK, Couto Alves A, Lyon HN, Ferreira MAR, Strachan DP, Zhao JH, Abramson MJ, et al. Genome-wide association studies of asthma in population-based cohorts confirm known and suggested loci and identify an additional association near HLA. PLoS One. 2012;7(9):e44008.
37. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896–901.
38. Ahmad OS, Morris JA, Mujammami M, Forgetta V, Leong A, Li R, Turgeon M, Greenwood CM, Thanassoulis G, Meigs JB, et al. A Mendelian randomization study of the effect of type-2 diabetes on coronary heart disease. Nat Commun. 2015;6:7060.
39. Palmer TM, Sterne JA, Harbord RM, Lawlor DA, Sheehan NA, Meng S, Granell R, Smith GD, Didelez V. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. Am J Epidemiol. 2011;173(12):1392–403.
40. Amirian ES, Zhou R, Wrensch MR, Olson SH, Scheurer ME, Il'yasova D, Lachance D, Armstrong GN, McCoy LS, et al. Approaching a scientific consensus on the association between allergies and glioma risk: a report from the Glioma International Case-Control Study. Cancer Epidemiol Biomarkers Prev. 2016;25(2):282–90.
41. Scheurer ME, El-Zein R, Thompson PA, Aldape KD, Levin VA, Gilbert MR, Weinberg JS, Bondy ML. Long-term anti-inflammatory and antihistamine medication use and adult glioma risk. Cancer Epidemiol Biomarkers Prev. 2008;17(5):1277–81.

42. Amirian ES, Marquez-Do D, Bondy ML, Scheurer ME. Antihistamine use and immunoglobulin E levels in glioma risk and prognosis. Cancer Epidemiol. 2013;37(6):908–12.
43. Razavi S-M, Lee KE, Jin BE, Aujla PS, Gholamin S, Li G. Immune evasion strategies of glioblastoma. Front Surg. 2016;3:11.
44. Gustafson MP, Lin Y, New KC, Bulur PA, O'Neill BP, Gastineau DA, Dietz AB. Systemic immune suppression in glioblastoma: the interplay between CD14(+)HLA-DR(lo/neg) monocytes, tumor factors, and dexamethasone. Neuro Oncol. 2010;12(7):631–44.
45. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015;44(2):512–25.
46. Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. Epidemiology. 2017;28(1):30–42.
47. Il'yasova D, McCarthy B, Marcello J, Schildkraut JM, Moorman PG, Krishnamachari B, Ali-Osman F, Bigner DD, Davis F. Association between glioma and history of allergies, asthma, and eczema: a case-control study with three groups of controls. Cancer Epidemiol Biomarkers Prev. 2009;18(4):1232–8.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit


BioMed Central

BJC British Journal of Cancer

www.nature.com/bjc

ARTICLE

Epidemiology

Influence of obesity-related risk factors in the aetiology of glioma

Linden Disney-Hogg[1], Amit Sud[1], Philip J. Law[1], Alex J. Cornish[1], Ben Kinnersley[1], Quinn T. Ostrom[2], Karim Labreche[1], Jeanette E. Eckel-Passow[3], Georgina N. Armstrong[4], Elizabeth B. Claus[5,6], Dora Il'yasova[7,8,9], Joellen Schildkraut[8,9], Jill S. Barnholtz-Sloan[3], Sara H. Olson[10], Jonine L. Bernstein[10], Rose K. Lai[11], Anthony J. Swerdlow[1,12], Matthias Simon[13], Per Hoffmann[14,15], Markus M. Nöthen[15,16], Karl-Heinz Jöckel[17], Stephen Chanock[18], Preetha Rajaraman[18], Christoffer Johansen[19,20], Robert B. Jenkins[21], Beatrice S. Melin[22], Margaret R. Wrensch[23,24], Marc Sanson[25,26], Melissa L. Bondy[4] and Richard S. Houlston[1,27]

BACKGROUND: Obesity and related factors have been implicated as possible aetiological factors for the development of glioma in epidemiological observation studies. We used genetic markers in a Mendelian randomisation framework to examine whether obesity-related traits influence glioma risk. This methodology reduces bias from confounding and is not affected by reverse causation.

METHODS: Genetic instruments were identified for 10 key obesity-related risk factors, and their association with glioma risk was evaluated using data from a genome-wide association study of 12,488 glioma patients and 18,169 controls. The estimated odds ratio of glioma associated with each of the genetically defined obesity-related traits was used to infer evidence for a causal relationship.

RESULTS: No convincing association with glioma risk was seen for genetic instruments for body mass index, waist-to-hip ratio, lipids, type-2 diabetes, hyperglycaemia or insulin resistance. Similarly, we found no evidence to support a relationship between obesity-related traits with subtypes of glioma–glioblastoma (GBM) or non-GBM tumours.

CONCLUSIONS: This study provides no evidence to implicate obesity-related factors as causes of glioma.

British Journal of Cancer https://doi.org/10.1038/s41416-018-0009-x

INTRODUCTION

Glioma is the most common primary intracranial tumour, accounting for around 80% of all malignant brain tumours.[1] Thus far, few established risk factors for the development of glioma have been robustly identified.[2]

Obesity-related factors are increasingly being recognised as risk determinants for the development many of common cancers, such as those of the breast and colorectum.[3] Evidence from epidemiological observational studies, for obesity-related traits being a risk factor for the development of glioma have, however been inconsistent, with only a subset of studies reporting a significant association.[4–9] Furthermore, in contrast to most cancers, some studies have reported diabetes to be protective against glioma.[10–13] Obesity-related exposures are however inherently interrelated,[14, 15] and in traditional epidemiological studies it can be problematic to isolate specific risk factors that may exert a causal influence on disease from those that are merely associated with an underlying causal factor (i.e.

[1]Division of Genetics and Epidemiology, The Institute of Cancer Research, London SW7 3RP, UK; [2]Case Comprehensive Cancer Center, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA; [3]Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, MI 55905, USA; [4]Section of Epidemiology and Population Sciences, Department of Medicine, Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA; [5]School of Public Health, Yale University, New Haven, CT 06510, USA; [6]Department of Neurosurgery, Brigham and Women's Hospital, Boston, MA 02115, USA; [7]Department of Epidemiology and Biostatistics, School of Public Health, Georgia State University, Atlanta, GA 30303, USA; [8]Duke Cancer Institute, Duke University Medical Center, Durham, NC 27710, USA; [9]Cancer Control and Prevention Program, Department of Community and Family Medicine, Duke University Medical Center, Durham, NC 27710, USA; [10]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10017, USA; [11]Departments of Neurology and Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA; [12]Division of Breast Cancer Research, The Institute of Cancer Research, London SW7 3RP, UK; [13]Department of Neurosurgery, University of Bonn Medical Center, Sigmund-Freud-Str. 25, Bonn 53105, Germany; [14]Human Genomics Research Group, Department of Biomedicine, University of Basel, Basel 4031, Switzerland; [15]Department of Genomics, Life & Brain Center, University of Bonn, Bonn 53127, Germany; [16]Institute of Human Genetics, University of Bonn School of Medicine and University Hospital Bonn, Bonn 53127, Germany; [17]Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen, University of Duisburg-Essen, Essen 45147, Germany; [18]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA; [19]Danish Cancer Society Research Center, Survivorship, Danish Cancer Society, Copenhagen 2100, Denmark; [20]Oncology Clinic, Finsen Centre, Rigshospitalet, University of Copenhagen, Copenhagen 2100, Denmark; [21]Department of Laboratory Medicine and Pathology, Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, MI 55905, USA; [22]Department of Radiation Sciences, Umeå University, Umeå 901 87, Sweden; [23]Department of Neurological Surgery, School of Medicine, University of California, San Francisco, CA 94143, USA; [24]Institute of Human Genetics, University of California, San Francisco, CA 94143, USA; [25]Sorbonne Universités UPMC Univ Paris 06 INSERM CNRS, U1127, UMR 7225, ICM, Paris 75013, France; [26]AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de Neurologie 2-Mazarin, Paris 75013, France and [27]Division of Molecular Pathology, The Institute of Cancer Research, London SW7 3RP, UK
Correspondence: Richard S. Houlston (richard.houlston@icr.ac.uk)

Received: 16 October 2017 Revised: 5 January 2018 Accepted: 8 January 2018
Published online: 13 March 2018

© The Author(s) 2018

Published by Springer Nature on behalf of Cancer Research UK

Table 1. Metabolic risk factors for which genetic instruments were developed and evaluated in relation to disease risk

Trait	SNPs[a]	Mean (SD)	Units	PVE (%)	References
Two hour post-challenge glucose	7	5.6 (1.7)	mmol/l	1.7	24
BMI	75	27.0 (4.6)	kg/m^2	2.4	21
Fasting glucose	33	5.2 (0.8)	mmol/l	4.8	24
Fasting insulin	12	56.9 (44.4)	pmol/l	1.2	24
HDL cholesterol	54	53.3 (15.5)	mg/dl	13.7	23
LDL cholesterol	26	133.6 (38.0)	mg/dl	14.6	23
Type-2 diabetes	34	—	—	1.6	25
Total cholesterol	37	213.3 (42.6)	mg/dl	15.0	23
Triglycerides	24	140.9 (87.8)	mg/dl	11.7	23
WHR	33	1.1 (0.1)	cm/cm	0.7	22

BMI body mass index, *HDL* high-density lipoprotein, *LDL* low-density lipoprotein, *PVE* proportion of variance explained, *SD* standard deviation, *SNP* single-nucleotide polymorphism, *WHR* waist–hip ratio
[a]Number of SNPs used after quality control

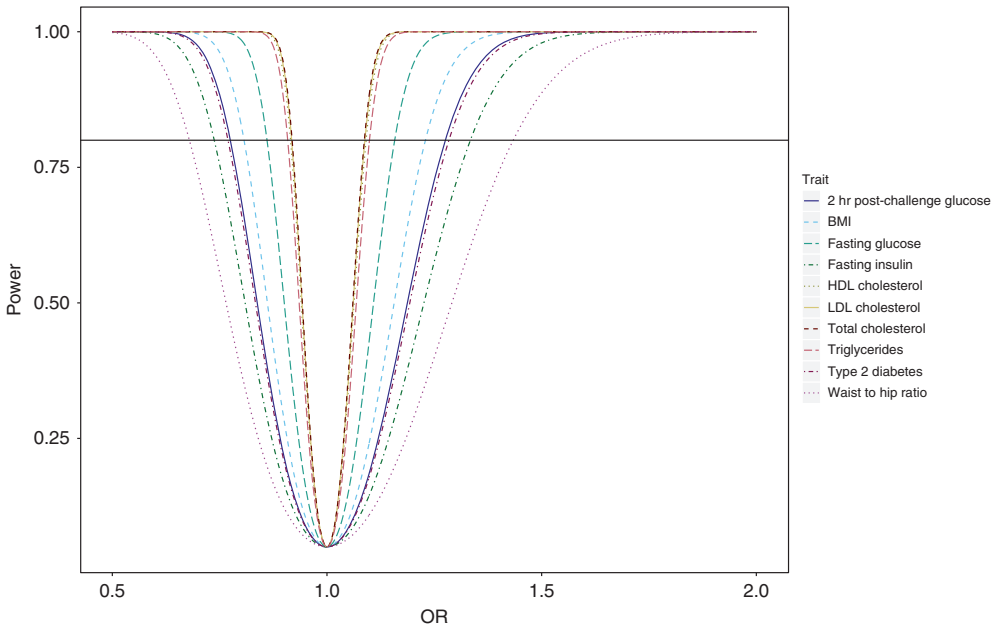


Fig. 1 Study power against OR for each obesity-related trait and all glioma ($P = 0.05$, two-sided). A line indicating a power of 80% is shown. BMI body mass index, HDL high-density lipoprotein, LDL low-density lipoprotein, OR odds ratio

confounded). In addition, findings can be affected by reverse causation.

Mendelian randomisation (MR) is an analytical approach to the traditional epidemiological study whereby genetic markers are used as proxies or instrumental variables (IVs) of environmental and lifestyle-related risk factors.[16] Such genetic markers cannot be influenced by reverse causation and can act as unconfounded markers of exposures provided the variants are not associated with the disease through an alternative mechanism.[16] Under these circumstances, the association between a genetic variant (or set of variants) and outcome of interest implies a causal relationship between the risk factor and outcome. MR has therefore been compared to a natural randomised controlled trial, circumventing some of the limitations of epidemiological observational studies.[17] However, as IVs used in MR often explain a small proportion of the exposure phenotypic variance, large sample sizes are required to have sufficient power.[18]

To gain insight into the aetiology of glioma, we have examined the role of obesity-related risk factors in glioma using an MR-based framework. Specifically, we identified genetic variants associated with 10 key obesity-related risk factors from external genetic association studies. We implemented two-sample MR[19] to estimate associations between these genetic variants with glioma risk using genome-wide association study (GWAS) data from the Glioma International Case-Control Consortium study (GICC).[20]

MATERIALS AND METHODS

Two-sample MR was undertaken using GWAS data. Ethical approval was not sought for this specific project because all data came from the summary statistics of published GWAS, and no individual-level data were used.

Genetic instruments for obesity and related risk factors

Genetic instruments were identified as a panel of single-nucleotide polymorphisms (SNPs) identified from recent meta-analyses or largest studies published to date. Specifically: (i) SNPs for body mass index (BMI) and waist-to-hip ratio (WHR) were

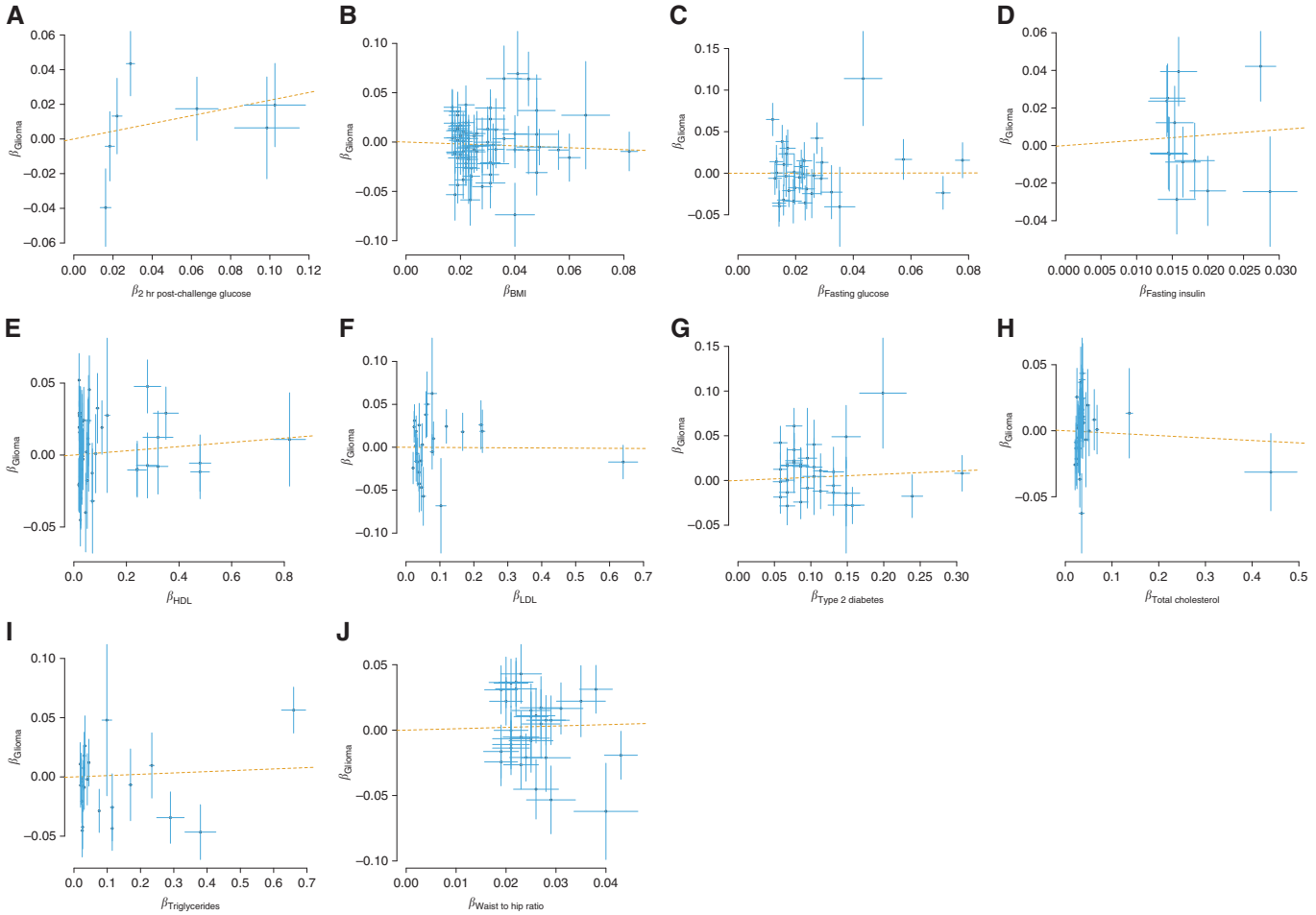


Fig. 2 SNP-specific effects for risk of all glioma. For each figure, the effect size of the respective measure for: **a** 2-h post-challenge glucose, **b** BMI, **c** fasting glucose, **d** fasting insulin, **e** HDL cholesterol, **f** LDL cholesterol, **g** type-2 diabetes, **h** total cholesterol, **i** triglycerides and **j** WHR is plotted against the effect for all glioma. Error bars represent one SD. The GSMR estimate is plotted as a dashed line for reference. BMI body mass index, GSMR generalised summary data-based Mendelian randomisation, HDL high-density lipoprotein, LDL low-density lipoprotein, SD standard deviation, WHR waist–hip ratio

identified from the Genetic Investigation of ANthropometric Traits (GIANT) consortium;[21, 22] (ii) SNPs for circulating high-density and low-density lipoprotein cholesterol (HDL and LDL), total cholesterol and triglycerides, were identified from the Global Lipids Genetic Consortium (GLGC);[23] (iii) SNPs for factors related to hyperglycaemia and hyperinsulinemia—fasting glucose, fasting insulin and 2-h post-challenge glucose, were obtained from the Meta-Analysis of Glucose and Insulin related traits Consortium (MAGIC)[24] and (iv) SNPs for type-2 diabetes were identified from.[25] For each SNP, we recovered the chromosome position, the effect estimate expressed in standard deviations (SD) of the trait per-allele along with the corresponding standard error (Supplementary Table 1). We restricted our analysis to SNPs associated at genome-wide significance (i.e. $P \leq 5.0 \times 10^{-8}$) in individuals with European ancestry. To avoid co-linearity between SNPs for each trait, we excluded SNPs that were correlated (i.e. $r^2 \geq 0.01$) within each trait, and only considered the SNPs with the strongest effect on the trait for inclusion in genetic risk scores (Supplementary Table 2). For type-2 diabetes, linkage disequilibrium (LD) scores with rs140730081 were calculated via a proxy SNP rs2259835 ($r^2 = 0.48$). After imposing these criteria, we obtained 7 SNPs for 2-h post-challenge glucose, 75 for BMI, 33 for fasting glucose, 13 for fasting insulin, 54 for HDL cholesterol, 26 for LDL cholesterol, 38 for type-2 diabetes, 39 for total cholesterol, 25 for triglycerides and 33 for WHR.

Glioma association results

To evaluate the association of each genetic instrument with glioma risk, we made use of data from the most recent meta-analysis of GWAS in glioma, comprising >10 million genetic variants (after imputation) in 12,488 glioma patients and 18,169 controls from eight independent GWAS data sets of individuals of European descent (Supplementary Table 3).[20] Comprehensive details of the genotyping and quality control of the seven GWAS have been previously reported.[20] To limit the effects of cryptic population stratification, association test statistics for six of the glioma GWAS were generated using principal components as previously detailed.[20] Gliomas are heterogeneous and different tumour subtypes, defined in part by malignancy grade (e.g. pilocytic astrocytoma World Health Organization (WHO) grade I, diffuse 'low-grade' glioma WHO grade II, anaplastic glioma WHO grade III and GBM WHO grade IV) can be distinguished.[26] For the sake of diagnostic brevity, we considered gliomas as being either GBM or non-GBM tumours.

Statistical analysis

The odds ratios (OR) of glioma per unit of SD increment for each obesity-related trait, were estimated using generalised summary data-based Mendelian randomisation (GSMR).[27] This approach performs a multi-SNP MR analysis, which is more powerful than other existing summary data-based MR methodologies.[28]

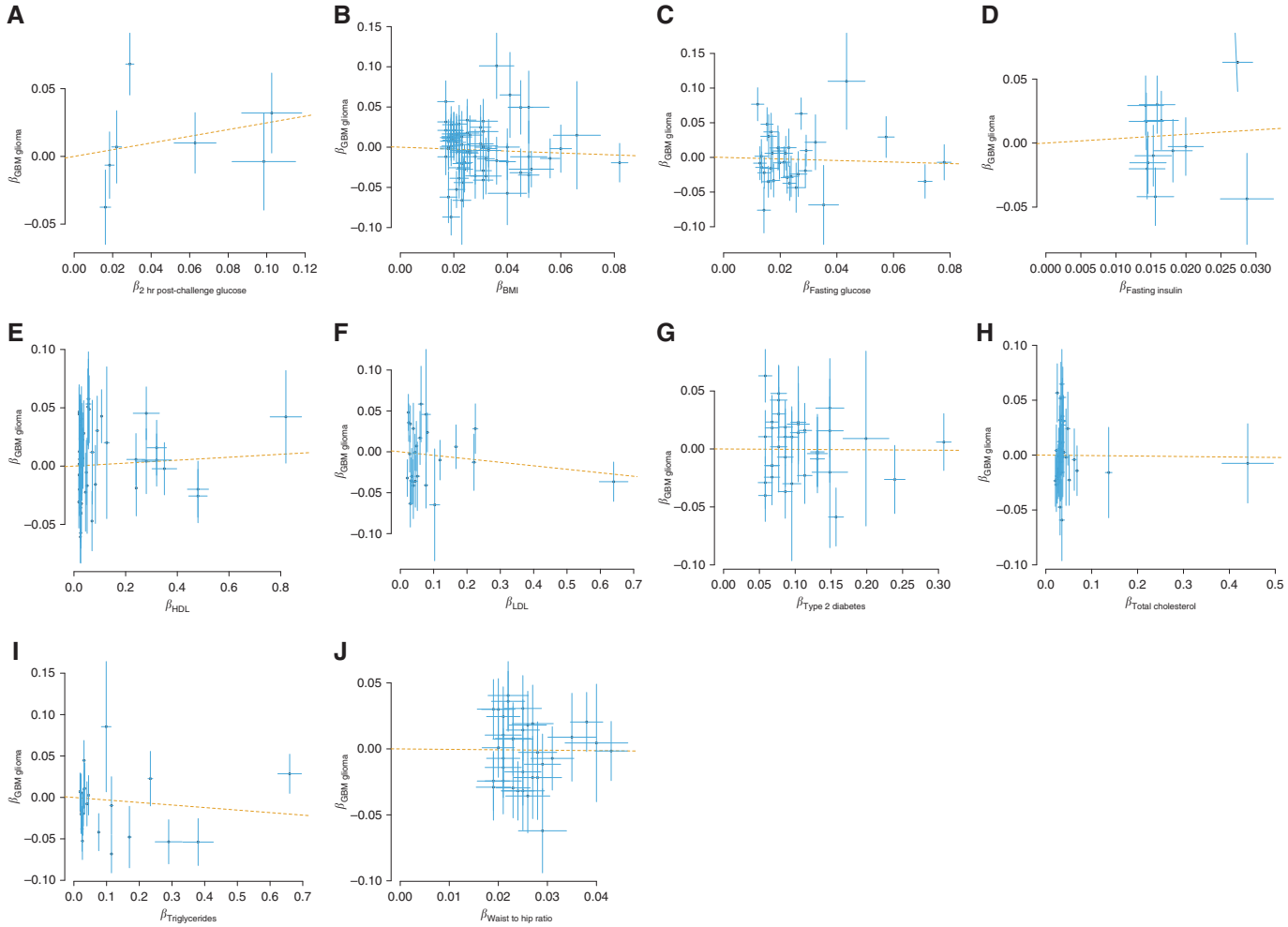


Fig. 3 SNP-specific effects for risk of GBM glioma. For each figure, the effect size of the respective measure for **a** 2-h post-challenge glucose, **b** BMI, **c** fasting glucose, **d** fasting insulin, **e** HDL cholesterol, **f** LDL cholesterol, **g** type-2 diabetes, **h** total cholesterol, **i** triglycerides and **j** WHR is plotted against the effect for GBM glioma. Error bars represent one SD. The GSMR estimate is plotted as a dashed line for reference. BMI body mass index, GBM glioblastoma mulitforme, GSMR generalised summary data-based Mendelian randomisation, HDL high-density lipoprotein, LDL low-density lipoprotein, SD standard deviation, WHR waist–hip ratio

Separation of signals of causality from horizontal pleiotropy (a single locus influencing affecting multiple phenotypes, also referred to as type-II pleiotropy) is a recognised issue in MR analyses and we therefore used a HEIDI-outlier test[27] to detect and eliminate genetic instruments that have apparent pleiotropic effects on both the obesity-related trait and glioma. A P value threshold of 0.01 for the HEIDI-outlier test was utilised as recommended by Zhu et al. The HEIDI-outlier test may also in theory detect additional violations of the assumptions of MR such as the exclusion restriction assumption. Given that glioma is a binary outcome and type-2 diabetes a binary exposure, the resulting causal effect estimate in this scenario represents the odds for glioma risk per unit increase in the log OR for type-2 diabetes.

For each statistical test, we considered a global significance level of $P < 0.05$ as being satisfactory to derive conclusions. To assess the robustness of our conclusions, we imposed a Bonferroni-corrected significance threshold of 0.0017 (i.e. 0.05/30, to correct for testing 10 traits over three outcomes). We considered a P value > 0.05 as non-significant (i.e. no association), a P value ≤ 0.05 as evidence for a potential causal association, and a P value ≤ 0.0017 as significant evidence for an association. Additionally, we defined the Bayesian false null probability (BFNP) using the Bayesian false discovery probability (BFDP) as per Wakefield[29] by BFNP = 1 − BFDP. Then to assess whether null results found could be considered reliable, we calculated the minimum prior probability of the alternative hypothesis for which the BFNP was >10%. The power of an MR investigation depends greatly on the proportion of variance in the risk factor that is explained by the respective IV. We estimated study power a priori using the methodology of Burgess.[30] Statistical analyses were undertaken using R software (Version 3.1.2).

RESULTS

In our data sets, there were missing data for one fasting insulin SNP (rs1530559), four type-2 diabetes SNPs (rs2972156, rs34706136, rs11257658, rs144613775) and one total cholesterol SNP (rs7570971). These SNPs were excluded from our analysis. Performing HEIDI-outlier analysis on the instruments for each trait

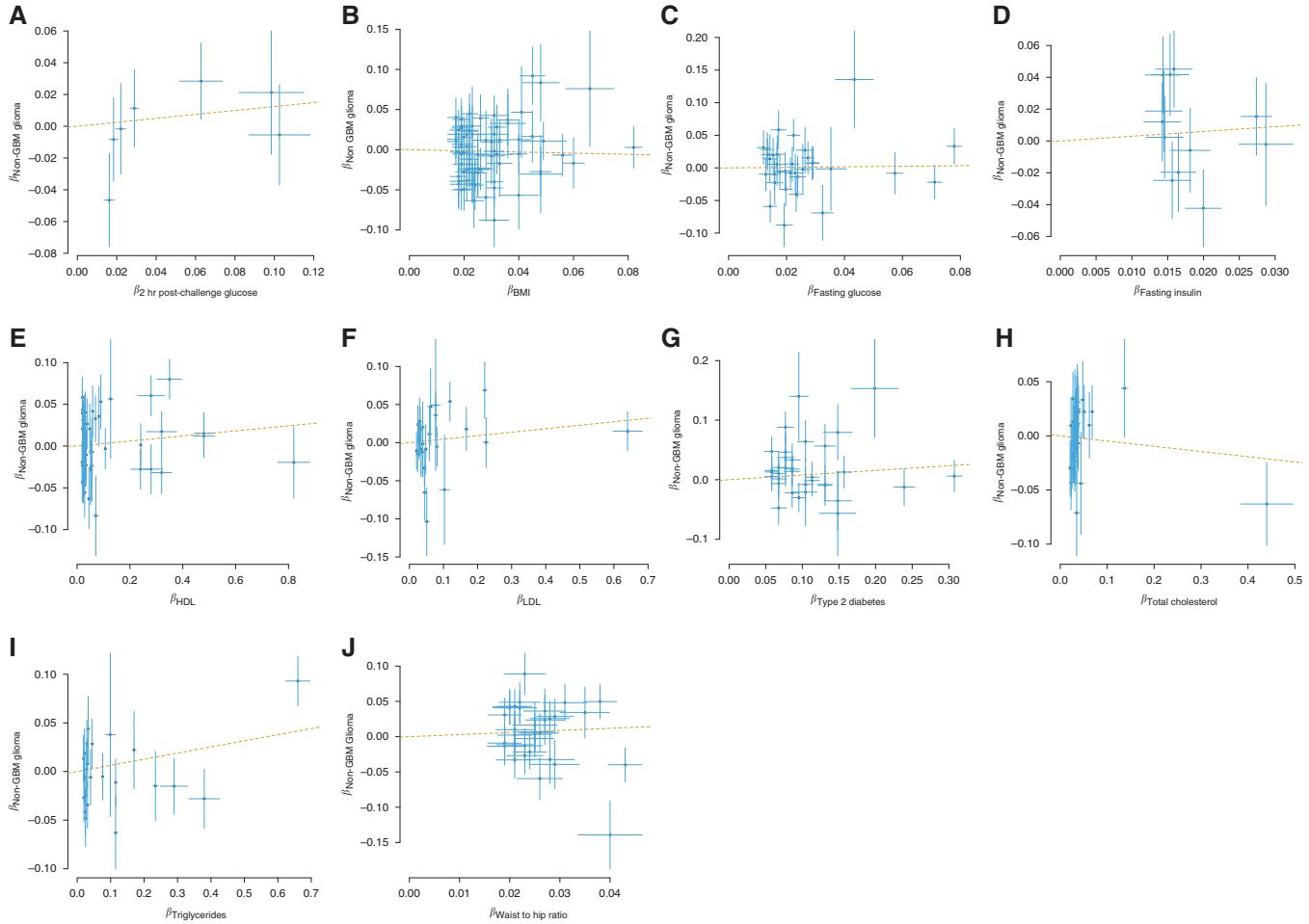


Fig. 4 SNP-specific effects for risk of non-GBM glioma. For each figure, the effect size of the respective measure for **a** 2-h post-challenge glucose, **b** BMI, **c** fasting glucose, **d** fasting insulin, **e** HDL cholesterol, **f** LDL cholesterol, **g** type-2 diabetes, **h** total cholesterol, **i** triglycerides and **j** WHR, is plotted against the effect for non-GBM glioma. Error bars represent one SD. The GSMR estimate is plotted as a dashed line for reference. BMI body mass index, GBM glioblastoma mulitforme, GSMR generalised summary data-based Mendelian randomisation, HDL high-density lipoprotein, LDL low-density lipoprotein, SD standard deviation, WHR waist–hip ratio

identified two SNPs as violating the assumptions of MR with respect to horizontal pleiotropy, rs11603023 for total cholesterol and rs5756931 for triglyceride, which were further excluded. Both SNPs are in LD with the lead SNP in glioma risk loci.

Subsequently, Table 1 details the number of SNPs used as an IV for each of the obesity-related traits, the mean and SD of the risk factor in the original discovery study, and the proportion of variance explained for each factor by the corresponding genetic instruments. Effect estimates for each SNP used as genetic instruments for each risk factor and disease risk are detailed in Supplementary Table 1. For BMI and LDL, the SNPs rs12016871 and rs9411489 have since merged with the SNPs rs9581854 and rs635634, respectively, and it is from these subsequent SNPs the associations with glioma were derived. Figure 1 shows the statistical power of genetic instruments for different levels of predicted ORs for each obesity-related trait.

Figure 2 shows a plot of the association of each IV with exposure against the association with glioma, together with the resulting GSMR estimate of the log OR. For each of the obesity-related traits under investigation, an approximately null estimate for effect was obtained, with the strongest association being shown by fasting insulin. Setting a threshold of $P \leq 0.05$, no statistically significant associations were shown for 2-h post-challenge glucose ($OR_{SD} = 1.25$, 95% confidence interval (CI) = 0.93–1.67), BMI ($OR_{SD} = 0.91$, 95% CI = 0.77–1.07), fasting glucose ($OR_{SD} = 1.00$, 95% CI = 0.78–1.3), fasting insulin ($OR_{SD} = 1.32$, 95% CI = 0.71–2.46), HDL cholesterol ($OR_{SD} = 1.01$, 95% CI = 0.98–1.05), LDL cholesterol ($OR_{SD} = 1.00$, 95% CI = 0.95–1.05), type-2 diabetes ($OR_{SD} = 1.04$, 95% CI = 0.97–1.11), total cholesterol ($OR_{SD} = 0.98$, 95% CI = 0.88–1.09), triglycerides ($OR_{SD} = 1.01$, 95% CI = 0.97–1.06) and WHR ($OR_{SD} = 1.11$, 95% CI = 0.84–1.46).

We explored the possibility that a relationship between an obesity-related trait and glioma might be subtype-specific, considering GBM and non-GBM separately. Figures 3 and 4 show corresponding plots of the association of each IV with exposure against the association with GBM and non-GBM glioma. The strongest association was provided by the relationship between increased triglyceride level and risk of non-GBM glioma ($OR_{SD} = 1.07$, 95% CI = 1.00–1.13, $P = 0.044$), albeit non-significant after adjustment for multiple testing (Table 2). Table 3 presents the minimum prior probabilities of an association required for each trait to have a BFNP ≥ 0.1. Where possible, the maximum likely OR has been taken from the largest value reported in observational studies.[7, 12, 31] In the event that this was not possible, an upper bound of 2 was chosen. If the 'true' maximum likely OR were lower, then the smallest required prior probability would in fact be

Table 2. GSMR results for the combined obesity-related IVs

Trait	All glioma		GBM		Non-GBM	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Two hour post-challenge glucose	1.25 (0.93–1.67)	0.132	1.28 (0.90–1.83)	0.173	1.13 (0.77–1.66)	0.525
BMI	0.91 (0.77–1.07)	0.247	0.89 (0.73–1.08)	0.237	0.93 (0.75–1.15)	0.510
Fasting glucose	1.00 (0.78–1.3)	0.974	0.89 (0.66–1.22)	0.484	1.04 (0.75–1.45)	0.809
Fasting insulin	1.32 (0.71–2.46)	0.374	1.41 (0.66–3.00)	0.377	1.35 (0.60–3.04)	0.471
HDL cholesterol	1.01 (0.98–1.05)	0.375	1.01 (0.97–1.05)	0.532	1.03 (0.99–1.08)	0.167
LDL cholesterol	1.00 (0.95–1.05)	0.939	0.96 (0.90–1.02)	0.197	1.05 (0.98–1.12)	0.195
Type-2 diabetes	1.04 (0.97–1.11)	0.290	1.00 (0.92–1.08)	0.933	1.08 (0.99–1.18)	0.076
Total cholesterol	0.98 (0.88–1.09)	0.736	1.00 (0.87–1.14)	0.949	0.95 (0.83–1.10)	0.505
Triglycerides	1.01 (0.97–1.06)	0.637	0.97 (0.92–1.03)	0.291	1.07 (1.00–1.13)	0.044
WHR	1.11 (0.84–1.46)	0.456	0.97 (0.69–1.35)	0.847	1.34 (0.94–1.93)	0.109

BMI body mass index, *CI* confidence interval, *GBM* glioblastoma multiforme, *GSMR* generalised summary data-based Mendelian randomisation, *HDL* high-density lipoprotein, *IV* instrumental variable, *LDL* low-density lipoprotein, *OR* odds ratio, *SD* standard deviation, *WHR* waist–hip ratio

Table 3. Prior probability of association required for BFNP > 0.1, for the combined obesity-related IVs

Trait	Glioma		References
	Maximum likely OR	Minimum required prior probability	
Two hour post-challenge glucose	2.00	0.10	N/A
BMI	1.27	0.11	8
Fasting glucose	1.57	0.18	31
Fasting insulin	2.00	0.12	N/A
HDL cholesterol	200	0.64	N/A
LDL cholesterol	2.00	0.61	N/A
Type-2 diabetes	0.60	0.31	12
Total cholesterol	2.00	0.41	N/A
Triglycerides	2.00	0.60	N/A
WHR	2.00	0.19	N/A

BFNP Bayesian false null probability, *BMI* body mass index, *HDL* high-density lipoprotein, *IV* instrumental variable, *LDL* low-density lipoprotein, *WHR* waist–hip ratio, *OR* odds ratio, *N/A* no observational data to inform maximum likely OR, value of 2 taken

lower. There is no current precedent for what value should be taken for the prior probability of an association, indeed attempting to sample published papers would produce an over estimation due to winners curse, but it is noted that a value of 10% would ensure all the results reported would have significance.

DISCUSSION

There is an abundance of studies that have implicated obesity and related traits (notably diabetes), as risk factors for all of the major common cancers, including breast, colorectal, oesophageal, pancreatic, ovarian and renal.[3] Furthermore, there is increasing evidence that obesity is likely to also be a risk factor for many of the less common tumours, such as those of the haematopoietic system.[3, 32] The mechanistic basis of how obesity and diabetes affects an increased cancer risk is poorly understood. The long-term metabolic consequences of obesity and its related traits are complex and several mechanisms have been suggested, including increased insulin and insulin-like growth factor signalling, chronic inflammation and signalling via adipokines.[33] Such mechanisms would be compatible with obesity and related traits having a generic effect on cancer risk.

Evidence for obesity influencing risk of glioma from previous observational studies has been mixed.[4, 6, 9] Intriguingly, in contrast to other cancers, an inverse relationship between both diabetes and increased HbA1c with risk of glioma has been reported in some but not all studies.[4–7, 9] Furthermore, in so far as it has been studied, anti-diabetic treatment has been reported to not influence glioma risk.[12] In terms of the wider spectrum of the metabolic syndrome, a study has linked elevated levels of triglyceride to risk of developing glioma.[9]

Our findings do not support a causal role for higher BMI and related metabolic risk factors, including diagnosis of type-2 diabetes and blood lipid levels, in influencing glioma risk. An important strength of our analysis is that by utilising the random allocation of genetic variants, we were able to overcome potential confounding, for example, from other interrelated traits.[14, 15] Furthermore, reverse causation and selection bias may have biased estimates from previously published observational studies. By exploiting data from large genetic consortia for multiple obesity-related traits and glioma risk has enabled us to more precisely test study hypotheses than if we had been reliant on individual-level data from a small study. The only obesity-related trait with a first-stage *F*-statistic <10 was WHR ($F = 6.75$) and therefore weak instrument bias for other traits is unlikely.[34] In addition, given that a poor outcome from glioma is almost universal, it is unlikely that survival bias will have influenced study findings materially. Finally, we have employed a Bayesian approach to interpret the significance of the null results while comparing our findings to published observational epidemiological studies. There is currently no precedent within the MR community as to what value is an accurate representation of the prior probability of association. If the true value is ~20%, then the null findings for 2 h post-challenge glucose, BMI, fasting glucose, fasting insulin and WHR all have a >10% chance of being false.

There are however potential limitations in our analysis that warrant further discussion. Firstly, the use of summary test statistics in two-sample MR analyses requires consideration of sample overlap, the winner's curse and genotype uncertainty.[35, 36] Sample overlap between the association studies of the exposure traits and outcome trait has the potential of inflating the type I error rate. The number of controls shared between the glioma GWAS and the anthropometric and lipid GWAS are, however <2% of the respective exposure sample size. Although we are unable to

calculate an exact number of glioma cases sampled in the exposure GWAS, given the lifetime risk of glioma is only 0.24%, very few numbers of glioma cases will have been analysed in the exposure trait studies. Hence, such sample overlap is unlikely to contribute to type I error rate inflation.[36] As the instrumental variables were discovered in the data used in this two-sample MR analysis, weak instrument bias will be accentuated due to winner's curse, thus attenuating the causal effect estimate towards the null.[36] Uncertainty with respect to genotyping or disease associations may diminish causal effect estimates.[36] However IVs used in this analysis are robust and only SNPs passing stringent quality control thresholds were used in the analysis. Secondly, MR is limited in the extent to which it can explore different life course models, such as when an exposure has a temporal relationship to the outcome risk.[35] Finally, our study does have limitations related to power. However, based on the relatively sizable fraction of variance explained by the genetic instruments for the majority of the obesity-related factors (Table 1), typically there was sufficient statistical power (>80%) to detect even modest odds ratios of 1.43, and close to complete statistical power (99%) to detect relative risks of 1.72 (Fig. 1).

In conclusion, our findings shed light on an issue for which the evidence to date has been mixed. Specifically, they provide evidence against obesity and related traits as significant risk factors for the development of glioma.

Availability of data and material

Genotype data from the GICC GWAS are available from the database of Genotypes and Phenotypes (dbGaP) under accession phs001319.v1.p1. In addition, genotypes from the GliomaScan GWAS can be accessed through dbGaP accession phs000652.v1.p1.

ACKNOWLEDGEMENTS

L.D.-H. was supported by a Wellcome Trust Summer Student bursary. A.S. is supported by a Cancer Research UK clinical fellowship and The Royal Marsden Hospital Haematology Research Fund. In the UK, funding was provided by Cancer Research UK (C1298/A8362 supported by the Bobby Moore Fund). The GICC was supported by grants from the National Institutes of Health, Bethesda, Maryland (R01CA139020, R01CA52689, P50097257, P30CA125123). The UK Interphone Study was supported by the European Commission Fifth Framework Program 'Quality of Life and Management of Living Resources' and the UK Mobile Telecommunications and Health Programme. The Mobile Manufacturers Forum and the GSM Association provided funding for the study through the scientifically independent International Union against Cancer (UICC).

AUTHOR CONTRIBUTIONS

R.S.H., A.S. and A.J.C. managed the project. L.D.-H., A.S., P.J.L., A.J.C. and R.S.H. drafted the manuscript. L.D.-H. performed statistical analyses. B.K., K.L., A.J.S. and R.H.S. provided UK data. M. Simon, P.H., M.M.N. and K.-H.J. provided German data. Q.T.O., J. E.E.-P., G.N.A., E.B.C., D.I., J.S., J.S.B.-S., S.H.O., J.L.B., R.K.L., C.J., R.B.J., B.S.M., M.R.W., M.L. B. and R.S.H. provided GICC data. S.C. and P.R. provided National Cancer Institute (NCI) data. M. Sanson provided French data. All authors reviewed the final manuscript.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41416-018-0009-x.

Competing interests: The authors declare no competing financial interests.

Ethics approval and consent to participate: Two-sample MR was undertaken using GWAS data. Ethical approval was not sought for this specific project because all data came from the summary statistics of published GWAS, and no individual-level data were used.

REFERENCES

1. Ostrom, Q. T. et al. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2006–2010. *NeuroOncology* **15**, ii1–56 (2013).
2. Ostrom, Q. T. et al. The epidemiology of glioma in adults: a "state of the science" review. *NeuroOncology* **16**, 896–913 (2014).
3. Kyrgiou, M. et al. Adiposity and cancer at major anatomical sites: umbrella review of the literature. *BMJ* **356**, j477 (2017).
4. Kaplan, S., Novikov, I. & Modan, B. Nutritional factors in the etiology of brain tumors: potential role of nitrosamines, fat, and cholesterol. *Am. J. Epidemiol.* **146**, 832–841 (1997).
5. Niedermaier, T. et al. Body mass index, physical activity, and risk of adult meningioma and glioma: a meta-analysis. *Neurology* **85**, 1342–1350 (2015).
6. Sergentanis, T. N. et al. Obesity and risk for brain/CNS tumors, gliomas and meningiomas: a meta-analysis. *PLoS ONE* **10**, e0136974 (2015).
7. Wiedmann, M. et al. Body mass index and the risk of meningioma, glioma and schwannoma in a large prospective cohort study (The HUNT Study). *Br. J. Cancer* **109**, 289–294 (2013).
8. Dai, Z.-F., Huang, Q.-L. & Liu, H.-P. Different body mass index grade on the risk of developing glioma: a meta-analysis. *Chin. Neurosurg. J.* **1**, 7 (2015).
9. Edlinger, M. et al. Blood pressure and other metabolic syndrome factors and risk of brain tumour in the large population-based Me-Can cohort study. *J. Hypertens.* **30**, 290–296 (2012).
10. Kitahara, C. M. et al. Personal history of diabetes, genetic susceptibility to diabetes, and risk of brain glioma: a pooled analysis of observational studies. *Cancer Epidemiol. Biomark. Prev.* **23**, 47–54 (2014).
11. Schwartzbaum, J. et al. Associations between prediagnostic blood glucose levels, diabetes, and glioma. *Sci. Rep.* **7**, 1436 (2017).
12. Seliger, C. et al. Diabetes, use of anti-diabetic drugs, and the risk of glioma. *NeuroOncology* **18**, 340–349 (2016).
13. Zhao, L., Zheng, Z. & Huang, P. Diabetes mellitus and the risk of glioma: a meta-analysis. *Oncotarget* **7**, 4483–4489 (2016).
14. Brown, C. D. et al. Body mass index and the prevalence of hypertension and dyslipidemia. *Obes. Res.* **8**, 605–619 (2000).
15. GBD 2015 Obesity Collaborators et al. Health effects of overweight and obesity in 195 countries over 25 years. *N. Engl. J. Med.* **377**, 13–27 (2017).
16. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 (2014).
17. Nitsch, D. et al. Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. *Am. J. Epidemiol.* **163**, 397–403 (2006).
18. Freeman, G., Cowling, B. J. & Schooling, C. M. Power and sample size calculations for Mendelian randomization studies using one genetic instrument. *Int. J. Epidemiol.* **42**, 1157–1163 (2013).
19. Pierce, B. L. & Burgess, S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* **178**, 1177–1184 (2013).
20. Melin, B. S. et al. Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nat. Genet.* **49**, 789–794 (2017).
21. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
22. Shungin, D. et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
23. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
24. Scott, R. A. et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
25. Gaulton, K. J. et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type-2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
26. Louis, D. N. et al. The 2016 World Health Organization Classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* **131**, 803–820 (2016).
27. Zhu Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
28. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
29. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).

30. Burgess, S. Sample size and power calculations in Mendelian randomization with a single instrumental variable and a binary outcome. *Int. J. Epidemiol.* **43**, 922–929 (2014).
31. Derr, R. L. et al. Association between hyperglycemia and survival in patients with newly diagnosed glioblastoma. *J. Clin. Oncol.* **27**, 1082–1086 (2009).
32. Yang, T. O. et al. Body size in early life and risk of lymphoid malignancies and histological subtypes in adulthood. *Int J. Cancer* **139**, 42–49 (2016).
33. Font-Burgada, J., Sun, B. & Karin, M. Obesity and cancer: the oil that feeds the flame. *Cell Metab.* **23**, 48–62 (2016).
34. Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
35. Lawlor, D. A. Commentary: two-sample Mendelian randomization: opportunities and challenges. *Int. J. Epidemiol.* **45**, 908–915 (2016).
36. Burgess, S., Davies, N. M. & Thompson, S. G. Bias due to participant overlap in two-sample Mendelian randomization. *Genet. Epidemiol.* **40**, 597–608 (2016).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018

Acta Neuropathologica (2018) 135:743–755
https://doi.org/10.1007/s00401-018-1825-z

ORIGINAL PAPER

Diffuse gliomas classified by 1p/19q co-deletion, *TERT* promoter and IDH mutation status are associated with specific genetic risk loci

Karim Labreche[1,2] · Ben Kinnersley[2] · Giulia Berzero[1,3,4] · Anna Luisa Di Stefano[1,3] · Amithys Rahimian[1] · Ines Detrait[1] · Yannick Marie[1] · Benjamin Grenier-Boley[5] · Khe Hoang-Xuan[1,3] · Jean-Yves Delattre[1,3] · Ahmed Idbaih[1,3] · Richard S. Houlston[2] · Marc Sanson[1,3]

Received: 10 November 2017 / Revised: 13 February 2018 / Accepted: 14 February 2018 / Published online: 19 February 2018
© The Author(s) 2018. This article is an open access publication

Abstract

Recent genome-wide association studies of glioma have led to the discovery of single nucleotide polymorphisms (SNPs) at 25 loci influencing risk. Gliomas are heterogeneous, hence to investigate the relationship between risk SNPs and glioma subtype we analysed 1659 tumours profiled for IDH mutation, *TERT* promoter mutation and 1p/19q co-deletion. These data allowed definition of five molecular subgroups of glioma: triple-positive (IDH mutated, 1p/19q co-deletion, *TERT* promoter mutated); *TERT*-IDH (IDH mutated, *TERT* promoter mutated, 1p/19q-wild-type); IDH-only (IDH mutated, 1p/19q wild-type, *TERT* promoter wild-type); triple-negative (IDH wild-type, 1p/19q wild-type, *TERT* promoter wild-type) and *TERT*-only (*TERT* promoter mutated, IDH wild-type, 1p/19q wild-type). Most glioma risk loci showed subtype specificity: (1) the 8q24.21 SNP for triple-positive glioma; (2) 5p15.33, 9p21.3, 17p13.1 and 20q13.33 SNPs for *TERT*-only glioma; (3) 1q44, 2q33.3, 3p14.1, 11q21, 11q23.3, 14q12, and 15q24.2 SNPs for IDH mutated glioma. To link risk SNPs to target candidate genes we analysed Hi-C and gene expression data, highlighting the potential role of *IDH1* at 2q33.3, *MYC* at 8q24.21 and *STMN3* at 20q13.33. Our observations provide further insight into the nature of susceptibility to glioma.

Introduction

Diffuse gliomas are the most common malignant primary brain tumour affecting adults with around 26,000 newly diagnosed cases each year in Europe [9]. Diffuse gliomas have traditionally been classified into oligodendroglial and astrocytic tumours and are graded II–IV, with the most common form—Glioblastoma (GBM) or glioma grade IV—typically having a median survival of only 15 months [2].

Despite glioma being an especially devastating malignancy little is known about its aetiology and aside from exposure to ionising radiation that accounts for very few cases no environmental or lifestyle factor has been unambiguously linked to risk [2]. Recent genome-wide association studies (GWAS) have, however, enlightened our understanding of glioma genetics identifying single-nucleotide polymorphisms (SNPs) at multiple independent loci influencing risk [22, 25, 35, 44, 49, 51, 63]. While understanding the functional basis of these risk loci offers the prospect of gaining insight into the development of glioma, few have been deciphered. Notable exceptions are the 17p13.1 locus, where the risk SNP rs78378222 disrupts *TP53* polyadenylation [51] and the 5p15.33 locus, where the risk SNP rs10069690

Karim Labreche and Ben Kinnersley are equally contributed.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00401-018-1825-z) contains supplementary material, which is available to authorized users.

✉ Richard S. Houlston
richard.houlston@icr.ac.uk

1 Sorbonne Universités UPMC Univ Paris 06, INSERM CNRS, U1127, UMR 7225, ICM, 75013 Paris, France

2 Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK

3 Service de neurologie 2-Mazarin, AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Paris, France

4 University of Pavia and C. Mondino National Institute of Neurology, Pavia, Italy

5 Univ. Lille, Inserm, Institut Pasteur de Lille, U1167-RID-AGE-Risk Factors and Molecular Determinants of Aging-Related Diseases, 59000 Lille, France

creates a splice-donor site leading to an alternate *TERT* splice isoform lacking telomerase activity [24].

Since the aetiological basis of glioma subtypes is likely to reflect different developmental pathways it is not perhaps surprising that subtype-specific associations have been shown for GBM (5p15.33, 7p11.2, 9p21.3, 11q14.1, 16p13.33, 16q12.1, 20q13.33 and 22q13.1) and for non-GBM glioma (1q44, 2q33.3, 3p14.1, 8q24.21, 10q25.2, 11q21, 11q23.2, 11q23.3, 12q21.2, 14q12 and 15q24.2) [35]. Recent large-scale sequencing projects have identified IDH mutation, *TERT* promoter mutation and 1p/19q co-deletion as cancer drivers in glioma. These findings have improved the subtyping of glioma [5, 12, 26, 27] and this information has been incorporated into the revised 2016 WHO classification of glial tumours [32]. Since these mutations are early events in glioma development, any relationship between risk SNP and molecular profile should provide insight into glial oncogenesis. Evidence for the existence of such subtype specificity is already provided by the association of the 8q24.21 (rs55705857) risk variant with 1p/19q co-deletion, IDH mutated glioma [13]. Additionally, it has been proposed that associations may exist between risk SNPs at 5p15.33, 9p21.3 and 20q13.33 and IDH wild-type glioma [10], as well as 17p13.1 and *TERT* promoter, IDH mutated glioma without 1p/19q co-deletion [12].

To gain a more comprehensive understanding of the relationship between the 25 glioma risk loci and tumour subtype we analysed three patient series totalling 2648 cases. Since generically the functional basis of GWAS cancer risk loci appear primarily to be through regulatory effects [53], we analysed Hi-C and gene expression data to gain insight into the likely target gene/s of glioma risk SNPs.

Materials and methods

Data sources

We analysed data from three non-overlapping case series: TCGA, French GWAS, French sequencing. Details of these datasets are provided below and are summarised in Table 1.

TCGA

Raw genotyping files (.CEL) for the Affymetrix Genome-wide version 6 array were downloaded for germline (i.e. normal blood) glioma samples from The Cancer Genome Atlas (TCGA, dbGaP study accession: phs000178.v1.p1). Controls were from publicly accessible genotype data generated by the Wellcome Trust Case–Control Consortium 2 (WTCCC2) analysis of 2699 individuals from the 1958 British birth cohort (1958-BC) [41]. Genotypes were generated using the Affymetrix Power Tools Release 1.20.5 using the

Table 1 Overview of TCGA, French GWAS and French seq series and mutation status of tumours

Dataset	Controls	Cases (GBM/non-GBM)	Case groupings														
			IDH status		*EGFR*		*CDKN2A*		Molecular subgroup				WHO 2016 classification				
			mut	wt	amp	wt	del	wt	IDH-only	*TERT*-IDH	*TERT*-only	Triple −ve	Triple +ve	Total	AstroIDH-mut	Astro IDH-wt	Oligo 1p19q	GBM IDH-mut	GBM IDH-wt	Total
TCGA	2648	521 (183/338)	293	228	246	270	254	262	100	4	45	10	65	224	166	51	116	10	171	514
French GWAS	1190	1423 (430/993)	366	498	118	628	173	573	169	46	309	141	85	750	188	214	95	27	233	757
French seq	5527	704 (181/523)	427	277	101	592	144	549	181	28	185	92	199	685	178	114	218	31	148	689
Total	9365	2648 (795/1854)	1086	1003	465	1490	571	1384	450	78	539	243	349	1659	532	379	429	68	552	1960

Amp amplified, *astro* astrocytoma, *del* deleted, *mut* mutated, *oligo* oligodendroglioma, *wt* wildtype

Birdseed (v2) calling algorithm (https://www.affymetrix.com/support/developer/powertools/changelog/index.html) and PennCNV [59]. After quality control (Supplementary Figs. 1, 2, Supplementary Table 1) there were 521 TCGA glioma cases and 2648 controls (Table 1). Glioma tumour molecular data (IDH mutation, 1p/19q co-deletion, *TERT* promoter mutation) were obtained from Ceccarelli et al. [6]. Further data (*EGFR* amplification/activating mutations, *CDKN2A* deletion) were obtained from the cBioportal for cancer genomics [15]. After adjustment for principal components there was minimal evidence of over-dispersion inflation ($\lambda = 1.01$; Supplementary Fig. 2).

French GWAS

The French-GWAS [25, 44] comprised 1423 patients with newly diagnosed grade II–IV diffuse glioma attending the Service de Neurologie Mazarin, Groupe Hospitalier Pitié-Salpêtrière Paris. The controls ($n = 1190$) were ascertained from the SU.VI.MAX (SUpplementation en VItamines et Mineraux AntioXydants) study of 12,735 healthy subjects (women aged 35–60 years; men aged 45–60 years) [19]. Tumours from patients were snap-frozen in liquid nitrogen and DNA was extracted using the QIAmp DNA minikit, according to the manufacturer's instructions (Qiagen, Venlo, LN, USA). DNA was analysed for large-scale copy number variation by comparative genomic hybridisation (CGH) array as previously described [16, 21]. For tumours not analysed by CGH array, 1p/19q co-deletion status was assigned using PCR microsatellites, and *EGFR*-amplification and *CDKN2A-p16-INK4a* homozygous deletion by quantitative PCR. *IDH1*, *IDH2* and *TERT* promoter mutation status was assigned by sequencing [26, 45].

French sequencing

Eight hundred and fifteen patients newly diagnosed grade II–IV diffuse glioma were ascertained through the Service de Neurologie Mazarin, Groupe Hospitalier Pitié-Salpêtrière Paris. Genotypes for the 25 risk SNPs were obtained by universal-tailed amplicon sequencing in conjunction with Miseq technology (Illumina Inc.). Genotypes were called using GATK (Genome Analysis ToolKit, version 3.6-0-g89b7209) software. Duplicated samples and individuals with low call rate (< 90%) were excluded ($n = 111$). Molecular profiling of tumour samples was carried out as per the French GWAS.

Unrelated French controls were obtained from the 3C Study (Group 2003) [17] a population-based, prospective study of the relationship between vascular factors and dementia being carried out in Bordeaux, Montpellier, and Dijon. Genotyping of controls was performed using Illumina Human 610-Quad BeadChips. To recover untyped genotypes imputation using IMPUTE2 software was performed using 1000 genomes multi-ethnic data (1000 G phase 1 integrated variant set release v3) as reference. SNPs genotypes were retained call rates were > 98%, Hardy–Weinberg equilibrium (HWE) P value $> 1 \times 10^{-6}$, minor allele frequency (MAF) > 1%. After quality control, 704 cases and 5527 controls were available for analysis (Table 1).

Statistical analysis

Test of association between SNP and glioma molecular subgroup was performed using SNPTESTv2.5 [33] under an additive frequentist model. Where appropriate, principal components, generated using common SNPs, were included in the analysis to limit the effects of cryptic population stratification that otherwise might cause inflation of test statistics. Eigenvectors for the TCGA study were inferred using smartpca (part of EIGENSOFTv2.4) [40] by merging cases and controls with phase II HapMap samples [25].

To ensure reliability when restricting cases to per-group low sample counts, imputed genotypes were thresholded at a probability > 0.9 (e.g. –method threshold in SNPtest) for the TCGA and French-GWAS studies. For the French-sequence study we used –method expected, as we were comparing genotypes from directly sequenced cases against imputed controls. We compared control frequencies to those from European 1000 genomes project to ensure the validity of this approach.

Meta-analyses were performed using the fixed-effects inverse-variance method based on the β estimates and standard errors from each study using META v1.6 [30]. Cochran's Q statistic was used to test for heterogeneity [20].

Risk allele number and age at diagnosis

For imputed SNPs a genotype probability threshold > 0.9 was used. The age and survival distribution of cases carrying additive combinations of risk alleles were assessed for the 25 SNPs across the molecular subgroups. Trend lines were estimated using linear regression in *R* and plotted using the *ggplot2* package [62]. Association between risk allele number and age was assessed using Pearson correlation.

Survival analysis

Survival plots were generated using the *survfit* package in *R* which computes an estimate of a survival curve for censored data using the Kaplan–Meier method. Log-rank tests were used to compare curves between groups and power to demonstrate a relationship between different groups and overall survival was estimated using sample size formulae for comparative binomial trials. The Cox proportional-hazards regression model was used to investigate the association

between survival and age, grade, molecular group and number of risk alleles. Individuals were excluded if they died within a month of surgery. Date of surgery was used as a proxy for the date of diagnosis.

Expression quantitative trait locus analysis

We searched for expression quantitative trait loci (eQTLs) in 10 brain regions using the V6p GTEx [31] portal (https://gtexportal.org/home/) as well as in whole blood using the blood eQTL browser [61] (https://molgenis58.target.rug.nl/bloodeqtlbrowser/).

Hi-C analysis

We examined for significant contacts between glioma risk SNPs and nearby genes using the HUGIn browser [34], which is based on analysis by Schmitt et al. [48]. We restricted the analysis to Hi-C data generated on H1 Embryonic Stem Cell and Neuronal Progenitor cell lines, as originally described in Dixon et al. [11]. Plotted topologically associating domain (TAD) boundaries were obtained from the insulating score method [8] at 40-kb bin resolution. We searched for significant interactions between bins overlapping the glioma risk SNP and all other bins within 1 Mb at each locus (i.e. "virtual 4C").

Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was carried out using version 3.0 with gene sets from Molecular Signatures Database (MSigDB) v6.0 [36, 52], restricted to the C2 canonical pathways sets ($n = 1329$). Analysis was carried out using default settings, with the exception of removing restrictions on gene set size. RSEM normalised mRNASeq expression data for 20,501 genes in 676 glioma cases from TCGA were downloaded from the Broad Institute TCGA GDAC (http://gdac.broadinstitute.org/). These were assigned molecular groupings using sample information from Supplementary Table 1 of Ceccarelli et al. [6].

Results

Descriptive characteristics of datasets

We studied three non-overlapping glioma case–control series of Northern European ancestry totalling 2648 cases and 9365 controls (Table 1). For 1659 of the 2648 cases information on tumour, 1p/19q co-deletion, *TERT* promoter and IDH mutation status was available (Fig. 1). Using these data allowed definition of five molecular subgroups of glioma: triple-positive (IDH mutated, 1p/19q co-deletion, *TERT* promoter mutated); *TERT*-IDH (IDH mutated, *TERT* promoter mutated, 1p/19q-wild-type); IDH-only (IDH mutated, 1p/19q wild-type, *TERT* promoter wild-type); *TERT*-only (*TERT* promoter mutated, IDH wild-type, 1p/19q wild-type) and triple-negative (IDH wild-type, 1p/19q wild-type, *TERT* promoter wild-type). As only 29 cases were classified as IDH mutation, 1p/19q co-deletion and *TERT* promoter wild-type, we restricted subsequent analyses to the five groups as above. Table 1 also shows grouping of the 1960 cases adopting the WHO 2016 classification of glial tumours into five categories (Astrocytoma with IDH mutation, IDH wild-type

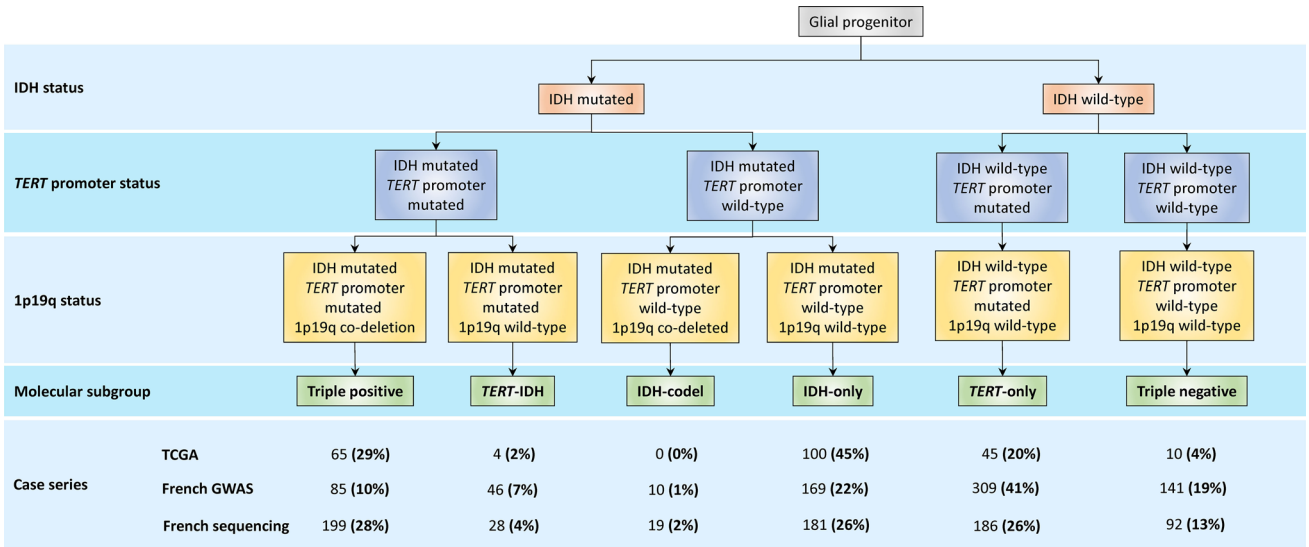


Fig. 1 Molecular classification of diffuse glioma and frequency of each subgroup in the TCGA, French-GWAS and French sequencing case series

astrocytoma, Oligodendroglioma with 1p/19q co-deletion, GBM with IDH mutation and IDH wild-type GBM) (Supplementary Table 2 [Online Resource 1]).

SNP selection

We analysed 25 SNPs, which had been reported to show the strongest genome-wide significant association with glioma in our recent meta-analysis of 12,496 cases and 18,190 controls [35] (Table 2). In the current analysis all of the SNPs exhibited a consistent direction of effect with that previously reported, albeit some weakly [Supplementary Fig. 4 (Online Resource 1), Supplementary Table 3 (Online Resource 2)].

Relationship between risk SNP and molecular subgroup

In the first instance, we examined whether the associations at the 25 risk loci were broadly defined by IDH status. We observed significant association for IDH mutated group with 1q44 (rs12076373), 2q33.3 (rs7572263), 3p14.1 (rs11706832), 8q24.21 (rs55705857), 11q21 (rs7107785), 11q23.3 (rs12803321), 14q12 (rs10131032), 15q24.2 (rs77633900) and 17p13.1 (rs78378222) risk SNPs. In addition, we found strong associations with IDH wild-type gliomas at 5p15.33 (rs10069690), 7p11.2 (rs75061358), 9p21.3 (rs634537), and 20q13.33 (rs2297440) (Supplementary Fig. 5 [Online Resource 1], Supplementary Table 3 [Online Resource 2]). Of particular note was the finding that many of the risk loci recently discovered which were reported to be associated with non-GBM (1q44, 2q33.3, 3p14.1, 11q21, 14q12, 15q24.2) [35] showed a strong association with IDH mutant glioma.

Following on from this we performed a more detailed stratified analysis based on classifying the glioma tumours into the five molecularly defined groups. We found a strong association with IDH mutated tumours at 8q24.21 (rs55705857), in particular with triple-positive glioma [$P = 1.27 \times 10^{-37}$, OR = 9.30 (6.61–13.08)], which corresponds to the WHO 2016 oligodendroglioma classification [Supplementary Fig. 6 (Online Resource 1), Supplementary Table 3 (Online Resource 2)]. Furthermore, we confirmed the previously reported associations at 5p15.33 (rs10069690), 9p21.3 (rs634537), 17p13.1 (rs78378222) and 20q13.33 (rs2297440) with *TERT*-only glioma in each of the three series [12]. Finally, we found suggestive evidence for an association between 22q13.1 (rs2235573) with *TERT*-only glioma, as well as 11q21 (rs7107785), 11q23.2 (rs648044), and 12q21.2 (rs1275600) with triple-positive glioma [Fig. 2, Supplementary Table 3 (Online Resource 2)].

In addition to data on 1p/19q co-deletion, *TERT* promoter and IDH mutation, for 1955 of the tumours we had information on *EGFR* amplification and *CDKN2A* deletion status (Table 1). Using these data we examined for an association with *EGFR* amplification and *CDKN2A* deletion, particularly focusing on the 7p11.2 (rs75061358 and rs11979158) and 9p21.3 (rs634537) risk SNPs in view of the fact that these loci map in or near *EGFR* and *CDKN2A,* respectively (Supplementary Figs. 7, 8 [Online Resource 1], Supplementary Table 3 [Online Resource 2]). At 7p11.2, the intergenic

Table 2 Overview of glioma risk SNPs at the 25 loci

Locus	SNP	Alleles	RAF	Reported subtype
1p31.3 [35]	rs12752552 [35]	**T**/C	0.87	GBM
1q32.1 [35]	rs4252707 [35]	G/**A**	0.22	Non-GBM
1q44 [35]	rs12076373 [35]	**G**/C	0.84	Non-GBM
2q33.3 [35]	rs7572263 [35]	**A**/G	0.76	Non-GBM
3p14.1 [35]	rs11706832 [35]	A/**C**	0.46	Non-GBM
5p15.33 [49]	rs10069690 [35]	C/**T**	0.28	GBM
7p11.2 [44]	rs75061358 [35]	T/**G**	0.10	GBM
7p11.2 [44]	rs11979158 [44]	**A**/G	0.83	GBM
8q24.21 [49]	rs55705857 [13, 22]	A/**G**	0.06	Non-GBM
9p21.3 [49, 63]	rs634537 [35]	T/**G**	0.41	GBM
10q24.33 [35]	rs11598018 [35]	**C**/A	0.46	Non-GBM
10q25.2 [25]	rs11196067 [25]	A/**T**	0.58	Non-GBM
11q14.1 [35]	rs11233250 [35]	**C**/T	0.87	GBM
11q21 [35]	rs7107785 [35]	**T**/C	0.48	Non-GBM
11q23.2 [25]	rs648044 [25]	**A**/G	0.39	Non-GBM
11q23.3 [49]	rs12803321 [35]	**G**/C	0.64	Non-GBM
12q21.2 [25]	rs1275600 [35]	**T**/A	0.60	Non-GBM
14q12 [35]	rs10131032 [35]	**G**/A	0.92	Non-GBM
15q24.2 [25]	rs77633900 [35]	G/**C**	0.09	Non-GBM
16p13.3 [35]	rs2562152 [35]	A/**T**	0.85	GBM
16p13.3 [35]	rs3751667 [35]	C/**T**	0.21	Non-GBM
16q12.1 [35]	rs10852606 [35]	T/**C**	0.71	GBM
17p13.1 [51]	rs78378222 [51]	T/**G**	0.01	All
20q13.33 [49, 63]	rs2297440 [35]	T/**C**	0.80	GBM
22q13.1 [35]	rs2235573 [35]	**G**/A	0.51	GBM

The risk allele frequency (RAF) is from European samples from 1000 genomes project. At 10q25.2, rs11599775 [35] failed sequencing so the originally reported SNP rs11196067 [25] was used

The risk allele is emboldened and the minor allele underlined

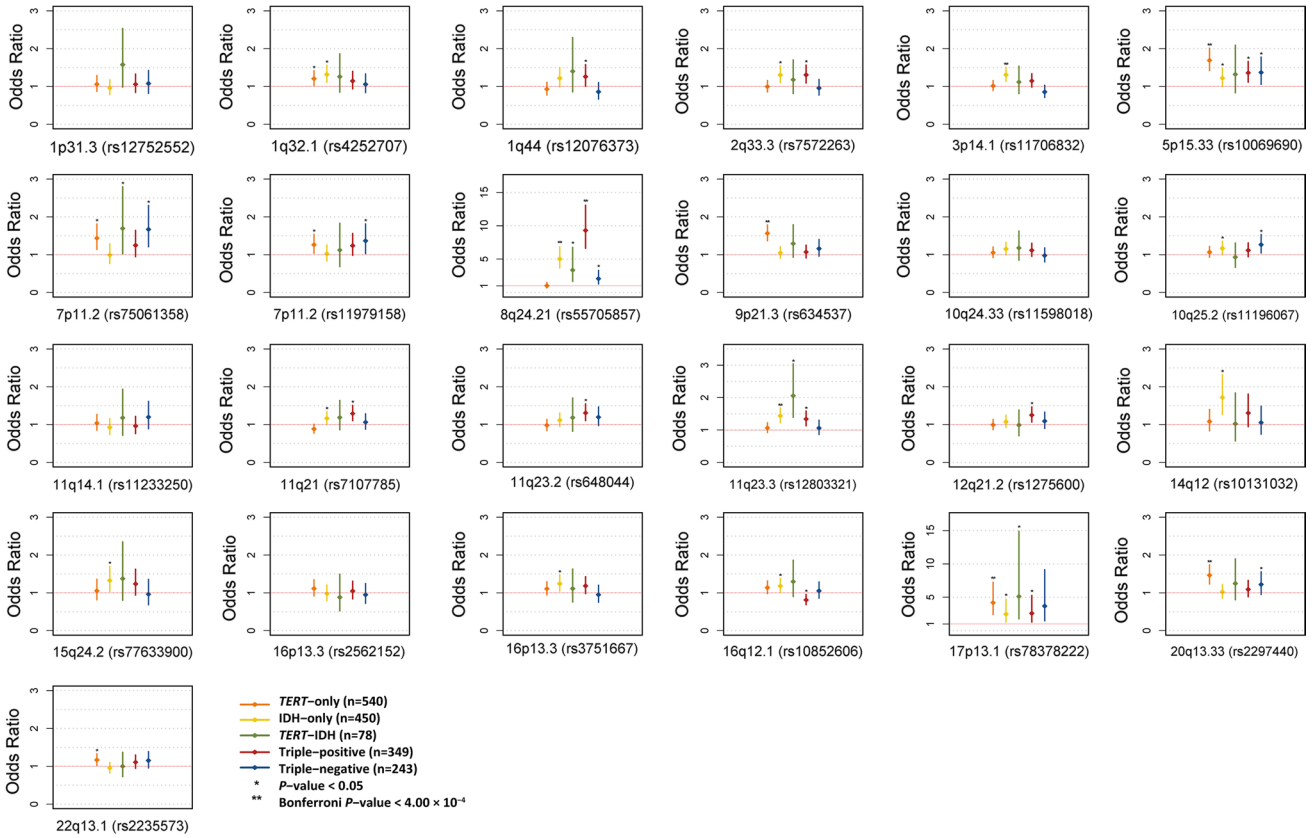


Fig. 2 Association between the 25 risk loci and glioma subgroup. Horizontal red line corresponds to an odds ratio of 1.0

variant rs75061358, which is located in the genomic vicinity of *EGFR,* was associated with *EGFR* amplified tumours and not those without amplification. There was a less strong association with EGFR amplification seen with the second independent signal at the locus defined by rs11979158, which is intronic within *EGFR* itself. At 9p21.3 rs634537, which is intronic within *CDKN2B-AS1* and in the vicinity of *CDKN2A* and *CDKN2B*, was not associated with *CDKN2A* deletion status. Low grade gliomas tend to be *EGFR* wild-type and *p16* wild-type tumours and, therefore, as anticipated many non-GBM risk SNPs were most strongly associated with these tumours; notably 2q33.3 (rs7572263), 3p14.1 (rs11706832), 8q24.21 (rs55705857), 10q25.2 (rs11196067), 11q23.3 (rs12803321) (Supplementary Figs. 7, 8 [Online Resource 1], Supplementary Table 3 [Online Resource 2]).

Polygenic contribution to age at diagnosis and patient survival

Patient survival by molecular subgroup in each of the three series was consistent with previous published reports [5, 12]; specifically, patients with triple-positive tumours had the best prognosis whilst those with *TERT*-only tumours had the worst outcome (Supplementary Fig. 3 [Online Resource 1]). We investigated whether an increased burden of glioma risk alleles might be associated with earlier age at diagnosis (i.e. indicative of influence on glioma initiation) or survival (indicative of influence on glioma progression). There was a slight albeit, non-significant trend towards decreased age at diagnosis with increased risk allele number in the IDH-only, *TERT*-only and triple-positive molecular subgroup, but with decreased risk allele number in the *TERT*-IDH and Triple-negative tumours (Supplementary Fig. 9 [Online Resource 1]). We found no overall relationship between age and risk allele number, or for the individual molecular groups (Supplementary Table 4 [Online Resource 1]). Examining each SNP individually, only rs55705857 at 8q24.21 was nominally associated with age (Supplementary Table 4 [Online Resource 1]).

We used Cox Proportional-Hazards Regression to investigate whether burden of glioma risk was associated with survival, with each risk allele coded as 0, 1 or 2. As expected, age, grade and all molecular group (Triple-negative, Triple-positive, *TERT*-only, IDH-only and *TERT*-IDH) were strongly associated with decreased survival. Intriguingly, the number of risk alleles was associated with increased survival (Supplementary Table 5 [Online Resource 1]; $P < 10^{-4}$) with 1q32.1

(rs4252707), 11q23.3 (rs12803321) and 11q21 (rs7107785) each being nominally associated with survival, independent of age and molecular subgroup. Considering the relationship between burden of glioma risk alleles and survival in each molecular subgroup a consistent association with increased survival was shown in Triple-positive, Triple-negative and *TERT*-only molecular groups but not in IDH-only and *TERT*-IDH groups.

Biological inference of risk loci

Since genomic spatial proximity and chromatin looping interactions are fundamental for the regulation of gene expression [42], we interrogated physical interactions at respective risk loci in embryonic stem cells and neuronal progenitor cells using Hi-C data. We also sought to gain insight into the possible biological mechanisms for associations by performing expression quantitative trait locus (eQTL) analysis using mRNA expression data in 10 brain regions using the GTEx portal.

We identified significant Hi-C contacts from the genomic regions which encompass 14 of the 25 risk loci implicating a number of presumptive candidate genes. For two of these, candidacy was supported by eQTL data. (Table 3; Supplementary Fig. 10 [Online Resource 1]; Supplementary Table 6 [Online Resource 3]). Notably at 2q33.3, there was a significant looping interaction between the risk SNP and *IDH1/IDH1-AS1*, as well as with *EGFR/EGFR-AS1* at 7p11.2, *CDKN2A/CDKN2B* at 9p21.3, *NFASC* at 1q32.1 and *LRIG1* at 3p14.1. At the 8q24.21 gene desert Hi-C data revealed a significant interaction between the risk SNP rs55705857 and *MYC*, as well as lincRNAs in the region such as *PCAT1/PCAT2*. Additionally, the risk SNP rs12803321 at 11q23.3 was significantly associated with *PHLDB1* expression in the brain.

Pathway analysis

To potentially gain further insight into the biological basis of subtype associations, we performed a gene-set enrichment analysis (GSEA) analysing gene expression data from TCGA (Supplementary Table 7 [Online Resource 4]). While we did not identify any significantly altered gene sets (at FDR q value < 0.1), the most significantly expressed genes in subgroups was upregulation of PI3K signalling shown in 1p/19q co-deleted tumours (Supplementary Table 7 [Online Resource 4]).

Discussion

Our findings provide further support for subtype-specific associations for glioma risk loci. Specifically, we confirm the strong relationship between the 8q24.21 (rs55705857) risk variant and Triple-positive glioma. Moreover, we substantiate the proposed specific associations between 5p15.33 (rs10069690) and 20q13.33 (rs2297440) variants with *TERT* promoter mutations, 9p21.3 (rs634537) with *TERT*-only glioma, as well as 17p13.1 (rs78378222) with *TERT*-IDH glioma. Other loci such as 1q32.1 (rs4252707) and 10q25.2 (rs11196067) appear to have more generic effects.

Although preliminary, and in part speculative, our analysis delineates potential candidate disease mechanisms across the 25 glioma risk loci (Table 3; Fig. 3). First, maintenance of telomeres is central to cell immortalization [57], and is generally considered to require mutually exclusive mutations in either the *TERT* promoter or *ATRX*. The risk alleles at 5p15.33 (*TERT*) and 10q24.33 (*OBFC1*) are associated with increased leukocyte telomere length, thereby supporting a relationship between SNP genotype and biology [56, 57, 66]. While dysregulation of the telomere gene *RTEL1* has traditionally been assumed to represent the functional basis of the 20q13.33 locus, the glioma risk SNP does not map to the locus associated with telomere length [7, 35]. Intriguingly, our analysis instead implicates *STMN3* at 20q13.33, whose over-expression promotes growth in GBM cells [68], suggesting an alternative mechanism by which the risk SNP influences glioma development. With respect to the 5p15.33 (*TERT*) and 10q24.33 (*OBFC1*) loci, it is unclear whether the effect on glioma risk is solely due to telomeres or is pleiotropic and involves multiple factors. For example, rs10069690 at 5p15.33 is strongly associated with *TERT*-only glioma, yet the *TERT* promoter mutation increases telomerase activity without necessarily affecting telomere length [6]. An intriguing hypothesis to test would, therefore, be to examine the impact of allele-specific effects of rs10069690 on telomere length in the context of gliomas carrying the *TERT* promoter mutation.

Second, the *EGFR-AKT* pathway involves *EGFR* at 7p11.2, *LRIG1* at 3p14.1, *PHLDB1* at 11q23.3 and *AKT3* at 1q44. We showed a significant interaction between the risk SNP rs11979158 at 7p11.2 and *EGFR*, consistent with a cis-regulatory effect on gene expression. Although the mechanistic basis of the 7p11.2 locus has long been suspected to involve *EGFR* and is highly associated with classical GBM, emerging evidence suggests that additional components of the *EGFR-AKT* signalling pathway are implicated by non-GBM SNPs. At the IDH-only associated locus 3p14.1, *LRIG1* is highly expressed in the brain and negatively regulates the epidermal growth factor receptor (EGFR) signalling pathway [18]. Reduced *LRIG1* expression is linked to tumour aggressiveness, temozolomide resistance and radioresistance [60, 65]. Downstream components of *EGFR-AKT* signalling are implicated at 11q23.3 via *PHLDB1*, as well as 1p31.3 via *JAK1* and 1q44 via *AKT3*. The risk allele of rs12803321 is associated with increased expression of *PHLDB1*, an insulin-responsive protein that enhances Akt

Table 3 Candidate gene basis of glioma risk loci

Locus	SNP	Molecular group	IDH, *EGFR*, *CDKN2A* status	eQTL (tissue)/Hi-C	Commentary
1p31.3	rs12752552	–	–	*JAK1* (brain)/*RAVER2, JAK1, UBE2U, CACHD1*	JAK1 is involved in actomyosin contractility in tumour cells and stroma to aid metastasis [46]
1q32.1	rs4252707	*TERT*-only*, IDH-only*	IDHmut*, *EGFR*wt*, *CDKN2A*wt*	*NFASC*	NFASC is a cell adhesion molecule involved in axon subcellular targeting and synapse formation during neural development [1]
1q44	rs12076373	TP*	IDHmut**	*AKT3, ZBTB18, SDCCAG8*	AKT3 is highly expressed in brain, regulates cell signalling in response to insulin and growth factors [4], involved in regulation of normal brain size [28]
2q33.3	rs7572263	IDH-only*, TP*	IDHmut**, *EGFR*wt*, *CDKN2A*wt*	*IDH1, IDH1-AS1*	IDH mutant protein overexpression increases glioma cell radiation sensitivity [29]
3p14.1	rs11706832	IDH-only**	IDHmut**, EGFRwt*, *CDKN2A*wt*	*LRIG1* (blood), *SLC25A26* (blood)/*LRIG1*	–
5p15.33	rs10069690	*TERT*-only**, IDH-only*, TP*, TN*	IDHmut*, IDHwt**, *EGFR*amp**, *EGFR*wt*, *CDKN2A*del*, *CDKN2A*wt**	–	rs10069690 affects *TERT* splicing [24]
7p11.2	rs75061358	*TERT*-only*, *TERT*-IDH*, TN*	IDHwt**, *EGFR*amp**, *CDKN2A*wt*	–	–
7p11.2	rs11979158	*TERT*-only*, TN*	IDHwt*, *EGFR*amp*, *EGFR*wt*, *CDKN2A*del*, *CDKN2A*wt*	*EGFR, EGFR-AS1*	–
8q24.21	rs55705857	IDH-only**, *TERT*-IDH*, TP**, TN*	IDHmut**, *EGFR*wt*, *CDKN2A*wt**, *CDKN2A*del**	*PCAT1, PCAT2, CASC8, CASC11, MYC, PVT1*	–
9p21.3	rs634537	*TERT*-only**	IDHwt**, *EGFR*amp*, *EGFR*wt*, *CDKN2A*del*, *CDKN2A*wt**	*CDKN2A, CDKN2B-AS1*	–
10q24.33	rs11598018	–	IDHmut*, *EGFR*wt*	*GSTO1, GSTO2 SH3PXD2A*	Correlated SNP to rs11598018 associated with telomere length likely through *OBFC1* [7]
10q25.2	rs11196067	IDH-only*, TN*	IDHmut*, IDHwt*, *EGFR*wt*, *CDKN2A*wt*	*TCF7L2, VTI1A, HABP2*	TCF7L2 modifies beta-catenin signalling and controls oligodendrocyte differentiation [69]
11q14.1	rs11233250	–	–	–	–
11q21	rs7107785	IDH-only*, TP*	IDHmut**, *EGFR*wt*, *CDKN2A*del*	*RP11-712B9.2* (brain)	–
11q23.2	rs648044	TP*	IDHmut*, *EGFR*wt**, *CDKN2A*wt**	*NNMT, ZBTB16*	NNMT is upregulated in GBM, NAD metabolism important in glioma [23]
11q23.3	rs12803321	IDH-only**, *TERT*-IDH*, TP*	IDHmut**, *EGFR*wt**, *CDKN2A*wt**, *CDKN2A*del*	*PHLDB1* (brain)	PHLDB1 is an insulin-responsive protein that enhances Akt activation [70]
12q21.2	rs1275600	TP*	IDHmut*, *EGFR*wt**, *CDKN2A*wt**, *CDKN2A*del*	*KRR1, GLIPR1*	GLIPR1 is targeted by TP53 [43]
14q12	rs10131032	IDH-only*	IDHmut**, *EGFR*wt*, *CDKN2A*del*, *CDKN2A*wt*	*NPAS3*	NPAS3 is a tumour suppressor for astrocytoma [37]

Table 3 (continued)

Locus	SNP	Molecular group	IDH, *EGFR*, *CDKN2A* status	eQTL (tissue)/Hi-C	Commentary
15q24.2	rs77633900	IDH-only*	IDHmut**, *EGFR*wt*, *CDKN2A*wt*	*SCAPER*	–
16p13.3	rs2562152	–	–	–	–
16p13.3	rs3751667	IDH-only*	IDHmut*, *EGFR*Ramp*, *EGFR*wt*, *CDKN2A*wt*	*RP11-161M6.2* (brain), *SOX8* (blood)	SOX8 is strongly expressed in brain and may be involved in neural development [47]
16q12.1	rs10852606	IDH-only*, TP* (−ve)	–	*HEATR3* (brain)	HEATR3 may be involved in NOD2-mediated NF-kappa B signalling [67]
17p13.1	rs78378222	*TERT*-only**, IDH-only*, *TERT*-IDH*, TP*	IDHmut**, IDHwt*, *EGFR*amp*, *EGFR*wt*, *CKDN2A*wt**, *CDKN2A*del*	–	rs78378222 affects *TP53* 3′UTR poly-adenylation processing [51]
20q13.33	rs2297440	*TERT*-only**, TN*	IDHwt**, *EGFR*Ramp**, *EGFR*wt*, *CDKN2A*del*, *CDKN2A*wt*	*STMN3* (brain), *LIME1* (blood), *ZGPAT* (blood), *EEF1A2* (blood)	Overexpression of STMN3 promotes growth in GBM cells [68]
22q13.1	rs2235573	*TERT*-only*	IDHwt*	*CTA-228A9.3* (brain)	–

TN triple negative (i.e. IDH-wildtype, *TERT* promoter wildtype, 1p/19q wildtype), *TP* triple positive (i.e. IDH-mutation, *TERT* promoter mutation and 1p/19q co-deletion)

*$P < 0.05$, **significant after adjustment for multiple comparisons

activation [70]. *AKT3* at 1q44 is highly expressed in the brain and appears to respond to EGF in a PI3K dependent manner [38], with GBM cells containing amplified AKT3 having enhanced DNA repair and resistance to radiation and temozolomide [54]. The risk allele of rs12752552 at 1p31.3 is associated with increased *JAK1* expression in brain tissue. Since *JAK1* can be activated by EGF phosphorylation, it may be involved in astrocyte formation [3, 39, 50]. The 3p14.1 and 11q23.3 loci are strongly associated with *EGFR* amplification negative gliomas, with a consistent albeit non-significant trend at 1p31.3 and 1q44, consistent with elevated upstream *EGFR* activation masking their functional effects.

Third, the *NAD* pathway involves *IDH1* at 2q33.3 and *NNMT* at 11q23.2. At 2q33.3 we detected a significant Hi-C interaction between the glioma risk SNP rs7572263 and *IDH1*/*IDH1-AS1*. Overexpression of *IDH1* mutant proteins has been reported to sensitize glioma cells to radiation [29], providing an interesting mechanism to test the allele-specific effects of this SNP. IDH mutation causes de-regulation of NAD signalling [64]. Interestingly, therefore, at 11q23.2 which is strongly associated with IDH mutated gliomas, the most convincing molecular mechanism is via *NNMT*, which encodes nicotinamide *N*-methyltransferase and is highly expressed in GBM relative to normal brain, causing methionine depletion-mediated DNA hypomethylation and accelerated tumour growth [23, 55].

Fourth, genes with established roles in neural development may be involved. While the risk SNP rs4252707 at 1q32.1 is within the intron of *MDM4*, the strongest evidence for a mechanistic effect was with *NFASC*. Neurofascin is involved in synapse formation during neural development [1] and, therefore, represents an attractive functional candidate for the association with glioma. Additionally at 16p13.3 and 20q13.33, implicated genes *SOX8* and *STMN3* are strongly expressed in the brain and thought to play a role in neural development [47, 68]. At 10q25.2, implicated gene *TCF7L2* modifies beta-catenin signalling and controls oligodendrocyte differentiation [69]. Intriguingly, 10q25.2 has previously been reported to be a risk locus for colorectal cancer [58], a tumour driven by wnt signalling, however, the risk SNP is not correlated with rs11196067 raising the possibility of tissue-specific regulation across the wider region.

Finally, the p53 pathway is involved at 17p13.1, where the risk SNP rs78378222 affects *TP53* 3′UTR poly-adenylation processing. In addition, the p53 target GLIPR1 [43] is implicated at 12q21.2. Moreover, 12q21.2 is most strongly associated with Triple-positive glioma, which does not feature *TP53* mutation, consistent with wild-type p53 protein being required for the SNP to exert a functional effect.

As with many cancers, the exact point at which the risk SNPs exert their functional impact on glioma oncogenesis still remains to be elucidated, and we did not demonstrate a relationship between increased risk allele number and age

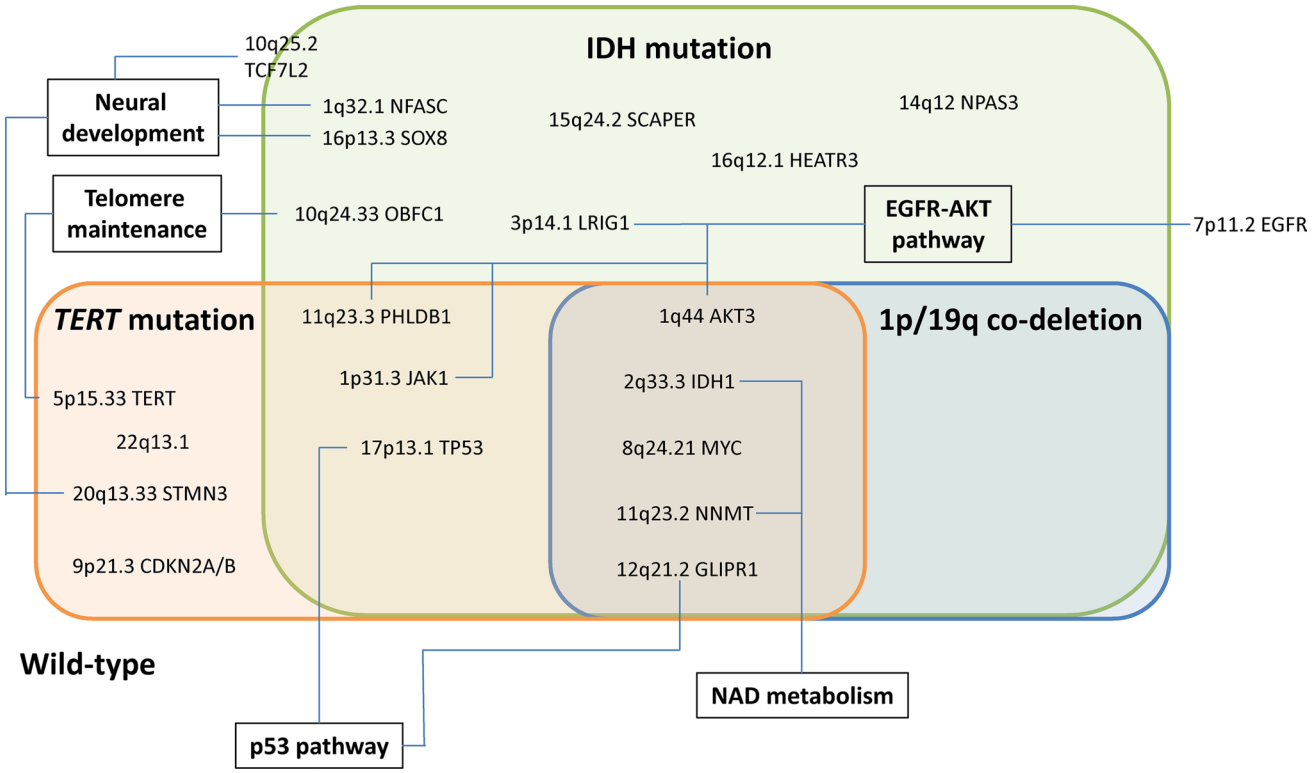


Fig. 3 Summary of the relationship between glioma risk with molecular subgroup and associated biological pathways. The extent of the evidence supporting each candidate gene (ranging from an established role in glioma to largely speculative) is summarised in Table 3

at diagnosis. Surprisingly we found a significant association between increasing risk allele number and improved outcome. This result was consistent across the prognostic molecular groups, consistent with our observations not being due to an over-representation of the more favourable prognostic groups among patients with a higher burden of risk alleles. In addition, the distribution of risk allele numbers did not differ across the four groups ($P = 0.3$, ANOVA test). Examining the impact of an individual SNP's impact on survival did not reveal any loci strongly associated with outcome. Collectively our findings suggest that, independent of other prognostic factors, the greater the number of risk alleles carried, the better the outcome.

In conclusion, we performed the most comprehensive association study between molecular subgroup and the 25 recently identified glioma risk loci to date. While confirming previous observations, we show that the majority of risk loci are associated with IDH mutation. Through the integration of Hi-C and eQTL data, we have additionally sought to define candidate target genes underlying the associations. Collectively our observations highlight pathways critical to glioma susceptibility, notably neural development and NAD metabolism, as well as EGFR-AKT signalling. Intriguingly, we show here that the number of risk alleles is consistently associated with better outcome. Functional investigation in tumour and neural progenitor-based systems will be required to more fully elucidate these molecular mechanisms. Notably, IDH mutant tumours have been shown to reshape 3D chromatin organisation and may reveal new regulatory interactions [14].

Our current analysis is based on defining glioma subgroups using only three primary markers. Given the extent of the missing heritability for glioma further expansion of GWAS by international consortia [35] is likely to result in the identification of additional risk variants. Additional molecular sub-grouping glioma resulting from ongoing large-scale tumour sequencing projects is likely to provide for further insights into glial oncogenesis and ultimately may suggest targets for novel therapeutic strategies.

Acknowledgements In France, funding was provided by the Ligue Nationale contre le Cancer, the fondation ARC, the Institut National du Cancer (INCa; PL046), the French Ministry of Higher Education and Research and the program "Investissements d'avenir" ANR-10-IAIHU-06. This study was additionally supported by funding to Mark Lathrop, including a Grant from Génome Québec, le Ministère de l'Enseignement supérieur, de la Recherche, de la Science et de la Technologie (MESRST) Québec and McGill University. We are grateful to Philippe Amouyel for providing access to control genotypes from the 3C study. KL is supported by l'Association pour la Recherche sur les Tumeurs Cérébrales (ARTC) and Institute CARNOT—Institut du Cerveau et de la Moelle Epinière (ICM). In the UK, funding was

provided by Cancer Research UK (C1298/A8362 supported by the Bobby Moore Fund), the Wellcome Trust and the DJ Fielding Medical Research Trust. The results here are in part based on data generated by the TCGA Research Network: http://cancergenome.nih.gov/. In the UK10K data generation and access was organised by the UK10K consortium and funded by the Wellcome Trust. Finally, we are grateful to all the patients and individuals for their participation and we would also like to thank the clinicians and other hospital staff, cancer registries and study staff in respective centers who contributed to the blood sample and data collection.

Author contributions KL and BK performed bioinformatics and statistical analysis. MS and RSH designed the study. GB, ALDS, AR, ID performed sequencing, YM performed the genotyping of the 25 SNPs on the second French series, MS, KHX, GB, ALDS, JYD, AI collected the clinical data. BG-N provided 3C control genotype data. All authors contributed to the final manuscript.

Compliance with ethical standards

French Tumour and blood samples were stored in the Onconeurotek tumorbank (certified NF S96 900), and received the authorization for genetic analysis from ethical committee (CPP Ile de France VI, ref A39II and 2013-1962), and French Ministry for research (AC 2013-1962).

Conflict of interest The authors declare no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Ango F, di Cristo G, Higashiyama H, Bennett V, Wu P, Huang ZJ (2004) Ankyrin-based subcellular gradient of neurofascin, an immunoglobulin family protein, directs GABAergic innervation at purkinje axon initial segment. Cell 119:257–272. https://doi.org/10.1016/j.cell.2004.10.004
2. Bondy ML, Scheurer ME, Malmer B, Barnholtz-Sloan JS, Davis FG, Il'yasova D, Kruchko C, McCarthy BJ, Rajaraman P, Schwartzbaum JA et al (2008) Brain tumor epidemiology: consensus from the Brain Tumor Epidemiology Consortium. Cancer 113:1953–1968. https://doi.org/10.1002/cncr.23741
3. Bonni A, Sun Y, Nadal-Vicens M, Bhatt A, Frank DA, Rozovsky I, Stahl N, Yancopoulos GD, Greenberg ME (1997) Regulation of gliogenesis in the central nervous system by the JAK-STAT signaling pathway. Science 278:477–483
4. Brodbeck D, Cron P, Hemmings BA (1999) A human protein kinase Bgamma with regulatory phosphorylation sites in the activation loop and in the C-terminal hydrophobic domain. J Biol Chem 274:9133–9136
5. Brat DJ, Verhaak RG, Aldape KD, Yung WK, Salama SR, Cooper LA, Rheinbay E, Miller CR, Vitucci M, Cancer Genome Atlas Research Network et al (2015) Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med 372:2481–2498. https://doi.org/10.1056/nejmoa1402121
6. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, Morozova O, Newton Y, Radenbaugh A, Pagnotta SM et al (2016) Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell 164:550–563. https://doi.org/10.1016/j.cell.2015.12.028
7. Codd V, Nelson CP, Albrecht E, Mangino M, Deelen J, Buxton JL, Hottenga JJ, Fischer K, Esko T, Surakka I et al (2013) Identification of seven loci affecting mean telomere length and their association with disease. Nat Genet 45:422–427. https://doi.org/10.1038/ng.2528
8. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ (2015) Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature 523:240–244. https://doi.org/10.1038/nature14450
9. Crocetti E, Trama A, Stiller C, Caldarella A, Soffietti R, Jaal J, Weber DC, Ricardi U, Slowinski J, Brandes A et al (2012) Epidemiology of glial and non-glial brain tumours in Europe. Eur J Cancer 48:1532–1542. https://doi.org/10.1016/j.ejca.2011.12.013
10. Di Stefano AL, Enciso-Mora V, Marie Y, Desestret V, Labussiere M, Boisselier B, Mokhtari K, Idbaih A, Hoang-Xuan K, Delattre JY et al (2013) Association between glioma susceptibility loci and tumour pathology defines specific molecular etiologies. Neuro Oncol 15:542–547. https://doi.org/10.1093/neuonc/nos284
11. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W et al (2015) Chromatin architecture reorganization during stem cell differentiation. Nature 518:331–336. https://doi.org/10.1038/nature14222
12. Eckel-Passow JE, Lachance DH, Molinaro AM, Walsh KM, Decker PA, Sicotte H, Pekmezci M, Rice T, Kosel ML, Smirnov IV et al (2015) Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. N Engl J Med 372:2499–2508. https://doi.org/10.1056/NEJMoa1407279
13. Enciso-Mora V, Hosking FJ, Kinnersley B, Wang Y, Shete S, Zelenika D, Broderick P, Idbaih A, Delattre JY, Hoang-Xuan K et al (2013) Deciphering the 8q24.21 association for glioma. Hum Mol Genet 22:2293–2302. https://doi.org/10.1093/hmg/ddt063
14. Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, Suva ML, Bernstein BE (2016) Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature 529:110–114. https://doi.org/10.1038/nature16490
15. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E et al (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 6:pl1. http://stke.sciencemag.org/content/6/269/pl1.full
16. Gonzalez-Aguilar A, Idbaih A, Boisselier B, Habbita N, Rossetto M, Laurenge A, Bruno A, Jouvet A, Polivka M, Adam C et al (2012) Recurrent mutations of MYD88 and TBL1XR1 in primary central nervous system lymphomas. Clin Cancer Res Off J Am Assoc Cancer Res 18:5203–5211. https://doi.org/10.1158/1078-0432.CCR-12-0845
17. Group CS (2003) Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. Neuroepidemiology 22:316–325
18. Gur G, Rubin C, Katz M, Amit I, Citri A, Nilsson J, Amariglio N, Henriksson R, Rechavi G, Hedman H et al (2004) LRIG1 restricts growth factor signaling by enhancing receptor ubiquitylation and degradation. EMBO J 23:3270–3281. https://doi.org/10.1038/sj.emboj.7600342
19. Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, Malvy D, Roussel AM, Favier A, Briancon S (2004) The SU. VI. MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. Arch Intern Med 164:2335–2342. https://doi.org/10.1001/archinte.164.21.2335
20. Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. BMJ 327:557–560. https://doi.org/10.1136/bmj.327.7414.557

21. Idbaih A, Marie Y, Lucchesi C, Pierron G, Manie E, Raynal V, Mosseri V, Hoang-Xuan K, Kujas M, Brito I et al (2008) BAC array CGH distinguishes mutually exclusive alterations that define clinicogenetic subtypes of gliomas. Int J Cancer J Int Cancer 122:1778–1786. https://doi.org/10.1002/ijc.23270
22. Jenkins RB, Xiao Y, Sicotte H, Decker PA, Kollmeyer TM, Hansen HM, Kosel ML, Zheng S, Walsh KM, Rice T et al (2012) A low-frequency variant at 8q24.21 is strongly associated with risk of oligodendroglial tumors and astrocytomas with IDH1 or IDH2 mutation. Nat Genet 44:1122–1125. https://doi.org/10.1038/ng.2388
23. Jung J, Kim LJY, Wang X, Wu Q, Sanvoranart T, Hubert CG, Prager BC, Wallace LC, Jin X, Mack SC et al (2017) Nicotinamide metabolism regulates glioblastoma stem cell maintenance. JCI Insight 2. https://doi.org/10.1172/jci.insight.90019
24. Killedar A, Stutz MD, Sobinoff AP, Tomlinson CG, Bryan TM, Beesley J, Chenevix-Trench G, Reddel RR, Pickett HA (2015) A common cancer risk-associated allele in the hTERT locus encodes a dominant negative inhibitor of telomerase. PLoS Genet 11:e1005286. https://doi.org/10.1371/journal.pgen.1005286
25. Kinnersley B, Labussiere M, Holroyd A, Di Stefano AL, Broderick P, Vijayakrishnan J, Mokhtari K, Delattre JY, Gousias K, Schramm J et al (2015) Genome-wide association study identifies multiple susceptibility loci for glioma. Nat Commun 6:8559. https://doi.org/10.1038/ncomms9559
26. Labussiere M, Di Stefano AL, Gleize V, Boisselier B, Giry M, Mangesius S, Bruno A, Paterra R, Marie Y, Rahimian A et al (2014) TERT promoter mutations in gliomas, genetic associations and clinico-pathological correlations. Br J Cancer 111:2024–2032. https://doi.org/10.1038/bjc.2014.538
27. Labussiere M, Idbaih A, Wang XW, Marie Y, Boisselier B, Falet C, Paris S, Laffaire J, Carpentier C, Criniere E et al (2010) All the 1p19q codeleted gliomas are mutated on IDH1 or IDH2. Neurology 74:1886–1890. https://doi.org/10.1212/WNL.0b013e3181e1cf3a
28. Lee JH, Huynh M, Silhavy JL, Kim S, Dixon-Salazar T, Heiberg A, Scott E, Bafna V, Hill KJ, Collazo A et al (2012) De novo somatic mutations in components of the PI3 K-AKT3-mTOR pathway cause hemimegalencephaly. Nat Genet 44:941–945. https://doi.org/10.1038/ng.2329
29. Li S, Chou AP, Chen W, Chen R, Deng Y, Phillips HS, Selfridge J, Zurayk M, Lou JJ, Everson RG et al (2013) Overexpression of isocitrate dehydrogenase mutant proteins renders glioma cells more sensitive to radiation. Neuro Oncol 15:57–68. https://doi.org/10.1093/neuonc/nos261
30. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, Berrettini W, Knouff CW, Yuan X, Waeber G et al (2010) Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet 42:436–440. https://doi.org/10.1038/ng.572
31. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N et al (2013) The genotype-tissue expression (GTEx) project. Nat Genet 45:580–585. https://doi.org/10.1038/ng.2653
32. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, Ohgaki H, Wiestler OD, Kleihues P, Ellison DW (2016) The 2016 World Health Organization classification of tumors of the central nervous system: a summary. Acta Neuropathol 131:803–820. https://doi.org/10.1007/s00401-016-1545-1
33. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913. https://doi.org/10.1038/ng2088
34. Martin JS, Xu Z, Reiner AP, Mohlke KL, Sullivan P, Ren B, Hu M, Li Y (2017) HUGIn: Hi-C unifying genomic interrogator. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx359
35. Melin BS, Barnholtz-Sloan JS, Wrensch MR, Johansen C, Il'yasova D, Kinnersley B, Ostrom QT, Labreche K, Chen Y, Armstrong G et al (2017) Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. Nat Genet 49:789–794. https://doi.org/10.1038/ng.3823
36. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E et al (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34:267–273. https://doi.org/10.1038/ng1180
37. Moreira F, Kiehl TR, So K, Ajeawung NF, Honculada C, Gould P, Pieper RO, Kamnasaran D (2011) NPAS3 demonstrates features of a tumor suppressive role in driving the progression of Astrocytomas. Am J Pathol 179:462–476. https://doi.org/10.1016/j.ajpath.2011.03.044
38. Okano J, Gaslightwala I, Birnbaum MJ, Rustgi AK, Nakagawa H (2000) Akt/protein kinase B isoforms are differentially regulated by epidermal growth factor stimulation. J Biol Chem 275:30934–30942. https://doi.org/10.1074/jbc.M004112200
39. Park OK, Schaefer TS, Nathans D (1996) In vitro activation of Stat3 by epidermal growth factor receptor kinase. Proc Natl Acad Sci USA 93:13704–13708
40. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190. https://doi.org/10.1371/journal.pgen.0020190
41. Power C, Elliott J (2006) Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol 35:34–41. https://doi.org/10.1093/ije/dyi183
42. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES et al (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159:1665–1680. https://doi.org/10.1016/j.cell.2014.11.021
43. Ren C, Ren CH, Li L, Goltsov AA, Thompson TC (2006) Identification and characterization of RTVP1/GLIPR1-like genes, a novel p53 target gene cluster. Genomics 88:163–172. https://doi.org/10.1016/j.ygeno.2006.03.021
44. Sanson M, Hosking FJ, Shete S, Zelenika D, Dobbins SE, Ma Y, Enciso-Mora V, Idbaih A, Delattre JY, Hoang-Xuan K et al (2011) Chromosome 7p11.2 (EGFR) variation influences glioma risk. Hum Mol Genet 20:2897–2904. https://doi.org/10.1093/hmg/ddr192
45. Sanson M, Marie Y, Paris S, Idbaih A, Laffaire J, Ducray F, El Hallani S, Boisselier B, Mokhtari K, Hoang-Xuan K et al (2009) Isocitrate dehydrogenase 1 codon 132 mutation is an important prognostic biomarker in gliomas. J Clin Oncol Off J Am Soc Clin Oncol 27:4150–4154. https://doi.org/10.1200/JCO.2009.21.9832
46. Sanz-Moreno V, Gaggioli C, Yeo M, Albrengues J, Wallberg F, Viros A, Hooper S, Mitter R, Feral CC, Cook M et al (2011) ROCK and JAK1 signaling cooperate to control actomyosin contractility in tumor cells and stroma. Cancer Cell 20:229–245. https://doi.org/10.1016/j.ccr.2011.06.018
47. Schepers GE, Bullejos M, Hosking BM, Koopman P (2000) Cloning and characterisation of the Sry-related transcription factor gene Sox8. Nucleic Acids Res 28:1473–1480
48. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL et al (2016) A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep 17:2042–2059. https://doi.org/10.1016/j.celrep.2016.10.061
49. Shete S, Hosking FJ, Robertson LB, Dobbins SE, Sanson M, Malmer B, Simon M, Marie Y, Boisselier B, Delattre JY et al (2009) Genome-wide association study identifies five

susceptibility loci for glioma. Nat Genet 41:899–904. https://doi.org/10.1038/ng.407
50. Shuai K, Ziemiecki A, Wilks AF, Harpur AG, Sadowski HB, Gilman MZ, Darnell JE (1993) Polypeptide signalling to the nucleus through tyrosine phosphorylation of Jak and Stat proteins. Nature 366:580–583. https://doi.org/10.1038/366580a0
51. Stacey SN, Sulem P, Jonasdottir A, Masson G, Gudmundsson J, Gudbjartsson DF, Magnusson OT, Gudjonsson SA, Sigurgeirsson B, Thorisdottir K et al (2011) A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. Nat Genet 43:1098–1103. https://doi.org/10.1038/ng.926
52. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102:15545–15550. https://doi.org/10.1073/pnas.0506580102
53. Sud A, Kinnersley B, Houlston RS (2017) Genome-wide association studies of cancer: current insights and future perspectives. Nat Rev Cancer 17:692–704. https://doi.org/10.1038/nrc.2017.82
54. Turner KM, Sun Y, Ji P, Granberg KJ, Bernard B, Hu L, Cogdell DE, Zhou X, Yli-Harja O, Nykter M et al (2015) Genomically amplified Akt3 activates DNA repair pathway and promotes glioma progression. Proc Natl Acad Sci USA 112:3421–3426. https://doi.org/10.1073/pnas.1414573112
55. Ulanovskaya OA, Zuhl AM, Cravatt BF (2013) NNMT promotes epigenetic remodeling in cancer by creating a metabolic methylation sink. Nat Chem Biol 9:300–306. https://doi.org/10.1038/nchembio.1204
56. Walsh KM, Codd V, Rice T, Nelson CP, Smirnov IV, McCoy LS, Hansen HM, Elhauge E, Ojha J, Francis SS et al (2015) Longer genotypically-estimated leukocyte telomere length is associated with increased adult glioma risk. Oncotarget 6:42468–42477. https://doi.org/10.18632/oncotarget.6468
57. Walsh KM, Wiencke JK, Lachance DH, Wiemels JL, Molinaro AM, Eckel-Passow JE, Jenkins RB, Wrensch MR (2015) Telomere maintenance and the etiology of adult glioma. Neuro Oncol 17:1445–1452. https://doi.org/10.1093/neuonc/nov082
58. Wang H, Burnett T, Kono S, Haiman CA, Iwasaki M, Wilkens LR, Loo LW, Van Den Berg D, Kolonel LN, Henderson BE et al (2014) Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. Nat Commun 5:4613. https://doi.org/10.1038/ncomms5613
59. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17:1665–1674. https://doi.org/10.1101/gr.6861907
60. Wei J, Qi X, Zhan Q, Zhou D, Yan Q, Wang Y, Mo L, Wan Y, Xie D, Xie J et al (2015) miR-20a mediates temozolomide-resistance in glioblastoma cells via negatively regulating LRIG1 expression. Biomed Pharmacother 71:112–118. https://doi.org/10.1016/j.biopha.2015.01.026
61. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, Christiansen MW, Fairfax BP, Schramm K, Powell JE et al (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 45:1238. https://doi.org/10.1038/Ng.2756
62. Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer, New York City
63. Wrensch M, Jenkins RB, Chang JS, Yeh RF, Xiao Y, Decker PA, Ballman KV, Berger M, Buckner JC, Chang S et al (2009) Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. Nat Genet 41:905–908. https://doi.org/10.1038/ng.408
64. Yan H, Parsons DW, Jin G, McLendon R, Rasheed BA, Yuan W, Kos I, Batinic-Haberle I, Jones S, Riggins GJ et al (2009) IDH1 and IDH2 mutations in gliomas. N Engl J Med 360:765–773. https://doi.org/10.1056/NEJMoa0808710
65. Yang JA, Liu BH, Shao LM, Guo ZT, Yang Q, Wu LQ, Ji BW, Zhu XN, Zhang SQ, Li CJ et al (2015) LRIG1 enhances the radiosensitivity of radioresistant human glioblastoma U251 cells via attenuation of the EGFR/Akt signaling pathway. Int J Clin Exp Pathol 8:3580–3590
66. Zhang C, Doherty JA, Burgess S, Hung RJ, Lindstrom S, Kraft P, Gong J, Amos CI, Sellers TA, Monteiro AN et al (2015) Genetic determinants of telomere length and risk of common cancers: a Mendelian randomization study. Hum Mol Genet 24:5356–5366. https://doi.org/10.1093/hmg/ddv252
67. Zhang W, Hui KY, Gusev A, Warner N, Ng SM, Ferguson J, Choi M, Burberry A, Abraham C, Mayer L et al (2013) Extended haplotype association study in Crohn's disease identifies a novel, Ashkenazi Jewish-specific missense mutation in the NF-kappaB pathway gene, HEATR3. Genes Immun 14:310–316. https://doi.org/10.1038/gene.2013.19
68. Zhang Y, Ni S, Huang B, Wang L, Zhang X, Li X, Wang H, Liu S, Hao A (2015) Overexpression of SCLIP promotes growth and motility in glioblastoma cells. Cancer Biol Ther 16:97–105. https://doi.org/10.4161/15384047.2014.987037
69. Zhao C, Deng Y, Liu L, Yu K, Zhang L, Wang H, He X, Wang J, Lu C, Wu LN et al (2016) Dual regulatory switch through interactions of Tcf7l2/Tcf4 with stage-specific partners propels oligodendroglial maturation. Nat Commun 7:10883. https://doi.org/10.1038/ncomms10883
70. Zhou QL, Jiang ZY, Mabardy AS, Del Campo CM, Lambright DG, Holik J, Fogarty KE, Straubhaar J, Nicoloro S, Chawla A et al (2010) A novel pleckstrin homology domain-containing protein enhances insulin-stimulated Akt phosphorylation and GLUT4 translocation in adipocytes. J Biol Chem 285:27581–27589. https://doi.org/10.1074/jbc.M110.146886

Identification of novel recurrent ETV6-IGH fusions in primary central nervous system lymphoma

Aurélie Bruno PhD[1*], Karim Labreche MsC[1*], Maïlys Daniau MsC[1,2*], Blandine Boisselier MsC[3], Guillaume Gauchotte MD PhD[4], Louis Royer-Perron MD[1], Amithys Rahimian MsC[1,5], Frédéric Lemoine PhD[6], Pierre de la Grange PhD[6], Justine Guégan MsC[7], Franck Bielle MD PhD[1,5,8,] Marc Polivka MD[9], Clovis Adam MD[10], David Meyronet MD PhD[11], Dominique Figarella-Branger MD PhD[12], Chiara Villa MD PhD[13], Fabrice Chrétien MD PhD[14], Sandrine Eimer MD PhD[15], Frédéric Davi MD PhD[16], Audrey Rousseau MD PhD[3], Caroline Houillier MD[17,18], Carole Soussain MD PhD[18,19], Karima Mokhtari MD[1,5,8], Khê Hoang-Xuan MD PhD[1,17,18], Agusti Alentorn MD PhD[1,17]

(1) Groupe Hospitalier Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, Sorbonne Universités, UPMC, University Paris 06, Institut du Cerveau et de la Moelle épinière, INSERM U1127, CNRS UMR 7225, 47 Boulevard de l'Hôpital, 75013 Paris, France

(2) Institut du Cerveau et de la Moelle épinière, Plateforme iGenSeq, 47 Boulevard de l'Hôpital, 75013 Paris, France

(3) Département de pathologie cellulaire et tissulaire, CHU d'Angers, 4, rue Larrey, 49933 Angers, France

(4) Neuropathologie, CHRU Nancy, 29 Avenue du Maréchal de Lattre de Tassigny, 54000 Nancy

(5) Onconeurotek, Groupe Hospitalier Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, 47 Boulevard de l'Hôpital, 75013 Paris, France

(6) Genosplice, Institut du Cerveau et de la Moelle épinière, 47 Boulevard de l'Hôpital, 75013 Paris, France

© The Author(s) 2018. Published by Oxford University Press on behalf of the Society for Neuro-Oncology. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

(7) Institut du Cerveau et de la Moelle épinière, ICONICS (bioinformatic and biostatistics core facility), 47 Boulevard de l'Hôpital, 75013 Paris, France

(8) Neuropathologie, Groupe Hospitalier Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, 47 Boulevard de l'Hôpital, 75013 Paris, France

(9) Hôpital Lariboisière, Assistance Publique-Hôpitaux de Paris, Service d'Anatomie et Cytologie Pathologiques, 2 Rue Ambroise Paré, 75010 Paris, France

(10) Centre Hospitalier Universitaire Bicêtre, Assistance Publique-Hôpitaux de Paris, Service d'anatomopathologie, 78 Rue du Général Leclerc, 94275 Le Kremlin-Bicêtre Cedex, France

(11) Hospices Civils de Lyon, Hôpital Neurologique, 59 Boulevard Pinel, 69677 Bron, France and INSERM U842, Université Lyon 1, 69003 Lyon, France

(12) Centre Hospitalier Universitaire La Timone, Laboratoire d'anatomie pathologique-neuropathologique and Tumorothèque de l'Assistance Publique-Hôpitaux de Marseille (AC 2013-1786), 264 Rue Saint-Pierre, 13385 Marseille Cedex 5, France ; Centre de Recherches en Oncologie biologique et Onco-pharmacologie, INSERM U911, Université Aix-Marseille, 27 Boulevard Jean Moulin, 13385 Marseille Cedex 5, France

(13) Hôpital Foch, Service d'anatomie pathologique, 40 Rue Worth, 92151 Suresnes, France

(14) Centre hospitalier Sainte Anne, Université Paris Descartes, Sorbonne Paris Cité, 1 Rue Cabanis, 75014 Paris, France

(15) Centre Hospitalier Universitaire Bordeaux, Service de Pathologie, Site Pellegrin, 146 Rue Léo Saignat Victor Segalen University, 33076 Bordeaux Cedex, France

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

(16) Hématologie, Hôpital Pitié-Salpêtrière, AP-HP, Paris, France; and UMR_S 1138, Sorbonne Universités, UPMC, University Paris 06, 47 Boulevard de l'Hôpital, 75013, Paris, France

(17) Service de Neurologie Mazarin, Groupe Hospitalier Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, 47 Boulevard de l'Hôpital, 75013 Paris, France

(18) Réseau Expert National LOC (Lymphomes Oculo-Cérébraux)

(19) Hôpital René Huguenin, Institut Curie, Service d'Hématologie, 35 rue Dailly, 92210 Saint Cloud, France

Corresponding author: Agusti Alentorn MD PhD, e-mail: agusti.alentorn@aphp.fr

Groupe Hospitalier Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, Université Pierre et Marie Curie-Paris 6, Institut du Cerveau et de la Moelle épinière, INSERM U1127, CNRS UMR 7225, 47 Boulevard de l'Hôpital, 75013 Paris, France

*These authors contributed equally to the manuscript.

Running title: Novel recurrent ETV6-IgH fusions in PCNSL

Funding

This work is part of the national program Cartes d'Identité des Tumeurs® (CIT) funded and developed by the Ligue nationale contre le cancer. Institut National du Cancer, Association pour la recherche sur les tumeurs cérébrales (ARTC), Cancéropôle Île-de-France "Emergence 2015-1" (2015-1-EMERG-05-INSERM 6-1), Ligue nationale contre le cancer (Comité du Val d'Oise, R14044DD), Ligue Nationale contre le cancer "Recherche épidemiologique" (N° PRE2015.LNCC), Fondation pour la Recherche Médicale (FDT20140930968) and program "Investissements d'avenir" ANR-10-IAIHU-06. Institut National du Cancer (INCa) (Réseau

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

Expert National LOC, Lymphomes Oculo-Cérébraux). This study was supported by Lymphomes Oculo-Cérébraux (LOC) study group network (réseau national de centres experts des lymphomes primitifs du système nerveux central.

Authors disclosures: The authors declare no conflicts of interest.

Abstract

Background: Primary central nervous system lymphoma (PCNSL) represents a particular entity within non-Hodgkin lymphomas and is associated with poor outcome. The present study addresses the potential clinical relevance of chimeric transcripts in PCSNL discovered by using RNA-sequencing (RNA-Seq).

Methods: Seventy-two immunocompetent and newly diagnosed PCNSL cases were included in the present study. Among them, six were analyzed by RNA-seq to detect new potential fusion transcripts. We confirmed the results in the remaining 66 PCNSL. The gene fusion was validated by fluorescence in situ hybridization (FISH) using formalin-fixed paraffin-embedded (FFPE) samples. We assessed the biological and clinical impact of one new gene fusion.

Results: We identified a novel recurrent gene fusion *ETV6-IgH*. Overall, *ETV6-IgH* was found in 13 out of 72 PCNSL (18%). No fusion conserved an intact functional domain of ETV6 and *ETV6* was significantly underexpressed at gene level, suggesting an ETV6 haploinsufficiency mechanism. The presence of the gene fusion was also validated by FISH in FFPE samples. Finally, PCNSL samples harboring *ETV6-IgH* showed a better prognosis in multivariate analysis, p-value=0.03, HR=0.33, 95% interval confidence (IC95) [0.12-0.88]. The overall survival at 5 years was of 69% for PCNSL harboring *ETV6-IgH* vs 29% for samples without this gene fusion.

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

Conclusions: *ETV6-IgH* is a new potential surrogate marker of PCNSL with favorable prognosis with *ETV6* haploinsuffiency as a possible mechanism. The potential clinical impact of *ETV6-IgH* should be validated in larger prospective studies.

Keywords: Primary CNS lymphoma, RNA sequencing, fusion gene, ETV6-IGH, haploinsufficiency

Importance of the study

Primary central nervous system lymphoma (PCNSL) is a rare entity with heterogenous clinical evolution. Chimeric genes are interesting molecular markers because they may allow to detect novel oncogenic pathways and could be used as a biomarkers. We analyzed 6 fresh-frozen PCNSL by RNA-Seq and we have detected a recurrent chimeric fusion involving ETV6-IGH. The prevalence of this gene fusion has been established using 66 fresh-frozen PCNSL samples by direct sequencing. We have analyzed the potential functional impact of this gene fusion by western blot of transfected COS-7 cells with ETV6-IGH gene fusion. Finally, we found that PCNSL harboring this chimeric gene are associated with a better prognosis in the multivariate analysis as well as low *ETV6* expression, suggesting a haploinsufficiency mechanism.

Introduction

Primary central nervous system lymphoma (PCNSL) is an intriguing entity currently classified according to World Health Organization (WHO) criteria as a diffuse large B-cell lymphoma (DLBCL) restricted to the CNS.[1] PCNSL are extranodal, malignant non-Hodgkin lymphomas that are confined to the brain, eyes, leptomeninges, or spinal cord, in the absence of systemic lymphoma.[1] The particular tropism of PCNSL to the central nervous system (CNS) as well as the reason why this neoplasm exclusively manifest in the immunoprivileged brain in the absence of systemic spread is still unclear.[2] Although PCNSL is associated with a

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

dismal prognosis, the prognosis has been substantially improved by using high-dose methotrexate.[3] However, treatment of this disease remains challenging because remissions are frequently of short-lasting with substantial toxicity.[4]

The rarity of this disease and the small amount of tissue obtained in the vast majority of cases from stereotaxic biopsies has delayed understanding of the oncogenesis of PCNSL. The expression profiling of PCNSL with expression of *BCL6*, *IRF4* together with an aberrant somatic hypermutation (aSHM) indicates that PCNSL cells belong to a late germinal center B cell.[2,5] We and others have reported recurrent copy number aberrations using high-density CGH or SNP arrays and described the mutational landscape of PCNSL using whole-exome sequencing (WES).[6–12] The most striking alterations reported to date are (i) frequent chromosomal deletions affecting *HLA* locus (6p21.32), 6q22 chromosome and *CDKN2A* locus (9p21.3) and (ii) somatic mutations in genes involved in B-cell receptor/Toll-like receptor/NF-κB pathways, especially *MYD88* and *CD79B*.[6,13-15]

The present study addresses the potential clinical relevance of chimeric transcripts in PCSNL discovered by using RNA-Seq. We have identified several new fusion genes and we have focused on the most frequent one involving *ETV6* and *IgH*, as a novel gene fusion that could be potentially used as a prognostic marker in PCNSL.

Material and methods

PCNSL samples

Seventy-two immunocompetent (HIV negative and no history of immunosuppressive drugs or organ transplantation) and newly diagnosed PCNSL cases homogenously treated with high-dose methotrexate regimen (3.5g/m^2) were included in the present study. Tumors were selected on the basis of fresh frozen tissue availability. All tumors were PCNSL classified as CD20+ DLBCL according to the WHO criteria [1] and demonstrated to contain at least 80%

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

tumor cells. For all cases, systemic lymphoma was excluded by extensive investigation. This project was approved by the local ethics committee (CPPRB Pitié-Salpêtrière). Written consent for sample collection and genetic analysis was obtained from all the participants. Details about PCNSL cases investigated in the present study are provided in Supplementary Supplementary Table S1.

RNA extraction and quality assessment

Total RNA from cryopreserved samples was extracted using the iPrep Trizol® Plus RNA kit (Life Technologies). Tumor lysis was first performed in Trizol (Invitrogen) lysis buffer and using FastPrep system (MP Biomedicals). After chloroform addition, total RNA was purified using iPrep Trizol® Plus RNA kit. RNA was quantified using a NanoDrop spectrophotometer, and the quality, depending on RNA Integrity Number (RIN), RNA concentration and 28S:18S rRNA ratio, was assessed using an Agilent BioAnalyzer.

RNA sequencing

RNA sequencing was performed for cases with a minimal amount of RNA of 1.5µg and a RNA Integrity Number (RIN) of at least 7. Library was prepared using the TruSeq Stranded mRNA kit protocol (Illumina technology) with an input total RNA of 1µg. Capture of polyadenylated RNA was realized using oligo dT beads. Captured RNA was fragmented in approximatively 400bp. After DNA synthesis, Illumina adaptors ligation and library amplification by PCR, 100bp paired-end sequencing was performed on an Illumina HiSEQ 2000.

Data analysis and detection of putative fusion transcripts

Data analysis was realized by GenoSplice technology (ICM, France). Data Quality control was performed using FastQC v0.10.1

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Fusion transcripts were detected using three different approaches: tophat-fusion, defuse, and EASANA-fusion (Genosplice). We only considered chimeric transcripts that were commonly detected by at least two three algorithms. Further details on the bioinformatics analysis are found in the supplementary methods.

ETV6 expression in PCNSL and transfected cells

ETV6 expression was assessed by quantitative PCR. Primer and probes were synthesized using Universal Probe Library (UPL, Roche) software (primers and probes are provided in suppl data). The qPCR was performed on LightCycler 480 (Roche) and using the following conditions : 10 minutes at 95°C for 1 cycle, 10 seconds at 95°C, 30 seconds at 60°c and 1 second at 72°C for 45 cycles, 30 seconds at 40°C. Expression levels were normalized to (PPIA) and relative expression of ETV6 was calculated using the ΔΔCt method.

Cell culture

Monkey kidney COS-7 cell line were obtained from the American Type Culture Collection and supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin (15140, ThermoFisher Scientific). The cells were cultured in a humidified incubator with 5% CO2 at 37°C.

Plasmid construction

ETV6 wt, ETV6-IgHG4, ETV6-delta (a truncated version of ETV6 without IgH) and Green Fluorescent Protein (GFP) control were cloned into lentiviral vector control using a CMV-3HA-pPGK-puromycin selection. COS-7 cells expressing ETV6 wt, ETV6 delta (ETV6 lacking the last four exons), ETV6-IgHG4 or GFP control were generated by lentiviral transduction and subsequent puromycin selection.

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

Western blot analysis

Immunoblotting of COS-7 cells was performed with the following antibodies: anti-HA (ab18181, Abcam, diluted 1:5000) and anti-ETV6 (ab151698, Abcam, diluted 1:5000) in at first, and then anti-cyclophylin B (PA1-027A, Pierce, diluted 1:2000). After the overnight incubation at 4°C with primary antibodies IRDye 680RD Goat anti-Rabbit IgG (Li-Cor, diluted 1:5000), membranes were washed again and scanned on Odyssey CLx Imaging System. Scan settings were high quality, 169µm resolution, intensity 5 for both channels without focus offset. Further details are provided in the supplementary methods.

Interphase FISH on Formalin-fixed Paraffin-embedded (FFPE) sections

ETV-IgH fusion was confirmed using 3µm FFPE tissue section using *ETV6-IgH* positive PCNSL samples detected by RNA-Seq or by Sanger sequencing that were deparaffinized with the histology FISH Accessory Kit (Dako). Slides were visualized using a fluorescence scanner (Pathscan, Excilone). Hybridizing signals in at least 100 non-overlapping nuclei were counted. The presence of the breakapart probe signal in greater than 15% of tumor cells was defined as positive for ETV6-IgH fusions.

Direct sequencing of *MYD88* and *CD79B* somatic mutations

The hotspots mutations of *MYD88* (L265P) and *CD79B* (Y196) were investigated by Sanger as previously described.[8] Shortly, the amplifications conditions were 94°C for 3 min followed by 45 cycles of 94°Cx15 sec, 60°Cx45 and 72°Cx1 min, with a final step at 72°C for 8 min. The somatic DNA was amplified using the following primers: for MYD88 L265P TGTGTGAGTGAATGTGTGCC (forward) and GAGTCCAGAACCAAGATTTGGT and for CD79B Y196 CACCCCTCTCCCTGGCCCTC (forward) and CGGGACCACACCCCAACCAC (reverse).

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

Validation

The validation of the putative fusion transcripts identified by RNA-Seq was performed using RT-PCR. Five hundred nanograms of total RNA were retrotranscribed using the Maxima first strand cDNA synthesis kit (Thermo Scientific) following the manufacturer's instructions. PCR was performed using primers designed according to predicted fusion transcript sequence with the forward primer located within the 5 prime end of *ETV6* transcript and the reverse primer within the 3 prime end of *IGH* transcript. Primer sequences are listed in the Supplementary Table S2. The amplification conditions were as already described.[8] The purified sequences were addressed to GATC Biotech for conventional Sanger sequencing.

All transcriptome sequencing data have been deposited at the Gene Expression Omnibus (GEO), which is hosted by the National Center for Biotechnology Information (NCBI), under the accession code GSE81816.

The investigation of additional cases with *ETV6-IgH* fusion gene was assessed using an optimized RT-PCR assay. Further details are provided in supplementary methods en Supplementary Table S2.

Statistical Analyses

We applied unpaired Wilcoxon Mann-Whitney test for comparing *ETV6* expression levels obtained by qRT-PCR, age and Karnofsky Perfomance Status (KPS), both as a continuous variables, in PCNSL samples according to *ETV6-IgH* status.

Kaplan-Meier analysis and the log-rank test were used to explore differences between overall survival according to *ETV6-IgH* status, age (≥60 vs <60 years) and Karnofsky Performance Status (KPS) (≥70 vs <70%). Cox proportional hazards regression models were used to obtain hazard ratios (HR) with Wald 95% confidence intervals (CI) for the relationship between OS

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

and *ETV6-IgH* status, age, and KPS in the patient cohorts. We assessed the proportionality of the hazards for Cox regression with the Schoenfeld residuals. All p-values were two-sided and p-values less than 0.05 were interpreted as statistically significant.

Analyses were performed using R statistical software, version 3.3 (Free Software Foundation available at http://www.r-project.org).

Results

Gene fusion identification using RNA-Seq

We collected a cohort of 6 PCNSL samples on which we performed transcriptome sequencing with the aim of identifying new chimer alterations. We applied 3 different gene fusion algorithms and only those fusion genes detected by all of them were further considered.

We identified a total of 1827 putative fusion transcripts in the 6 PCNSL samples (Supplementary Table S3).

Thirty-two putative fusions involving 57 distinct genes were commonly detected by at least 2 out of the 3 fusion detection algorithms (Figure 1, Supplementary Figure S1 and Figure S2, Table S3) including 3 inter-chromosomal and 29 intra-chromosomal fusions. Only 3 fusions were commonly detected by the 3 pipelines: *SSR2-GON4L*, *ETV6-IgH* and *WHSC2-LETM1*. Among them, we selected the most frequent chimeric transcript *ETV6-IgH* detected in 2 cases out of the 6 investigated by RNA-Seq. This fusion raised our interest because *ETV6* is frequently involved in different hematological diseases, it has a prominent role in hematopoietic stem cell homeostasis.[16,17] In addition, focal deletions of *ETV6* locus and recurrent somatic mutations have been recently identified in 2 different PCNSL studies.[18,19] In the same line, there are many studies suggesting that *ETV6* could act in some setting as a tumor suppressor gene.[20]

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

Validation of the *ETV6-IGH* gene fusion by sequencing in 66 PCNSL

The fusion is a somatic genomic event as *ETV6* break-apart FISH and FISH with custom *ETV6* and *IgH* probes revealed rearrangements in the respective chromosomal regions in the tumor cells, but not in surrounding nontumoral cells (Figure 2E).

We next performed reverse transcriptase PCR (RT-PCR) using primers specific for the chimeric transcript to identify additional tumors bearing the fusion in a set of 66 PCNSL, in addition to the 6 PCNSL tested by RNA-seq. We identified 11 additional tumors carrying the fusion (Supplementary Table S1). All breakpoints that we identified on *ETV6* were located on the 5 prime side of the transcript - *i.e* before the third exon - while *IGH* breakpoints were distributed all along the transcripts. Some preferential clustering of breakpoints were identified at the ends of exons 1 and 2 for *ETV6* and in the middle of exon 4 for *IgHG4* (Figure 2B-2D and Supplementary Figure S3). The predicted fusion proteins indicated that none preserved an entire functional domain of ETV6 protein (Supplementary Figure S3). Four ETV6-IgH proteins were predicted to conserve a part of the PNT (or pointed) domain responsible for protein-protein interactions including one conserving more than half of its domain.

Clinical impact of ETV6 gene fusion

Age and sex were equally distributed in PCNSL carrying *ETV6-IgH* and in *ETV6* wild-type (wt) PNCSL counterparts (Supplementary Table S4A). Interestingly, univariate survival analysis pinpointed that PCNSL harboring *ETV6-IgH* had a better prognosis than their *ETV6* wt counterparts (p = 0.04, Figure 3A). Moreover, multivariate analysis using Cox proportional hazards model confirmed that *ETV6-IgH* was independently associated with favorable prognosis after adjusting for age and KPS (p-value=0.03, HR=0.33, 95% interval confidence

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

(IC95) [0.12-0.88]) (Supplementary Table S4B) with an OS at 5 years of 69% for PCNSL harboring *ETV6-IgH* vs 29%.

We also analyzed the prognostic impact of *ETV6* expression. Patients with high *ETV6* expression levels (according to the median) had lower KPS compared to the low *ETV6* expression samples (p=0.02) and age was equally distributed (Supplementary Table S4C). Low *ETV6* expression in the overall cohort was associated with a better prognosis in univariate (p-value=0.007, Supplementary Figure S4) and in multivariate analysis (p-value=0.01, HR=0.44 [0.24-0.83], Supplementary Table 4D) with an OS at 5 years of 55% for PCNSL with low *ETV6* expression levels vs 20%. However, when only *ETV6* wild-type (wt) samples (i.e. without *ETV6* fusion) were analyzed, we did not find any prognostic impact of *ETV6* gene expression (p=0.17, Supplementary Figure S5).

Functional impact of ETV6-IGH fusion

In most of the cases *ETV6* fusions involved the first 2 exons, potentially altering the expression *ETV6*. To validate this prediction, we transduced COS-7 cells with *ETV6-IgHG4* and *ETV6*-Delta lentiviruses and we performed western blot of COS-7 cells to determine the expression. We also transduced this cell line with either an empty vector, or a virus containing normal *ETV6* (*ETV6* wt) (Figure 3B). We did not find any difference in ETV6 protein expression compared to the different ETV6 constructions (Figure 3B). We next analyzed the expression of *ETV6* in COS-7 cells using qRT-PCR showing a underexpression of *ETV6* 3p compared to control constructions ($p < 0.05$, data not shown), arguing in favor of a potential haploinsufficiency of *ETV6* expression. Likewise, qRT-PCR in PCNSL samples showed a significant *ETV6* underexpression in *ETV6-IgH* positive samples compared to those with *ETV6* wt ($p < 0.05$, Figure 3C).

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

Taken together this data suggest that *ETV6-IgH* leads to a single-allele loss of *ETV6* reducing its gene expression (Figure 3C) but without significantly modifying its protein expression (Figure 3B). Therefore, haploinsufficiency may have a potential impact in the mechanism involved in this gene fusion.

Correlation of ETV6 fusion with other molecular features

We have also screened the most frequent hotspot mutations described in PCNSL: *MYD88* L265P and CD79B Y196.[18,19] Overall, 29/72 (40.3%) harbored *MYD88* L265P, 19/72 (26.4%) CD79B Y196 mutation and 15/72 (20.8%) both of them, Supp Table S1. In addition, the distribution of *MYD88* L265P and *CD79B* Y196 mutations was similar according to *ETV6-IgH* gene fusion status. Indeed, according to *ETV6-IgH* gene fusion status, 5/13 (38.6%) harbored MYD88 L265P mutation vs 24/59 (40.7%), p-value = 1 Fisher's exact text, CD79B Y196 in 2/13 (15.4%) vs 13/59 (22%), p-value =0.7, and also in 2/13 (15.4%) vs 13/59 (22%) both of them (Supp Table S1).

Discussion

We have characterized a small cohort of PCNSL by RNA-seq to discover new chimeric transcripts. We have identified several potential interesting gene fusions and we have further estimated the frequency and the clinical impact in a larger series of PCNSL using fresh-frozen tissue. All *ETV6-IgH* gene fusions were validated by cDNA sequencing. Overall, we identified 13 cases with *ETV6-IgH* fusion gene in our whole-cohort of 72 PCNSL. We estimate the frequency of *ETV6-IgH* in PCNSL to approximately 18%. Therefore, *ETV6-IgH* is the most frequently reported fusion gene in PCNSL. We provide evidence that *ETV6-IgH* leads to a decrease of expression (at mRNA), suggesting a potential role of haploinsufficiency of *ETV6*. In the same line, there are many studies suggesting that *ETV6* could be a tumor suppressor gene also by an haploinsufficiency mechanism.[21] Haploinsufficiency occurs when

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

the amount of protein product created from the remaining wild-type allele is not sufficient for normal cellular function. Therefore *ETV6* could be considered as 'haplo-insufficient' to indicate that one copy of the gene is insufficient for proper function.[22]

The ETV6 protein contains two major domains, the HLH (helix-loop-helix) domain, encoded by exons 3 and 4, and the ETS domain, encoded by exons 6 through 8, with in between the internal domain encoded by exon 5. *ETV6* is a strong transcriptional repressor, acting through its HLH and internal domains.[16] This transcription factor is frequently rearranged in childhood pre-B acute lymphoblastic leukemia (ALL) and leukemia of myeloid or lymphoid origins.[23,24] It is important to emphasize that *ETV6* is known to be fused with a wide range of genes encoding receptor tyrosine kinases genes, transcription factors, homeobox genes, and many others.[25] Interestingly, the mentioned fusions, as the one described in this study, do not include the full-length ETV6 protein. Remarkably, several gene fusions involving *ETV6* have been associated with a haploinsufficiency mechanism.[26,27] Furthermore, even fusions of *ETV6* with the same target will not always have the same breakpoints in ETV6 protein.[25] Interestingly, in a recent PCNSL study, *ETV6* was found to be statistically significant associated as a target of aSHM phenotype in 22/41 of cases (53.7%).[19]

The fusion partner of *ETV6*, *IgH* is a frequently rearranged locus in DLBCL and PCNSL and in both diseases these rearrangements could be associated with aSHM.[28] IgH translocations have been found in 13% of PCNSL and are less frequent than in DLBCL (45%).[14] In addition, the most common *IgH* translocation partner in PCNSL is *BCL6* (80%) while in DLBCL is more frequently linked to *BCL2* (15%).[14]

Furthermore, *ETV6-IgH* samples harbored a favorable prognosis in multivariate analysis with an OS at 5 years of 69% for PCNSL harboring *ETV6-IgH* vs 29% for samples without this gene fusion after adjusting for age and KPS (Supp Table S4). Different prognostic scores

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

using clinical characteristics have been proposed but age and KPS seem to be the strongest independent predictors in PCNSL.[29] However, it should be noted that further molecular alterations might impact the clinical evolution of PCNSL. Accordingly, another gene fusion involving *ETV6*, *ETV6-RUNX1*, is the most frequent genomic aberration found in pre-B acute lymphoblastic leukemia (ALL), occurring in approximately 25% of cases, and is associated with favorable prognosis.[30] Different potential biomarkers of prognosis in PCNSL have been described during the last years. Overexpression of *BCL-6* was associated with improved survival compared to patients whose tumors did not express *BCL-6*.[31] However, other studies did not corroborate these findings.[32] More recently, recurrent somatic nonsynonymus mutations in *MYD88* and *CD79B* genes were found in approximately two thirds of PCNSL.[9,18,19] Interestingly, the blockade of B-cell-receptor (BCR) signals with an inhibitor of BTK kinase (ibrutinib) has shown clinical efficacy against activated B-cell DLBCL, notably in DLBCL with double mutations (*CD79B* and *MYD88*), showing a potential prediction biomarker for a target therapy.[33] In our study the distribution of double mutations of *MYD88* L265P and *CD79B* Y196 were equally distributed according to *ETV6-IgH* gene fusion status (2/13 (15.4%) vs 13/59 (22%), p-value = 0.7, Fisher's exact test. It is also important to highlight that all the patients included in this study were treated with high-dose methotrexate regimen without any prior chemotherapy treatment nor radiotherapy.[34]

We have validated the presence of *ETV6-IgH* gene fusion by FISH in FFPE samples. This technique could be used to detect this chimeric transcript in the clinical setting and to be screened in PCNSL samples in order to validate this potential new biomarker.

Recent studies have pinpointed recurrent chromosomal rearrangements in PCNSL with highly heterogeneous results.[18,19] Among the recently described gene fusions one study found: *BCL6-IgH* (17%) and PD ligand foci (*PD-L1* or *PD-L2*) translocations (6%).[18] We found a common gene fusion with this study involving *BCL6-IGL* (Supplementary Table S3).

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

Conversely, in another recent study, only one rare fusion gene was found in a series of 30 PCNSL.[19] These divergent results could be explained in part due to different pipeline analysis, NGS approaches and different tissue samples (i.e. fresh-frozen and FFPE). Interestingly, one of these studies using whole-exome and RNA-seq analysis of PCNSL had also identified inactivating alterations of *ETV6* in 3 out of 24 cases (12.5%), with deletions of exon 2 or exons 2-5 that modified the reading frame.[18] Therefore, it is tempting to speculate that these single-allele deletions of *ETV6* may be also involved in loss-of-function of this gene leading to a reduction of the amount of ETV6 within the cell as we showed in *ETV6-IgH* chimeric transcript. Furthermore, the mutational landscape of DLBCL using whole-genome analysis have also highlighted the presence of a rare gene fusion involving *ETV6* with a *IgH* in 1 out of 40 (2.5%) that was further validated by RNA-seq.[35] Consequently, we can hypothesize that due to the higher frequency found in this study, this gene fusion could be more frequently found in PCNSL (13 out of 72, 18% vs 1 out of 40, 2.5%, p-value = 0.017, Fisher's exact test).

It is worth mentioning that our study has some limitations. This is a small retrospective dataset and the potential clinical impact should be validated in larger prospective studies. The impact of intratumoral heterogeneity of *ETV6-IgH* has not been thoroughly assessed. Further studies analyzing larger cohort of PCNSL using FISH are warranted to better characterize the potential impact of intratumoral heterogeneity in *ETV6-IgH* gene fusion. It should be also noted that other genetic alterations (i.e. mutations and copy number alterations) of *ETV6* wild-type allele may modify the impact if this gene fusion. These alterations should be further evaluated in future studies. Finally, we cannot formally exclude a potential role of dominant-negative in *ETV6-IgH*. However, the loss of both oligomerization and DNA-binding domains in *ETV6-IgH* fusion make unlikely that this molecular mechanism has a major effect.

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

To the best of our knowledge, this is the first study showing a novel fusion gene in PCNSL that could be used as a potential biomarker to detect a subset of PCNSL patients with less severe disease.

Acknowledgments

We would like to thank the French LOC Network investigators.

Statistical analysis conducted by : Aurélie Bruno, PhD; Karim Labreche, MsC; Frédéric Lemoine, PhD; Pierre de la Grange, PhD; Justine Guégan, MsC and Dr Agusti Alentorn MD PhD

References

1. Campo E, Swerdlow SH, Harris NL, Pileri S, Stein H, Jaffe ES. The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. Blood . 2011;117(19):5019–5032.

2. Deckert M, Montesinos-Rongen M, Brunn A, Siebert R. Systems biology of primary CNS lymphoma: from genetic aberrations to modeling in mice. Acta Neuropathol . 2014;127(2):175–188.

3. Hoang-Xuan K, Bessell E, Bromberg J, et al. Diagnosis and treatment of primary CNS lymphoma in immunocompetent patients: guidelines from the European Association for Neuro-Oncology. Lancet Oncol . 2015;16(7):e322-332.

4. Ricard D, Idbaih A, Ducray F, Lahutte M, Hoang-Xuan K, Delattre J-Y. Primary brain tumours in adults. Lancet . 2012;379(9830):1984–1996.

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

5. Montesinos-Rongen M, Küppers R, Schlüter D, et al. Primary central nervous system lymphomas are derived from germinal-center B cells and show a preferential usage of the V4-34 gene segment. Am J Pathol . 1999;155(6):2077–2086.

6. Braggio E, McPhail ER, Macon W, et al. Primary central nervous system lymphomas: a validation study of array-based comparative genomic hybridization in formalin-fixed paraffin-embedded tumor specimens. Clin Cancer Res . 2011;17(13):4245–4253.

7. Braggio E, Van Wier S, Ojha J, et al. Genome-Wide Analysis Uncovers Novel Recurrent Alterations in Primary Central Nervous System Lymphomas. Clin Cancer Res . 2015;21(17):3986–3994.

8. Bruno A, Boisselier B, Labreche K, et al. Mutational analysis of primary central nervous system lymphoma. Oncotarget . 2014;5(13):5065–5075.

9. Gonzalez-Aguilar A, Idbaih A, Boisselier B, et al. Recurrent mutations of MYD88 and TBL1XR1 in primary central nervous system lymphomas. Clin Cancer Res . 2012;18(19):5203–5211.

10. Schwindt H, Vater I, Kreuz M, et al. Chromosomal imbalances and partial uniparental disomies in primary central nervous system lymphoma. Leukemia . 2009;23(10):1875–1884.

11. Sung CO, Kim SC, Karnan S, et al. Genomic profiling combined with gene expression profiling in primary central nervous system lymphoma. Blood . 2011;117(4):1291–1300.

12. Vater I, Montesinos-Rongen M, Schlesner M, et al. The mutational pattern of primary lymphoma of the central nervous system determined by whole-exome sequencing. Leukemia . 2015;29(3):677–685.

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

13. Booman M, Douwes J, Glas AM, et al. Mechanisms and effects of loss of human leukocyte antigen class II expression in immune-privileged site-associated B-cell lymphoma. Clin Cancer Res . 2006;12(9):2698–2705.

14. Cady FM, O'Neill BP, Law ME, et al. Del(6)(q22) and BCL6 rearrangements in primary CNS lymphoma are indicators of an aggressive clinical course. J Clin Oncol . 2008;26(29):4814–4819.

15. Montesinos-Rongen M, Godlewska E, Brunn A, Wiestler OD, Siebert R, Deckert M. Activating L265P mutations of the MYD88 gene are common in primary central nervous system lymphoma. Acta Neuropathol . 2011;122(6):791–792.

16. Bohlander SK. ETV6: a versatile player in leukemogenesis. Semin Cancer Biol . 2005;15(3):162–174.

17. Hock H, Meade E, Medeiros S, et al. Tel/Etv6 is an essential and selective regulator of adult hematopoietic stem cell survival. Genes Dev . 2004;18(19):2336–2341.

18. Chapuy B, Roemer MGM, Stewart C, et al. Targetable genetic features of primary testicular and primary central nervous system lymphomas. Blood . 2016;127(7):869–881.

19. Fukumura K, Kawazu M, Kojima S, et al. Genomic characterization of primary central nervous system lymphoma. Acta Neuropathol . 2016;131(6):865–875.

20. Van Vlierberghe P, Ambesi-Impiombato A, Perez-Garcia A, et al. ETV6 mutations in early immature human T cell leukemias. J Exp Med . 2011;208(13):2571–2579.

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

21. Fenrick R, Wang L, Nip J, et al. TEL, a putative tumor suppressor, modulates cell growth and cell morphology of ras-transformed cells while repressing the transcription of stromelysin-1. Mol Cell Biol . 2000;20(16):5828–5839.

22. Berger AH, Pandolfi PP. Haplo-insufficiency: a driving force in cancer. J Pathol . 2011;223(2):138–147.

23. Golub TR, Barker GF, Bohlander SK, et al. Fusion of the TEL gene on 12p13 to the AML1 gene on 21q22 in acute lymphoblastic leukemia. Proc Natl Acad Sci U S A . 1995;92(11):4917–4921.

24. Golub TR, Barker GF, Lovett M, Gilliland DG. Fusion of PDGF receptor beta to a novel ets-like gene, tel, in chronic myelomonocytic leukemia with t(5;12) chromosomal translocation. Cell . 1994;77(2):307–316.

25. De Braekeleer E, Douet-Guilbert N, Morel F, Le Bris M-J, Basinko A, De Braekeleer M. ETV6 fusion genes in hematological malignancies: a review. Leuk Res . 2012;36(8):945–961.

26. Panagopoulos I, Strömbeck B, Isaksson M, Heldrup J, Olofsson T, Johansson B. Fusion of ETV6 with an intronic sequence of the BAZ2A gene in a paediatric pre-B acute lymphoblastic leukaemia with a cryptic chromosome 12 rearrangement. Br J Haematol . 2006;133(3):270–275.

27. Belloni E, Trubia M, Mancini M, et al. A new complex rearrangement involving the ETV6, LOC115548, and MN1 genes in a case of acute myeloid leukemia. Genes Chromosomes Cancer . 2004;41(3):272–277.

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

28. Montesinos-Rongen M, Van Roost D, Schaller C, Wiestler OD, Deckert M. Primary diffuse large B-cell lymphomas of the central nervous system are targeted by aberrant somatic hypermutation. Blood . 2004;103(5):1869–1875.

29. Abrey LE, Ben-Porat L, Panageas KS, et al. Primary central nervous system lymphoma: the Memorial Sloan-Kettering Cancer Center prognostic model. J Clin Oncol . 2006;24(36):5711–5715.

30. Rubnitz JE, Downing JR, Pui CH, et al. TEL gene rearrangement in acute lymphoblastic leukemia: a new genetic marker with prognostic significance. J Clin Oncol . 1997;15(3):1150–1157.

31. Levy O, Deangelis LM, Filippa DA, Panageas KS, Abrey LE. Bcl-6 predicts improved prognosis in primary central nervous system lymphoma. Cancer. 2008;112(1):151–156.

32. Camilleri-Broët S, Crinière E, Broët P, et al. A uniform activated B-cell-like immunophenotype might explain the poor prognosis of primary central nervous system lymphomas: analysis of 83 cases. Blood . 2006;107(1):190–196.

33. Wilson WH, Young RM, Schmitz R, et al. Targeting B cell receptor signaling with ibrutinib in diffuse large B cell lymphoma. Nat Med . 2015;21(8):922–926.

34. Hoang-Xuan K, Taillandier L, Chinot O, et al. Chemotherapy alone as initial treatment for primary CNS lymphoma in patients older than 60 years: a multicenter phase II study (26952) of the European Organization for Research and Treatment of Cancer Brain Tumor Group. J Clin Oncol . 2003;21(14):2726–2731.

Accepted Manuscript

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

35. Morin RD, Mungall K, Pleasance E, et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. Blood . 2013;122(7):1256–1265.

Figure legends

Figure 1. Overview of the 32 putative chimeric transcripts identified by at least 2 fusion detection algorithms. Inner arcs represent rearrangements from the 6 cases analyzed by RNA-Seq. Interchromosomal fusions are shown in purple and intrachromosomal fusions are shown in red.

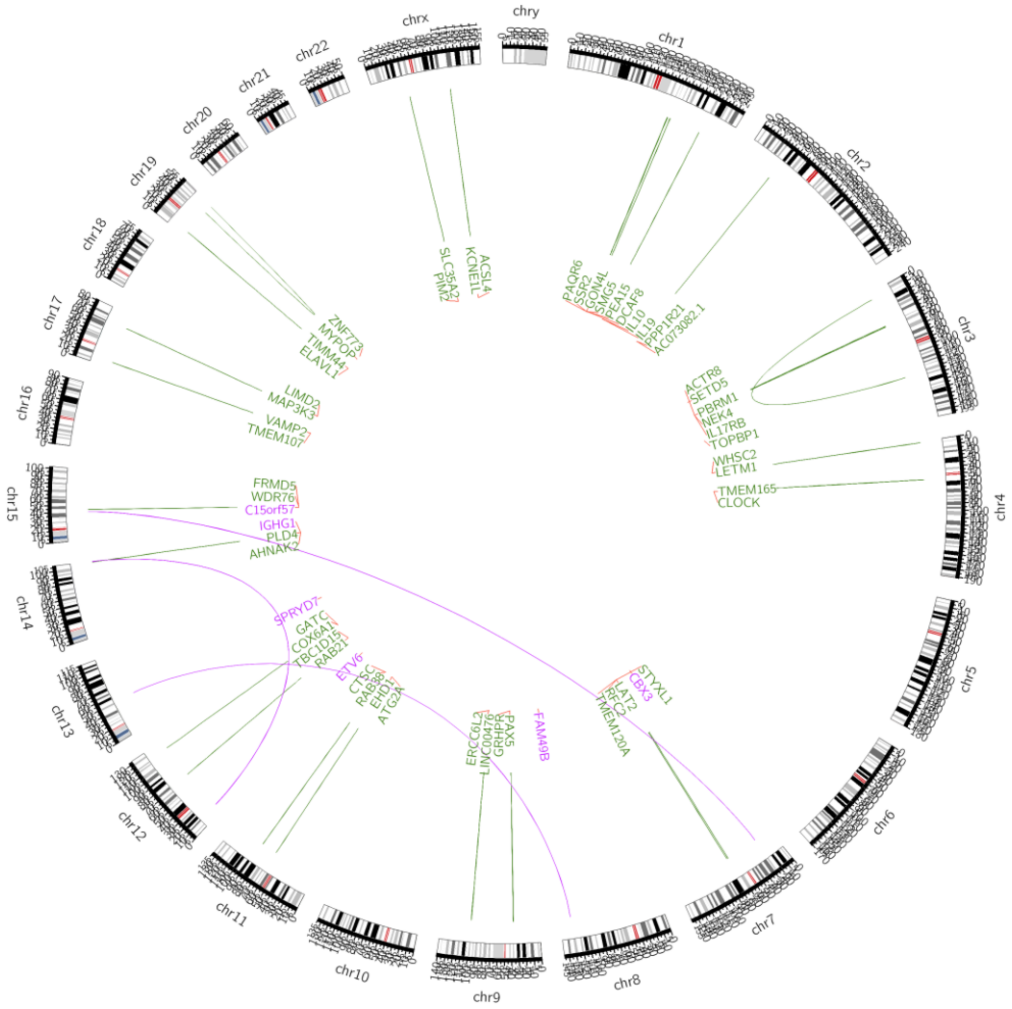
Figure 2. *ETV6-IgH* fusion transcripts identified by RNA sequencing of PCNSL and FISH. **(A)** *ETV6-IgH* specific PCR from cDNA derived from the 6 PCNSL cases of the RNA-Seq cohort showing two different ETV6 breakpoints (red arrows) detected in two patients. **(B)** and **(C)** Schematics of the two fusion transcripts identified in two cases using RNA-Seq. Regions corresponding to *ETV6* or *IgH* are shown in blue or purple, respectively. Vertical red lines show breakpoints and horizontal dotted lines indicate open reading frame for each fusion transcript. **(D)** Chromosomal rearrangements detected by FISH using custom *ETV6* and *IgH* probes showing (white arrows).

Figure 3. (A) Kaplan-Meier plot showing overall survival (OS) according to *ETV6-IgH* status. **(B)** Western Blot using COS-7 cell lines in the presence of and empty vector, no transfected cell line, transfection with *ETV6*, *ETV6-IgHG4* and *ETV6* truncated constructions using a lentivirus. **(C)** The boxes represent the median (black middle line) limited by the 25th (Q1) and 75th (Q3) percentiles of *ETV6* expression according to *ETV6-IgH* fusion status in arbitrary units. Significance of the differences of *ETV6* expression was determined using the Wilcoxon-Mann-Whitney test.

Accepted Manuscript

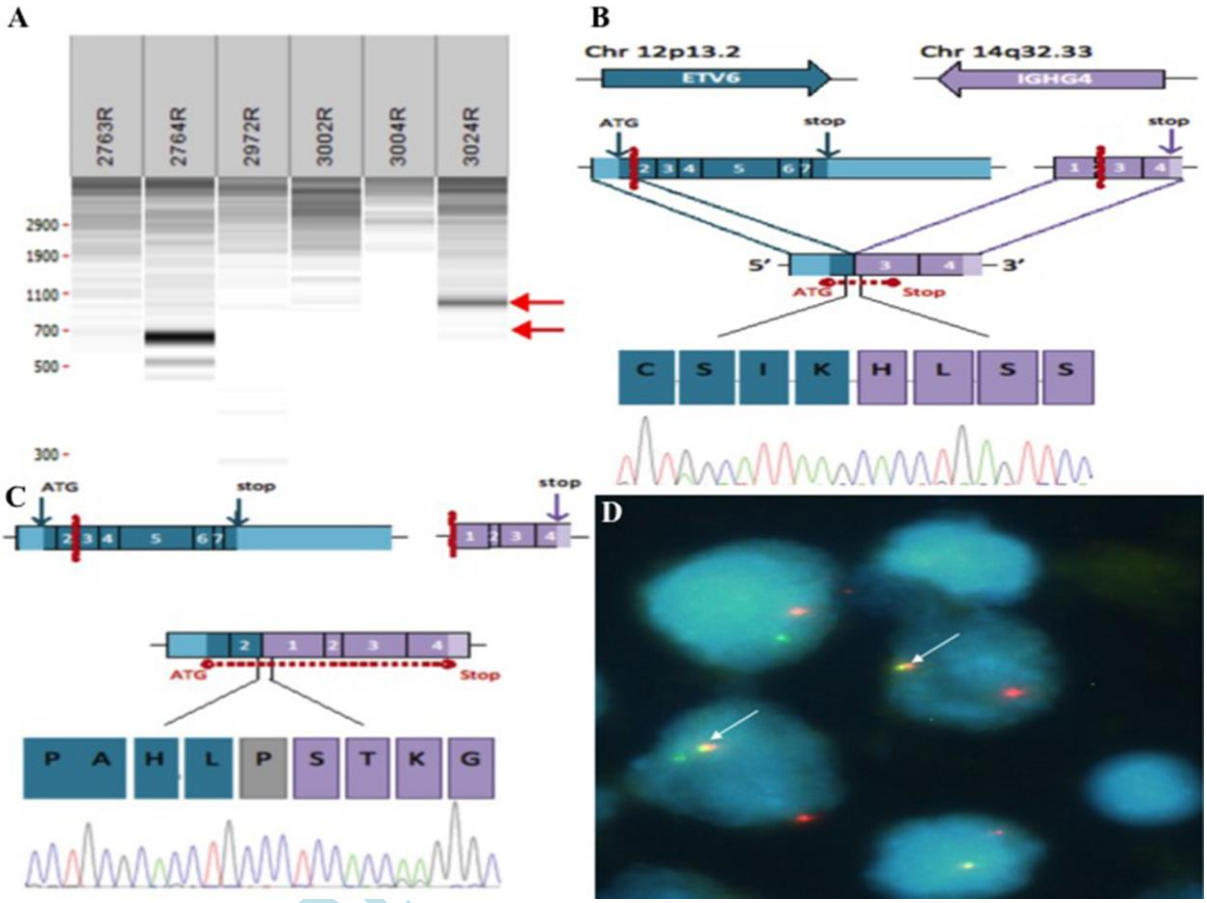
Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

Figure 1.



Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

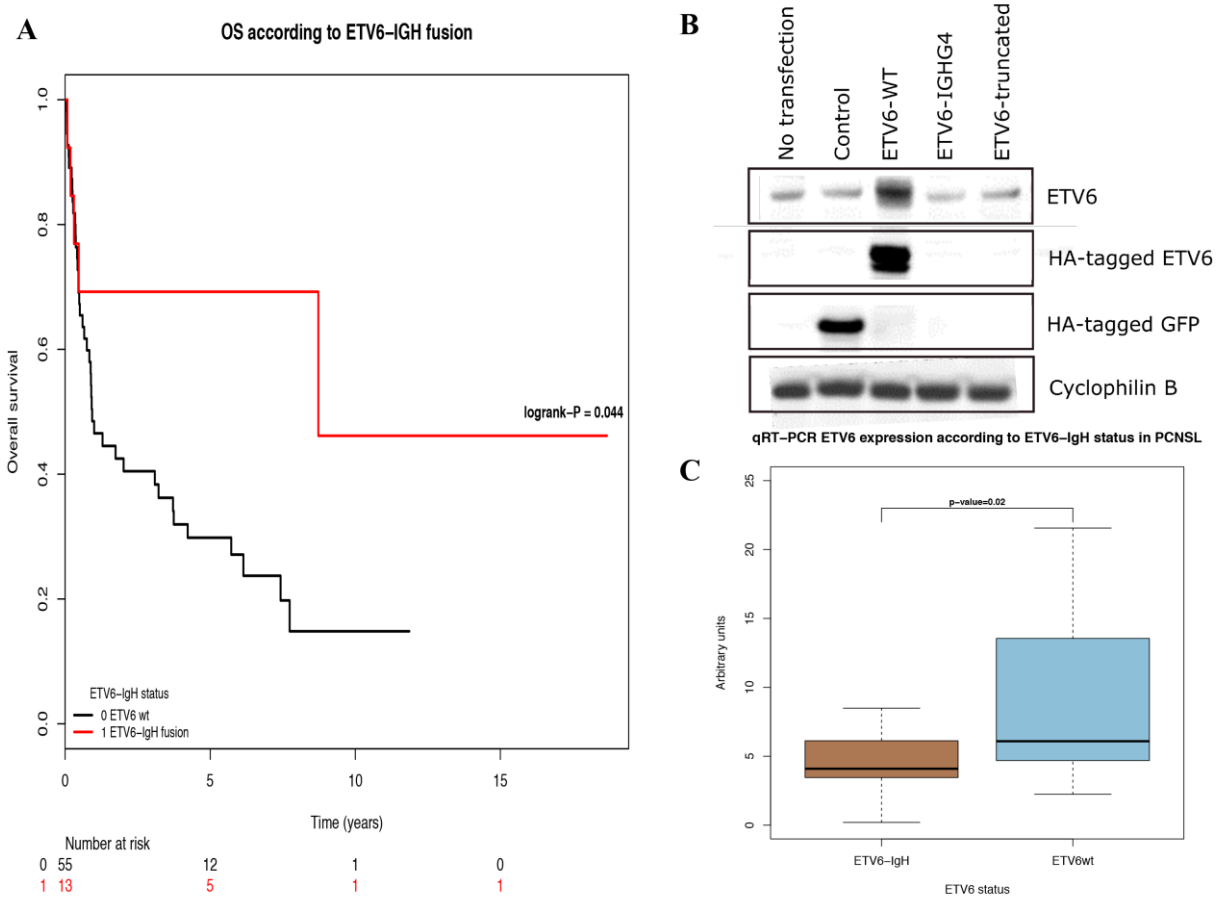
Figure 2.



Accepted

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

Figure 3.



Accepted

Downloaded from https://academic.oup.com/neuro-oncology/advance-article-abstract/doi/10.1093/neuonc/noy019/4843980
by University of California Santa Barbara/Davidson Library user
on 16 February 2018

SCIENTIFIC REPORTS

OPEN

Mendelian randomisation study of the relationship between vitamin D and risk of glioma

Received: 10 October 2017
Accepted: 24 January 2018
Published online: 05 February 2018

Hannah Takahashi[1], Alex J. Cornish[1], Amit Sud[1], Philip J. Law[1], Ben Kinnersley[1], Quinn T. Ostrom[2], Karim Labreche[1], Jeanette E. Eckel-Passow[3], Georgina N. Armstrong[4], Elizabeth B. Claus[5,6], Dora Il'yasova[7,8,9], Joellen Schildkraut[8,9], Jill S. Barnholtz-Sloan[2], Sara H. Olson[10], Jonine L. Bernstein[10], Rose K. Lai[11], Minouk J. Schoemaker[1], Matthias Simon[12], Per Hoffmann[13,14], Markus M. Nöthen[14,15], Karl-Heinz Jöckel[16], Stephen Chanock[17], Preetha Rajaraman[17], Christoffer Johansen[18], Robert B. Jenkins[19], Beatrice S. Melin[20], Margaret R. Wrensch[21,22], Marc Sanson[23,24], Melissa L. Bondy[4], Clare Turnbull[1,25,26] & Richard S. Houlston[1,27]

To examine for a causal relationship between vitamin D and glioma risk we performed an analysis of genetic variants associated with serum 25-hydroxyvitamin D (25(OH)D) levels using Mendelian randomisation (MR), an approach unaffected by biases from confounding. Two-sample MR was undertaken using genome-wide association study data. Single nucleotide polymorphisms (SNPs) associated with 25(OH)D levels were used as instrumental variables (IVs). We calculated MR estimates for the odds ratio (OR) for 25(OH)D levels with glioma using SNP-glioma estimates from 12,488 cases and 18,169 controls, using inverse-variance weighted (IVW) and maximum likelihood estimation (MLE) methods. A non-significant association between 25(OH)D levels and glioma risk was shown using

[1]Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK. [2]Case Comprehensive Cancer Center, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA. [3]Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, Minnesota, USA. [4]Department of Medicine, Section of Epidemiology and Population Sciences, Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, USA. [5]School of Public Health, Yale University, New Haven, Connecticut, USA. [6]Department of Neurosurgery, Brigham and Women's Hospital, Boston, Massachusetts, USA. [7]Department of Epidemiology and Biostatistics, School of Public Health, Georgia State University, Atlanta, Georgia, USA. [8]Duke Cancer Institute, Duke University Medical Center, Durham, North Carolina, USA. [9]Cancer Control and Prevention Program, Department of Community and Family Medicine, Duke University Medical Center, Durham, North Carolina, USA. [10]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [11]Departments of Neurology and Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. [12]Department of Neurosurgery, University of Bonn Medical Center, Sigmund-Freud-Str. 25, 53105, Bonn, Germany. [13]Human Genomics Research Group, Department of Biomedicine, University of Basel, Basel, Switzerland. [14]Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany. [15]Institute of Human Genetics, University of Bonn School of Medicine & University Hospital Bonn, Bonn, Germany. [16]Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen, University of Duisburg-Essen, Essen, Germany. [17]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, USA. [18]Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen, Denmark, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark. [19]Department of Laboratory Medicine and Pathology, Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, Minnesota, USA. [20]Department of Radiation Sciences, Umeå University, Umeå, Sweden. [21]Department of Neurological Surgery, School of Medicine, University of California, San Francisco, California, USA. [22]Institute of Human Genetics, University of California, San Francisco, California, USA. [23]Sorbonne Universités UPMC Univ Paris 06, INSERM CNRS, U1127, UMR 7225, ICM, F-75013, Paris, France. [24]AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de neurologie 2-Mazarin, Paris, France. [25]William Harvey Research Institute, Queen Mary University, London, UK. [26]Guys and St Thomas Foundation NHS Trust, Great Maze Pond, London, UK. [27]Division of Molecular Pathology, The Institute of Cancer Research, London, UK. Hannah Takahashi and Alex J. Cornish contributed equally to this work. Correspondence and requests for materials should be addressed to R.S.H. (email: richard.houlston@icr.ac.uk)

both the IVW (OR = 1.21, 95% confidence interval [CI] = 0.90–1.62, *P* = 0.201) and MLE (OR = 1.20, 95% CI = 0.98–1.48, *P* = 0.083) methods. In an exploratory analysis of tumour subtype, an inverse relationship between 25(OH)D levels and glioblastoma (GBM) risk was identified using the MLE method (OR = 0.62, 95% CI = 0.43–0.89, *P* = 0.010), but not the IVW method (OR = 0.62, 95% CI = 0.37–1.04, *P* = 0.070). No statistically significant association was shown between 25(OH)D levels and non-GBM glioma. Our results do not provide evidence for a causal relationship between 25(OH)D levels and all forms of glioma risk. More evidence is required to explore the relationship between 25(OH)D levels and risk of GBM.

While glioma accounts for around 80% of malignant primary brain tumours[1], thus far exposure to ionising radiation is the only well-established exogenous risk factor[2]. Vitamin D provides many health benefits, including increased bone strength and protection against autoimmune diseases and type 2 diabetes[3]. *In-vitro* studies have also suggested an anti-neoplastic role for vitamin D[4]. Several epidemiological studies have shown that vitamin D may indeed afford protection against the development of some cancers, including colon, prostate and breast cancer[5]. Associations in such observational studies do not however constitute evidence for a causal relationship and in some studies bias from confounding and reverse causation cannot be excluded.

Mendelian randomisation (MR) uses genetic markers as proxies for environmental exposures to determine the effect of the exposure on disease risk[6]. It therefore provides a strategy for establishing causal relationships where randomised control trials (RCTs) would involve either high cost or impractical study design. In the case of a possible relationship between vitamin D and glioma, the rarity of the cancer would limit any RCT to small sample sizes and would require lengthy follow up times.

We implemented two-sample MR analysis to examine the relationship between vitamin D and glioma risk in order to avoid the limitations of follow up time, reverse causation and confounding. Genotypes are randomly assigned at conception, thereby limiting confounding. Furthermore an individual's genotype will always be established before the onset of disease, excluding the possibility of reverse causation. The genotype is in part equivalent to a lifetime vitamin D deficiency, and hence a lifetime follow-up time in a RCT. We determine the relationship between vitamin D and glioma risk using genetic variants associated with 25(OH)D levels, rather than measuring 25(OH)D levels directly.

Genetic variants identified by the Study of Underlying Genetic Determinants of Vitamin D and Highly Related Traits (SUNLIGHT) Consortium[7] and the Canadian Multicentre Osteoporosis Study (CaMOS)[8] were used as an instrumental variable (IV). We performed an MR analysis to test for a causal relationship between 25(OH)D levels and glioma, using summary data from a recent genome-wide association study (GWAS) meta-analysis performed by the Glioma International Case-Control Consortium (GICC)[9].

Methods

Two-sample MR was undertaken using GWAS data. Ethical approval was not sought for this specific project because all data came from the summary statistics of previously published GWAS, and no individual-level data were used.

Genetic variant instruments for 25(OH)D level. Genetic variants used as IVs were selected from the previously published SUNLIGHT study[7]. The SUNLIGHT Consortium GWAS identified four genetic variants associated with lowered 25(OH)D levels in 33,996 individuals of European descent from 15 cohorts. These variants were rs2282679 in *GC* (vitamin D binding carrier protein), rs10741657 near *CYP2R1* (converter of vitamin D to the active ligand for the vitamin D receptor), rs12785878 near *DHCR7* (7-dehydrocholesterol synthesis from cholesterol, a precursor to vitamin D) and rs6013897 in *CYP24A1* (degrader of active 1,25-dihydroxyvitamin D3 to inactive vitamin D)[10]. The roles of GC, CYP2R1, DHCR7 and CYP24A1 in the vitamin D pathway are shown in Fig. 1. Association estimates (*per*-allele log-ORs) for SNPs were taken from previously published studies, which used data from the CaMOS study, a population based cohort study of 2,347 Canadians, genotyped and assayed for 25(OH)D levels[8,10,11]. None of the SNPs were in linkage disequilibrium (*i.e.* $r^2 \geq 0.001$). For each SNP, we recovered the chromosome position, risk allele, genetic locus, F-statistic and association estimates (Table 1). Standard errors (SE) were calculated from F-statistics calculated by previous studies, which derive from the CaMOS cohort[11]. The risk allele was taken to be the 25(OH)D decreasing allele. Allele frequencies for these SNPs were compared between the 25(OH)D and glioma data sets to ensure that the effect estimates were recorded with respect to the same allele. This study calculated the variants to account for about 2% of the variation in circulating 25(OH)D levels, and have a combined F-statistic of 12.57[12].

Glioma genotyping data. Association data between the four genetic variants and glioma were taken from the most-recent meta-analysis of GWAS in glioma[9], which related >10 million genetic variants (after imputation) to glioma (Supplementary Table 1). This meta-analysis comprised eight GWAS datasets of individuals of European descent: FRE, GER, GICC, MDA, GliomaScan (NIH), UCSF-Mayo, UCSF and UK (Supplementary Table 2). All diagnoses were confirmed in accordance with WHO guidelines. Full quality control details are provided in previously published work[9]. Gliomas are heterogeneous and different tumour subtypes, defined in part by malignancy grade (for example, pilocytic astrocytoma World Health Organization (WHO) grade I, diffuse 'low-grade' glioma WHO grade II, anaplastic glioma WHO grade III and glioblastoma (GBM) WHO grade IV) can be distinguished[13]. To avoid diagnostic ambiguity and for simplicity we considered glioma subtypes as being either GBM or non-GBM.

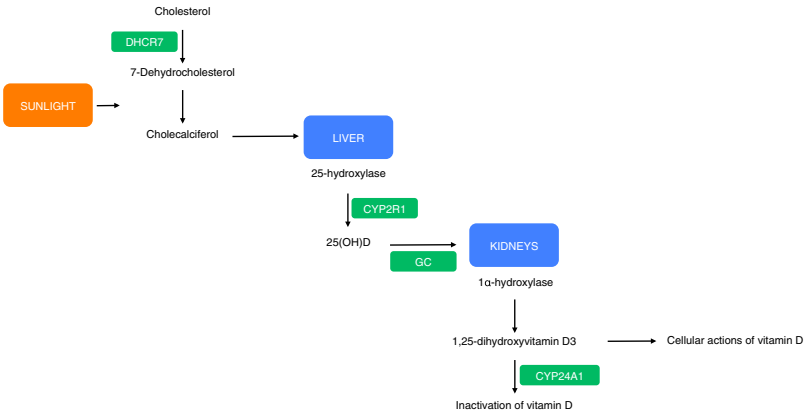


Figure 1. Effect of SNPs chosen as IVs on the vitamin D pathway. Genes that contain, or are in proximity to, variants chosen as IVs are highlighted green. *P* values for the association of these variants with 25(OH)D levels were 1.9×10^{-109} for *GC*, 2.1×10^{-27} for *DHCR7*, 3.3×10^{-20} for *CYP2R1*, and 6.0×10^{-10} for *CYP24A1*.

SNP ID	Chr	Locus	Base pair position	EA glioma	NEA glioma	EA 25(OH)D	NEA 25(OH)D	Effect on 25(OH)D	SE	F-statistic
rs2282679	4	*GC*	72608383	G	T	G	T	−0.047	0.013	13.38
rs10741657	11	Near *CYP2R1*	14914878	G	A	G	A	−0.052	0.012	18.78
rs12785878	11	Near *DHCR7*	71167449	T	G	G	T	−0.056	0.013	18.29
rs6013897	20	*CYP24A1*	52742479	A	T	A	T	−0.027	0.015	3.13

Table 1. Genetic variant instruments for 25(OH)D levels. EA, effect allele; NEA, non-effect allele; SE, standard error. Positions given using NCBI build 37. EA taken to be the 25(OH)D decreasing allele. Effect taken to be the per allele log OR effect on 25(OH)D.

Statistical analyses. We examined the association between circulating 25(OH)D levels and glioma (including subtypes) using MR on summary statistics using the inverse variance weighted (IVW) and maximum likelihood estimation (MLE) methods, as described by Burgess *et al.*[14]. The combined ratio estimate ($\hat{\beta}$) of all SNPs associated with 25(OH)D levels on glioma risk was calculated under a fixed-effects model:

$$\hat{\beta} = \sum_{i=1}^{k} \frac{X_k Y_k \sigma_Y^{-2}}{X_k^2 \sigma_Y^{-2}} \quad (1)$$

X_k is the association between SNP k with 25(OH)D levels, Y_k is the association between SNP k and glioma risk with standard error σ_Y. The standard error of this association is given by:

$$se(\hat{\beta}) = \sqrt{\sum_{i=1}^{k} \frac{1}{X_k^2 \sigma_Y^{-2}}} \quad (2)$$

We also conducted a likelihood based analysis using the same genetic summary data[15]. For this maximum likelihood estimate, a bivariate normal distribution for the genetic associations was assumed, and the R function *optim* was used to estimate β. $SE(\beta)$ was calculated using observed information.

With the estimates from the two analyses calculated for each of the eight cohorts in the glioma data, we performed a meta-analysis under a fixed-effect model to derive final odds ratios (ORs) and confidence intervals (CIs)[16].

To test whether the variants chosen as instruments were valid under MR assumptions, we examined the instruments for pleiotropy (multiple traits influenced by one gene) between the exposure and disease risk. This would be revealed as deviation from a linear relationship between SNPs and their effect size for 25(OH)D levels and glioma risk. We performed MR-Egger regression to test the average pleiotropic effect caused by the variants combined, as well as to provide a third association estimate between 25(OH)D level and glioma[17]. As per Dimitrakopoulou *et al.*[18], we further evaluated the presence of horizontal pleiotropy by conducting stratified MR analyses using only the genetic variants influencing vitamin D synthesis (rs12785878, rs10741657) and vitamin D metabolism (rs2282679, rs6013897). rs12785878 has been associated with non-European status[10] and we therefore also undertook a sensitivity analysis excluding rs12785878.

For each statistical test, we considered a global significance level of $P < 0.05$ as being satisfactory to derive conclusions. To assess the robustness of our conclusions, we imposed a conservative Bonferroni-corrected significance threshold of 0.017 (*i.e.* 0.05/3 tumour classifications).

	IVW method				MLE method			
	β	SE(β)	OR (95% CI)	*P* value	β	SE(β)	OR (95% CI)	*P* value
All glioma	0.189	0.148	1.21 (0.90–1.62)	0.201	0.184	0.106	1.20 (0.98–1.48)	0.083
GBM	−0.471	0.261	0.62 (0.37–1.04)	0.070	−0.479	0.186	0.62 (0.43–0.89)	0.010
Non-GBM	0.177	0.281	1.19 (0.69–2.07)	0.529	0.177	0.199	1.19 (0.81–1.76)	0.373

Table 2. MR estimates between multi-SNP risk scores of 25(OH)D levels and all glioma, GBM and non-GBM glioma using the IVW and MLE methods. IVW, inverse-variance weighted; MLE, maximum likelihood estimation; SE, standard error; OR, odds ratio; CI, confidence interval; GBM, glioblastoma.

	MR Egger slope		MR Egger intercept	
	Estimate (95% CI)	*P* value	Estimate (95% CI)	*P* value
All Glioma	0.072 (−0.121–0.264)	0.466	−0.001 (−0.019–0.017)	0.893
GBM	−0.097 (−0.272–0.078)	0.279	−0.013 (−0.039–0.012)	0.307
Non-GBM	0.160 (−0.114–0.434)	0.253	−0.005 (−0.035–0.026)	0.768

Table 3. MR-Egger test results for 25(OH)D levels and all glioma, GBM and non-GBM glioma. CI, confidence interval; GBM, glioblastoma.

The power of a MR investigation depends greatly on the proportion of variance in the risk factor that is explained by the IV. We therefore estimated study power to assess the strength of the results[19]. The detectable ORs at 80% power were 1.26 or 0.79 in the all glioma analysis, 1.34 or 0.75 in the GMB analysis and 1.35 or 0.74 in the non-GBM analysis. All power calculations were completed at a significance level of 0.05 and assumed the variants explained 2% of the total variance of 25(OH)D levels.

Data availability. Genotype data from the GICC GWAS are available from the database of Genotypes and Phenotypes (dbGaP; accession phs001319.v1.p1). Genotype data from the GliomaScan GWAS can also be accessed through dbGaP (accession phs000652.v1.p1). Data from the other studies are available upon request.

Results

The results of the IVW and MLE methods are summarised in Table 2. Results of the MR-Egger analysis are summarised in Table 3. Forest plots of all results from the IVW and MLE methods are shown in Figs 2 and 3. There was no evidence to support an association (*i.e.* $P > 0.05$) between circulating 25(OH)D levels and risk of all glioma using either the IVW (OR = 1.21, 95% CI = 0.90–1.62, $P = 0.201$) or MLE (OR = 1.20, 95% CI = 0.98–1.48, $P = 0.083$) methods. MR-Egger regression produced an intercept of −0.001 (95% CI = −0.019–0.017, $P = 0.893$) and therefore provided no evidence for pleiotropy amongst the genetic variants chosen as IVs (Supplementary Fig. 1). Hence there was no evidence of violation of MR assumptions.

We explored the possibility that a relationship between vitamin D and glioma may be subtype specific, considering GBM and non-GBM separately. We imposed a stronger significance threshold of $P = 0.017$ (*i.e.* 0.05/3), to correct for multiple testing. The MLE method identified an inverse relationship between 25(OH)D levels and risk of the GBM subtype, with an OR of 0.62 (95% CI = 0.43–0.89, $P = 0.010$). The IVW method provided a similar, but non-significant effect size (OR = 0.62, 95% CI = 0.37–1.04, $P = 0.070$). No evidence for an association between 25(OH)D levels and the non-GBM subtype was identified using either the IVW or MLE methods. MR-Egger regression provided intercepts of −0.013 (95% CI = −0.039–0.012, $P = 0.307$) for GBM and −0.005 (95% CI = −0.035–0.026, $P = 0.768$) for non-GBM, again providing no evidence of pleiotropy.

Stratified MR analyses using separate allelic scores for vitamin D synthesis and metabolism did not indicate the presence of horizontal pleiotropy (Supplementary Tables 3 and 4). To address the potential effects of population stratification, we undertook a MR sensitivity analysis excluding rs12785878, as this SNP has been associated with non-European status[10] (Supplementary Table 5). Excluding rs12785878, the inverse relationship between 25(OH)D levels and risk of the GBM subtype identified by the MLE method remains significant (OR = 0.51, 95% CI = 0.33–0.80, $P = 0.003$), thereby providing no evidence that this association is a result of population stratification.

Discussion

To our knowledge, this is the first MR study evaluating the effect of vitamin D on glioma risk undertaken. Overall our results do not provide evidence for an effect of vitamin D on risk of all forms of glioma. They do however raise the possibility for a protective role of vitamin D in GBM. While vitamin D and its metabolites have been shown to induce death of glioblastoma cells[20–22], only one epidemiological study has investigated the relationship between pre-diagnostic levels of 25(OH)D and glioma risk[23]. Researchers found that higher levels of 25(OH)D were protective against high-grade glioma in men over the age of 56 (OR = 0.59), although the reverse trend was shown in men under the age of 56, albeit at a borderline-significant level[23]. Excluding the possibility of post hoc data mining, such paradoxical findings would support distinct aetiologies between the GBM and non-GBM subtypes, as has been suggested previously[9].

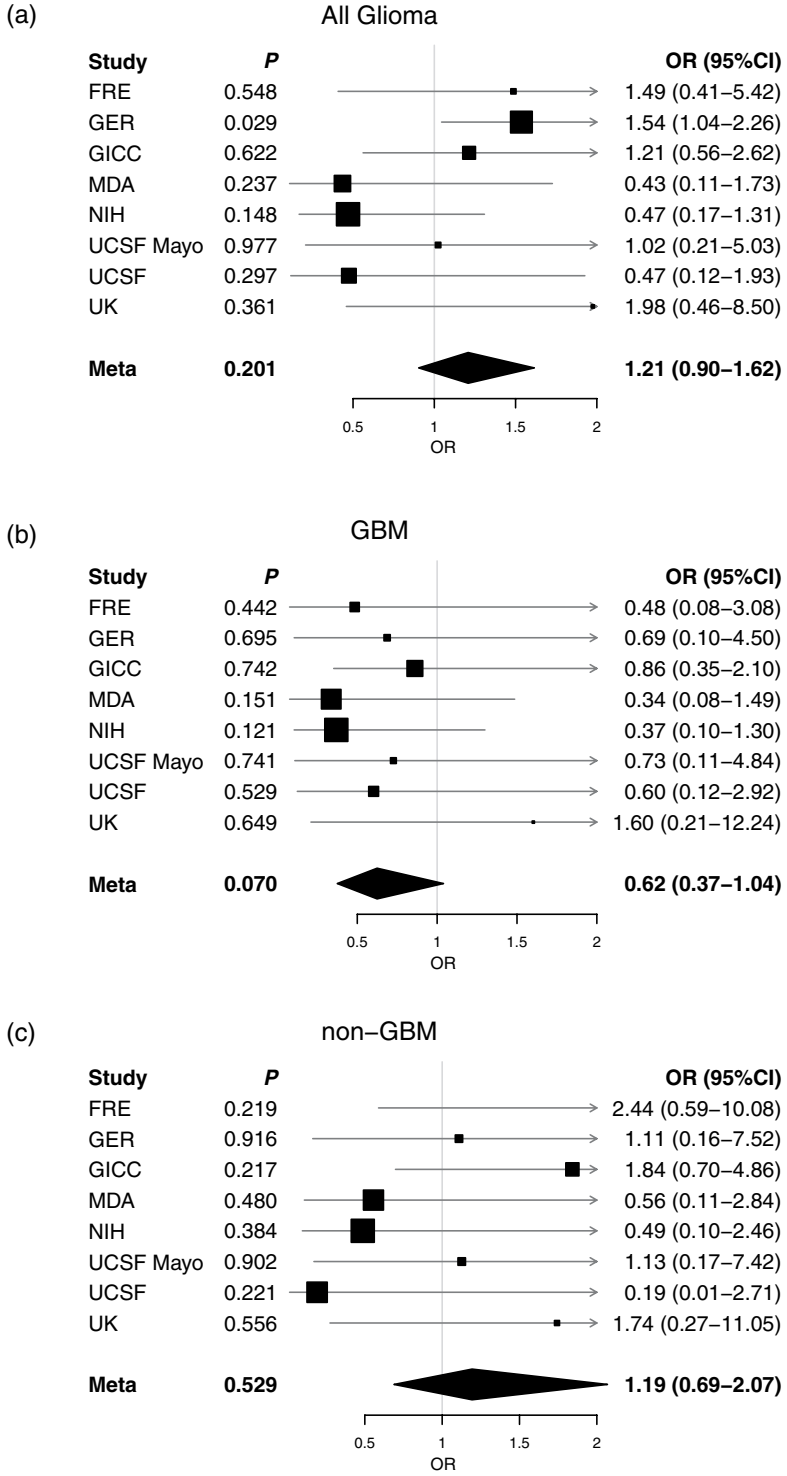


Figure 2. Individual cohort and meta-analysis ORs calculated using the IVW method. (**a**) All glioma, (**b**) GBM and (**c**) non-GBM glioma. Boxes are OR point estimates with area proportional to the weight of the study. Diamonds are overall summary estimates, with 95% CIs given by the width. Vertical line is null value (OR = 1.0).

Vital to the method of statistical analysis used herein is that none of the MR assumptions are violated. This requires that the variants chosen as IVs are (i) strongly associated with the exposure, (ii) are not associated with any confounding effects between exposure and outcome and (iii) are only associated with the outcome via the exposure. With regard to this study, the instruments chosen were associated with 25(OH)D levels at genome-wide significance levels. The MR-Egger test provided no evidence of horizontal pleiotropy, which we deemed sufficient to satisfy the third assumption. Furthermore, none of the four SNPs were in linkage disequilibrium

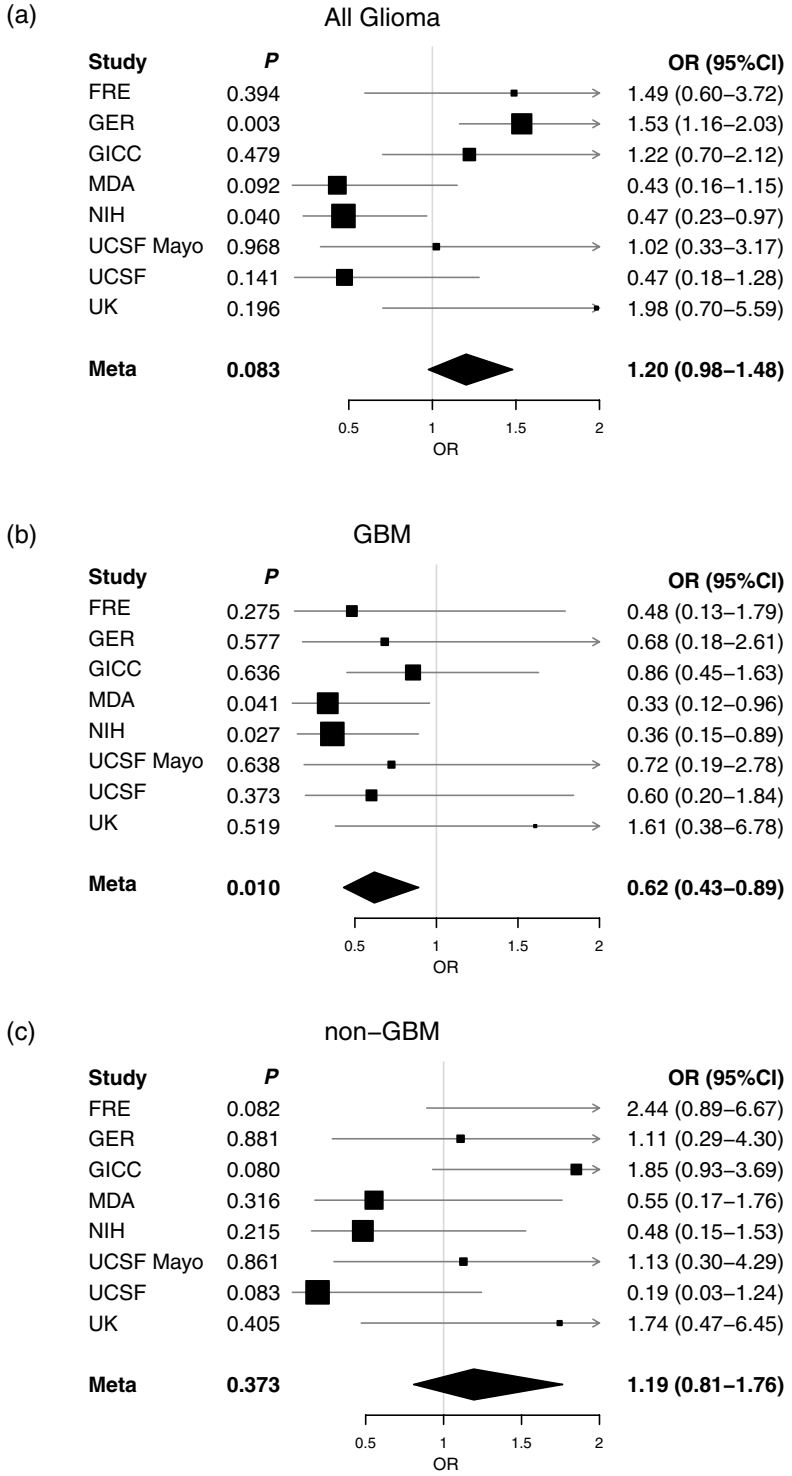


Figure 3. Individual cohort and meta-analysis ORs calculated using the MLE method. (**a**) All glioma, (**b**) GBM and (**c**) non-GBM glioma. Boxes are OR point estimates with area proportional to the weight of the study. Diamonds are overall summary estimates, with 95% CIs given by the width. Vertical line is null value (OR = 1.0).

(*i.e.* $r^2 \geq 0.001$) with any of the variants identified by Melin *et al.*[9] as being in the risk region for glioma. With regard to confounding factors, few risk factors are known for glioma, so it was not possible to entirely rule out the possibility of unknown confounding factors causing statistical bias. However it should also be noted that all four SNPs lie either within or near genetic loci whose function in vitamin D physiology is well understood[7], although a lack of knowledge of possible confounding factors means it was not possible to entirely rule out the possibility of confounding by unknown factors.

We acknowledge that a weakness of our study was in the small percentage of variability (around 2%) in 25(OH)D levels explained by the IV. Such a low value means any interpretation of these results as true indicators of the effect of total 25(OH)D levels on glioma risk are limited. This is quantified by the high ORs required for sufficient study power. Furthermore the study only accounts for circulating 25(OH)D levels and not for the action of 25(OH)D at the cellular level[11]. The genetic variants used as IVs in this MR analysis associate with 25(OH)D levels, rather than levels of the biologically active 1,25-dihydroxyvitamin D (1,25(OH)2D) and we therefore cannot explicitly comment on the relationship between 1,25(OH)2D and glioma. The low OR found in the GBM analysis should be noted however, given the fairly consistent indications of protective effects of 25(OH)D across all three methods. As is generally the case with MR, any findings should be viewed as a compliment to other future epidemiological studies, which test more robustly for associations between vitamin D and glioma and its subtypes.

In conclusion our MR analysis provides no evidence for an association between vitamin D and glioma, though findings raise the possibility of a potential association between vitamin D and GBM warranting further investigation.

References

1. Dolecek, T. A., Propp, J. M., Stroup, N. E. & Kruchko, C. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2005-2009. *Neuro Oncol* **14**(Suppl 5), v1–49 (2012).
2. Ostrom, Q. T. *et al*. The epidemiology of glioma in adults: a "state of the science" review. *Neuro Oncol* **16**, 896–913 (2014).
3. Wang, S. Epidemiology of vitamin D in health and disease. *Nutr Res Rev* **22**, 188–203 (2009).
4. Feldman, D., Krishnan, A. V., Swami, S., Giovannucci, E. & Feldman, B. J. The role of vitamin D in reducing cancer risk and progression. *Nat Rev Cancer* **14**, 342–57 (2014).
5. Toner, C. D., Davis, C. D. & Milner, J. A. The vitamin D and cancer conundrum: aiming at a moving target. *J Am Diet Assoc* **110**, 1492–500 (2010).
6. Sheehan, N. A., Didelez, V., Burton, P. R. & Tobin, M. D. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Med* **5**, e177 (2008).
7. Wang, T. J. *et al*. Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *Lancet* **376**, 180–8 (2010).
8. Langsetmo, L. *et al*. Calcium and vitamin D intake and mortality: results from the Canadian Multicentre Osteoporosis Study (CaMos). *J Clin Endocrinol Metab* **98**, 3010–8 (2013).
9. Melin, B. S. *et al*. Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nat Genet* **49**, 789–794 (2017).
10. Mokry, L. E. *et al*. Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study. *PLoS Med* **12**, e1001866 (2015).
11. Manousaki, D. *et al*. Vitamin D levels and susceptibility to asthma, elevated immunoglobulin E levels, and atopic dermatitis: A Mendelian randomization study. *PLoS Med* **14**, e1002294 (2017).
12. Pierce, B. L., Ahsan, H. & Vanderweele, T. J. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int J Epidemiol* **40**, 740–52 (2011).
13. Louis, D. N. *et al*. The2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* **131**, 803–20 (2016).
14. Burgess, S. *et al*. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol* **30**, 543–52 (2015).
15. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol* **181**, 251–60 (2015).
16. de Bakker, P. I. *et al*. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* **17**, R122–8 (2008).
17. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512–25 (2015).
18. Dimitrakopoulou, V. I. *et al*. Circulating vitamin D concentration and risk of seven cancers: Mendelian randomisation study. *BMJ* **359**, j4761 (2017).
19. Brion, M. J., Shakhbazov, K. & Visscher, P. M. Calculating statistical power in Mendelian randomization studies. *Int J Epidemiol* **42**, 1497–501 (2013).
20. Magrassi, L., Butti, G., Pezzotta, S., Infuso, L. & Milanesi, G. Effects of vitamin D and retinoic acid on human glioblastoma cell lines. *Acta Neurochir (Wien)* **133**, 184–90 (1995).
21. Magrassi, L. *et al*. Vitamin D metabolites activate the sphingomyelin pathway and induce death of glioblastoma cells. Acta Neurochir (Wien) **140**, 707–13; discussion 713–4 (1998).
22. Garcion, E., Wion-Barbot, N., Montero-Menei, C. N., Berger, F. & Wion, D. New clues about vitamin D functions in the nervous system. *Trends Endocrinol Metab* **13**, 100–5 (2002).
23. Zigmont, V. *et al*. Association Between Prediagnostic Serum 25-Hydroxyvitamin D Concentration and Glioma. *Nutr Cancer* **67**, 1120–30 (2015).

Acknowledgements

HT was supported by a Wellcome Trust Summer Student bursary. AS is supported by a Cancer Research UK clinical Fellowship. In the UK, funding was provided by Cancer Research UK (C1298/A8362) supported by the Bobby Moore Fund. The GICC was supported by grants from the National Institutes of Health, Bethesda, Maryland (R01CA139020, R01CA52689, P50097257, P30CA125123). The UK Interphone Study was supported by the European Commission Fifth Framework Program "Quality of Life and Management of Living Resources" and the UK Mobile Telecommunications and Health Programme. The Mobile Manufacturers Forum and the GSM Association provided funding for the study through the scientifically independent International Union against Cancer (UICC).

Author Contributions

R.S.H. and A.J.C. managed the project. H.T., A.J.C., A.S., P.J.L. and R.S.H. drafted the manuscript. H.T. and A.J.C. performed statistical analyses. B.K., K.L., M.J.S. and R.H.S. provided U.K. data. M. Simon, P.H., M.M.N. and K.-H.J. provided German data. Q.T.O., J.E.E.-P., G.N.A., E.B.C., D.I., J.S., J.S.B.-S., S.H.O., J.L.B., R.K.L., C.J., R.B.J., B.S.M., M.R.W., M.L.B. and R.S.H. provided GICC data. S.C. and P.R. provided National Cancer Institute (NCI) data. M. Sanson provided French data. All authors reviewed the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-20844-w.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018

Acta Neuropathol (2017) 134:691–703
DOI 10.1007/s00401-017-1743-5

METHODS PAPER

Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing

Philipp Euskirchen[1,2,3] · Franck Bielle[1,4,5] · Karim Labreche[1,6] · Wigard P. Kloosterman[7] · Shai Rosenberg[1] · Mailys Daniau[1] · Charlotte Schmitt[1] · Julien Masliah-Planchon[8] · Franck Bourdeaut[10] · Caroline Dehais[9] · Yannick Marie[1] · Jean-Yves Delattre[1,9] · Ahmed Idbaih[1,9]

Received: 3 March 2017 / Revised: 7 June 2017 / Accepted: 10 June 2017 / Published online: 21 June 2017
© The Author(s) 2017. This article is an open access publication

Abstract Molecular classification of cancer has entered clinical routine to inform diagnosis, prognosis, and treatment decisions. At the same time, new tumor entities have been identified that cannot be defined histologically. For central nervous system tumors, the current World Health Organization classification explicitly demands molecular testing, e.g., for 1p/19q-codeletion or IDH mutations, to make an integrated histomolecular diagnosis. However, a plethora of sophisticated technologies is currently needed to assess different genomic and epigenomic alterations and turnaround times are in the range of weeks, which makes standardized and widespread implementation difficult and hinders timely decision making. Here, we explored the potential of a pocket-size nanopore sequencing device for multimodal and rapid molecular diagnostics of cancer. Low-pass whole genome sequencing was used to simultaneously generate copy number (CN) and methylation profiles from native tumor DNA in the same sequencing run. Single nucleotide variants in *IDH1*, *IDH2*, *TP53*, *H3F3A,* and the *TERT* promoter region were identified using deep amplicon sequencing. Nanopore sequencing yielded ~0.1X genome coverage within 6 h and resulting CN and epigenetic profiles correlated well with matched microarray data. Diagnostically relevant alterations, such as 1p/19q codeletion, and focal amplifications could be recapitulated. Using ad hoc random forests, we could perform supervised pan-cancer classification to distinguish gliomas, medulloblastomas, and brain metastases of different primary sites. Single nucleotide variants in *IDH1*, *IDH2,* and *H3F3A* were identified using deep amplicon sequencing within minutes of sequencing. Detection of *TP53* and *TERT* promoter mutations shows that sequencing of entire genes and GC-rich regions is feasible. Nanopore sequencing allows same-day detection of structural variants, point mutations, and methylation profiling using a single device with negligible capital cost. It outperforms hybridization-based and current sequencing technologies with respect to time to diagnosis and required laboratory equipment and expertise, aiming to

Electronic supplementary material The online version of this article (doi:10.1007/s00401-017-1743-5) contains supplementary material, which is available to authorized users.

✉ Philipp Euskirchen
philipp.euskirchen@charite.de

✉ Ahmed Idbaih
ahmed.idbaih@aphp.fr

1 Inserm U 1127, CNRS UMR 7225, Sorbonne Universités, UPMC Univ Paris 06 UMR S 1127, Institut du Cerveau et de la Moelle épinière (ICM), Paris, France

2 Department of Neurology, Charité-Universitätsmedizin Berlin, Berlin, Germany

3 Berlin Institute of Health (BIH), Berlin, Germany

4 Service de Neuropathologie, AP-HP, Hôpitaux Universitaires La Pitié Salpêtrière-Charles Foix, Paris, France

5 OncoNeuroTek, Paris, France

6 Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK

7 Division of Biomedical Genetics, Center for Molecular Medicine, Department of Genetics, University Medical Center Utrecht, Utrecht, The Netherlands

8 Department of Genetics, Institut Curie, PSL Research University, Paris, France

9 Service de Neurologie, AP-HP, Hôpitaux Universitaires La Pitié Salpêtrière-Charles Foix, 2-Mazarin, Paris, France

10 Laboratory of Translational Research in Pediatric Oncology, Institut Curie, PSL Research University, Paris, France

make precision medicine possible for every cancer patient, even in resource-restricted settings.

Keywords Nanopore sequencing · Brain tumor · Glioma · Whole genome sequencing · Epigenomics · Molecular neuropathology

Introduction

Histomolecular classification of brain tumors has entered clinical routine diagnostics as the current World Health Organization (WHO) classification explicitly demands histological findings to be refined by molecular testing [20]. Thus, pathologists rely on timely and accurate molecular testing to make an integrated diagnosis using both in situ methods and genetic information. However, high turnaround time of current implementations delays integrated diagnosis by weeks. In addition, targeted next-generation sequencing panels, microarray-based analysis of copy number (CN), and epigenetic alterations all provide high-quality data and aid in the diagnosis and therapeutic management of patients (i.e., stratification or identification of actionable targets or inclusion in clinical trials), but their high capital cost, demanding workflows and need for highly skilled personnel hinder their widespread use. Here, we demonstrate that real-time molecular genomics using nanopore sequencing is both fast and reliable to aid diagnosing cancer by unsupervised classification of CN and methylation profiles.

Nanopore sequencing interprets changes in ionic currents observed when single DNA molecules pass through a nanometer-size protein pore. This has led to the development of handheld size devices that allow sequencing outside of classical laboratory settings and even in the field [27]. While overall throughput currently lacks behind other deep sequencing technologies, nanopores allow read analysis in real-time and selective sequencing [19], both of which allow rapid generation of data. In addition, nanopores are able to discriminate not only the nucleotides of a strand of DNA but also single base modifications such as 5-methylation of cytosine [29, 35]. This allows concurrent analysis of sequence identity and methylation using native DNA.

Materials and methods

Experimental design

We performed a retrospective observational study for molecular characterization of diagnostically relevant genetic alterations using nanopore sequencing. Patients were recruited at the Pitié-Salpêtrière university hospital and have given informed consent for research use of tumor material, including genotyping. All tumor samples have been molecularly characterized previously using short-read exome sequencing, Sanger sequencing, SNP array, and/or genome-wide methylation microarray [14, 30].

Nanopore whole genome sequencing

DNA quality of fresh-frozen tumor tissue was determined using NanoDrop (Thermo Fisher Scientific) and samples were quantified using a QuantiFluor dsDNA assay (Promega, Madison, WI, USA). For whole genome sequencing, libraries were prepared using Rapid 1D Sequencing Kit (SQK-RAD001, SQK-RAD002, or SQK-RBK001, Oxford Nanopore Technologies, UK) following the manufacturer's instructions. Briefly, 200 ng of tumor DNA was fragmented using a transposase and subjected to adapter ligation. Sequencing was performed using R9 or R9.4 flow cells on a MinION Mk 1B device (Oxford Nanopore) with the MinKNOW software (versions 1.0.5–1.5.12), respectively. For samples run with R9.4 sequencing chemistry, basecalling was performed using Albacore 1.1.0 (Oxford Nanopore). For R9 chemistries, online EPI2ME basecalling (Metrichor Ltd, Oxford, UK) was performed.

Template reads were exported as FASTA using nanopolish or poretools version 0.6 [18] and aligned to the hg19 human reference genome using BWA MEM 0.7.12 with the "−x ont2d" option [17]. Due to compatibility issues of data generated with R9 chemistries, only samples with R9.4 flow cells were used for copy number analysis and methylation-based classification.

Copy number analysis

For copy number analysis, the QDNAseq package version 1.8.0 [33] and R/Bioconductor, version 3.3, were used. Reads with a minimum mapping quality of 20 were sorted into 1000 kbp bins. Bins with missing reference sequence were excluded from analysis. To account for region- and technology-specific artifacts, public nanopore WGS data for the NA12878 human reference genome were processed identically and subtracted from the normalized tumor sample bin counts. Circular binary segmentation was performed as implemented in the DNAcopy package requiring an alpha value <0.05 to accept change points. Arm-level copy number calls were made by calculating the segment length weighted mean log ratio per chromosome arm.

Methylation analysis

To identify 5-methylation of cytosines, we used a recently published algorithm based on a hidden Markov model

which has been trained using in vitro methylated *E. coli* DNA [35]. Training models for R9 sequencing chemistries were kindly provided by Jared Simpson. We modified the original implementation of nanopolish 0.6.0 to allow methylation calling from different basecalling groups. For classification, the subset of CpG sites overlapping with sites covered by Illumina 450K BeadChip microarrays was used. Beta values in the training set were dichotomized using a cut-off value of 0.6.

Structural variant detection

For detection of structural variants in amplified regions, we aligned nanopore FASTQ files from sample 3427T to the human reference genome, build GRCh37, using LAST (version 744) with settings: $-Q$ 0. The *last-train* function was used with 1000 nanopore reads (~10 million bases) as input to adapt the alignment scoring parameters ($-p$) for error-prone nanopore reads. LAST alignment files (MAF) were converted to BAM files using the *maf-convert* function. BAM files were used as input for NanoSV [36] (available at https://github.com/mroosmalen/nanosv) with default settings.

Amplicon sequencing

Amplicons were designed to cover one or multiple exons of canonical transcripts of *IDH1*, *IDH2*, *TP53*, *H3F3A*, and the *TERT* promoter region. Oligonucleotide primers (Thermo Fisher Scientific) were then designed using Primer3 with the following non-default parameters (T_{min} 59 °C, T_{opt} 60 °C, T_{max} 61 °C, and maximum mononucleotide repeat length = 3) to yield product sizes of 489–2902 bp (Table S1).

25 ng of genomic DNA was amplified using 0.02 U/µl Q5 polymerase (New England Biolabs, Ipswich, MA, USA), 200 µM dNTPs, 500 nM forward and reverse primers, and Q5 reaction buffer with high GC enhancer in a total reaction volume of 20 µl. Thermal cycling was performed as follows: 98 °C initial denaturation for 2 min, followed by 30 cycles of denaturation at 98 °C for 10 s, annealing at 65 °C for 20 s and extension at 72 °C for 90 s, as well as a final extension at 72 °C for 2 min. Amplicons were analyzed using a Caliper LabChip GX DNA 5K assay (Perkin Elmer, Waltham, MA, USA). PCR products were purified using NucleoFast 96 PCR plates (Macherey–Nagel, Düren, Germany).

For amplicon sequencing, Ligation Sequencing Kit 1D (SQK-LSK108, Oxford) was used following the manufacturer's protocol. Briefly, 1 µg of pooled amplicon DNA was subjected to end repair and dA-tailing. 250 ng of end-repaired DNA (equivalent to 0.2 pmol of 2 kbp fragments) was then used as input for adapter ligation. For real-time monitoring of sequencing depth, reads were streamed to the BWA aligner using npReader [6] with jHDF5 2.11.0 and coverage was calculated using BEDTools [28]. For variant calling, reads were realigned on the event level and variants called using VarScan 2.4.3 [15]. Variants were annotated using SnpEff version 4.3i [9] and ExAC release 0.3.1 germline variants [16] before filtering for coding or hotspot mutations with a minimum mutant allele frequency >0.2.

Microarray methylation profiling

Samples for Illumina Infinium BeadChip 450K profiling were prepared as described before [14]. Briefly, 500 ng of DNA was subjected to bisulfite conversion. Hybridization and imaging were performed by IntegraGen (Evry, France). Raw IDAT files were preprocessed using the GenomeStudio software (Illumina, San Diego, CA, USA). Processed methylation data from previously characterized samples [14] were retrieved via ArrayExpress (accession E-MTAB-3903). Beta values were used for all the subsequent analysis steps.

Statistics

All data analysis was done using R/Bioconductor version 3.3 [13]. Hierarchical clustering was used for arranging probes in the depicted classification training set. Random forest classification as implemented in the R/randomForest package, version 4.6–12, was run with default parameters. Sequence concordance was calculated using the Genome Analysis Toolkit's Genotype Concordance tool, version 3.7 [21].

Data and material availability

Raw sequencing data are available via the European Genome–phenome Archive (accession EGAS00001002213). Microarray-based methylome data are available at ArrayExpress (E-MTAB-5797). TCGA data were retrieved from the UCSC Cancer Browser [11] or the TCGA FireBrowse website (http://www.firebrowse.org). Pipelines, scripts, and supplementary data to reproduce all results presented in this work are available at https://gitlab.com/pesk/glioma.nano-seq.

Results

To meet the needs of the WHO 2016 classification of CNS tumors, we designed 1-day workflows for CN, methylation, and point mutation profiling using nanopore sequencing (Fig. 1a). We first subjected tumor DNA from molecularly well-characterized brain tumors [14, 30] to

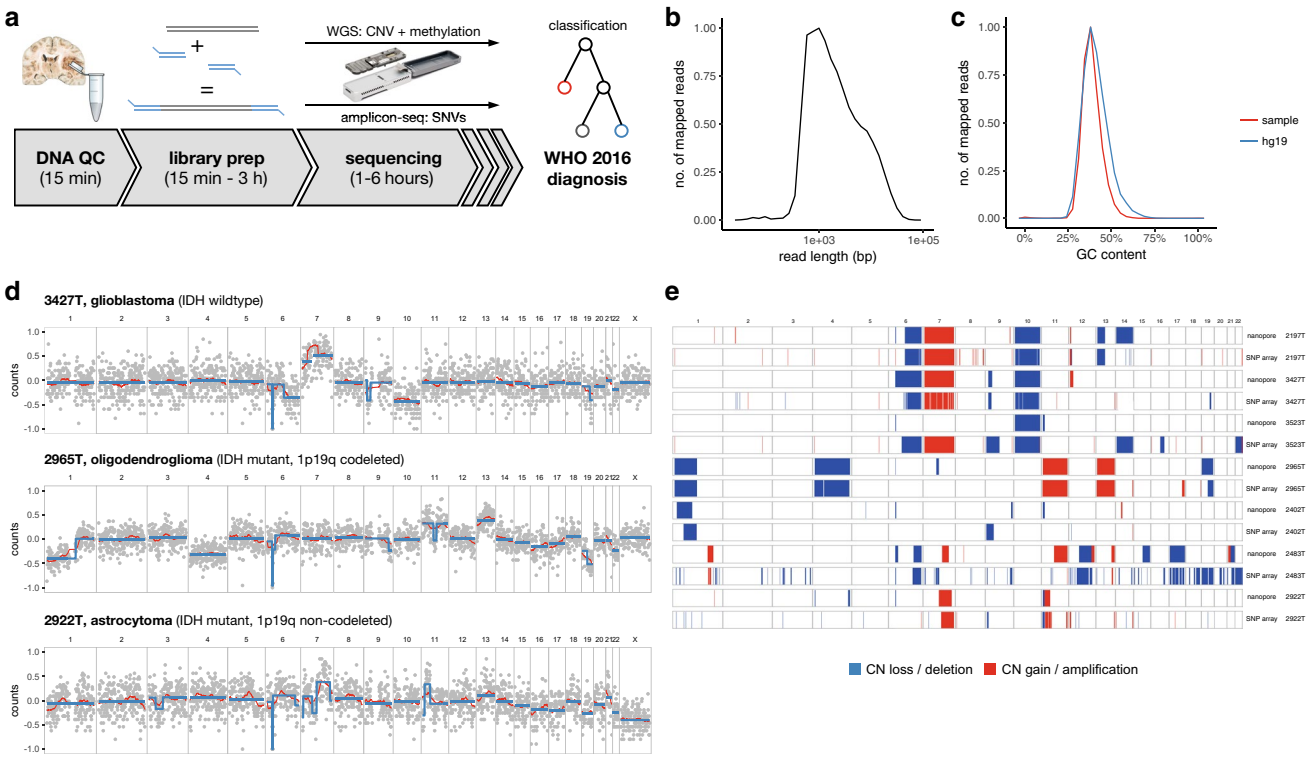


Fig. 1 Copy number profiling using nanopore low-pass whole genome sequencing. **a** Same-day workflows to simultaneously characterize copy number variation (CNV) and methylation profiles or single nucleotide variants, respectively. Tumor DNA is subjected to quality control (QC), and then, 250 ng input material is used for library preparation for either whole genome sequencing (WGS) or PCR-based deep amplicon sequencing. **b** Representative read length distribution of mapped reads. Note log scale on *X* axis. **c** Representative distribution of GC content of reads in comparison with the hg19 human reference genome. A randomly drawn subsample of the entire reference genome split into 1000 bp fragments is shown. **d** Copy number profile showing $\log_2$ transformed, normalized read counts per 1000 kbp window (*grey*) with running mean (*red*) and segmentation results (*blue*). **e** Comparison of nanopore WGS with matched SNP arrays. Heatmaps indicate copy number calls (losses and deletions in *blue*, and gains and amplifications in *red*) across the genome

low-pass whole genome sequencing (WGS) using a commercially available, handheld size nanopore sequencing device. With the aim of widespread implementation in routine diagnostics in mind, we used a transposon-based library preparation kit, which reduces sample preparation time to less than hour. In a cohort of 28 patients (Table 1), low-pass WGS for 6 h performed yielded a mean mapped read depth from <0.01X to 0.24X (Table S1), depending on the sequencing chemistry and input DNA fragment size. Nanopores decipher DNA sequence of single molecules as they present to the pore, generating long reads of variable length, whose distribution is determined by DNA extraction and fragmentation method. We observed typical mean read lengths around 2 kb (Fig. 1b). As library preparation does not involve PCR amplification, no GC bias is introduced and the GC content distribution of the reads resembles closely that of the human reference genome (Fig. 1c).

Copy number profiling

We then used WGS data to generate CN profiles. Reads were counted in 1000 kb windows, normalized and subjected to circular binary segmentation (Fig. 1c). No correction of GC bias or mappability is necessary for nanopore reads; however, the long reads cause alignment artifacts with current reference genomes in regions with repetitive sequence such as centromeres. Still, the resulting CN profiles closely resembled matched SNP array-based profiles (Fig. 1d). Importantly, codeletion of chromosome 1p/19q as a diagnostic criterion for oligodendrogliomas implemented in the 2016 WHO classification of CNS tumors was detected in three out of four affected samples (Fig. S1). The remaining sample did not yield sufficient read depth (<0.01) due to low input DNA quality (Table S1). High-level focal amplifications of *EGFR*, *PDGFRA*, and *CDK4* were detected in affected glioblastoma samples (Table 1).

Table 1 Clinical characteristics of patients in study

ID	Age at diagnosis	Sex	WHO 2016 integrated diagnosis	Nanopore sequencing performed	Nanopore methylation-based classification	Key alterations identified by nanopore sequencing
3523T	70	F	Glioblastoma, IDH-wildtype	WGS, amplicon	Not classifiable	pTERT C228T
2197T	58	F	Glioblastoma, IDH-wildtype	WGS, amplicon	Glioma, IDH-wildtype	TP53 p.S241F, pTERT C228T
3427T	72	F	Glioblastoma, IDH-wildtype	WGS, amplicon	Glioma, IDH-wildtype	pTERT C228T, CDKN2A^{loss}, EGFR^{amp}
2402T	58	M	Anaplastic oligodendroglioma, IDH-mutant, and 1p/19q-codeleted	WGS, amplicon	Not classifiable	IDH1 p.R132H, 1p/19q codeletion, pTERT C228T
2965T	29	F	Anaplastic oligodendroglioma, IDH-mutant and 1p/19q-codeleted	WGS, amplicon	Glioma, IDH-mutant	IDH1 p.R132H, 1p/19q codeletion, pTERT C228T
2483T	51	F	Anaplastic astrocytoma, IDH-mutant	WGS, amplicon	Glioma, IDH-mutant	IDH1 p.R132C TP53 p.R273C, p.R282Q
2922T	44	M	Diffuse astrocytoma, IDH-mutant	WGS	Glioma, IDH-mutant	N/D
6228T	33	F	Diffuse midline glioma, H3.3 K27M-mutant	WGS, amplicon	Classifiable	PDGFRA^{amp}
5337T	21	M	Glioma H3.3 G34R	WGS, amplicon	Glioma IDH-wildtype	H3F3A G34R, CDK4^{amp}, PDGFRA^{amp}
8347T	28	M	Desmoplastic/nodular medulloblastoma, SHH-activated and TP53 wild type	Amplicon	N/D	pTERT C228T
8372T	25	M	Classic medulloblastoma, non-WNT/non-SHH	WGS, amplicon	Medulloblastoma, group 4	pTERT C228T
MB683	7	F	Classic medulloblastoma, WNT-activated	WGS, amplicon	Medulloblastoma, WNT-activated	chr6 loss
8137T	48	M	Anaplastic oligodendroglioma, IDH-mutant and 1p/19q-codeleted	WGS, amplicon	Glioma, IDH-mutant	IDH2 p.R172 W, 1p/19q codeletion, pTERT C228T
8146T	N/A	F	Anaplastic oligodendroglioma, IDH-mutant and 1p/19q-codeleted	WGS, amplicon	Glioma, IDH-mutant	pTERT C228T
7382T	76	F	Glioblastoma, IDH-wildtype	WGS, amplicon	Glioma, IDH-wildtype	pTERT C228T, PDGFRA^{amp} TP53 p.V197M
7455T	45	M	Glioblastoma, IDH-wildtype	WGS, amplicon	Glioma, IDH-wildtype	pTERT C228T
8355T	56	M	Glioblastoma, IDH-wildtype	WGS	Not classifiable	N/D
8356T	73	F	Breast adenocarcinoma, GFAP+, S100+	WGS	Breast cancer	N/D
8357T	79	M	Neuro-endrocrine (prostate adeno) carcinoma, TTF1+	WGS	Lung cancer	N/D
8358T	63	F	Lung adenocarcinoma	WGS	Lung cancer	N/D
8359T	51	M	Bladder urothelial carcinoma	WGS, amplicon	Not classifiable	TP53 p.R280 K
8360T	65	F	Lung adenocarcinoma	Amplicon	N/D	TP53 p.I195T
4596T FFPE	44	F	Anaplastic oligodendroglioma, IDH-mutant and 1p/19q-codeleted	WGS, amplicon	Not classifiable	pTERT C228T

Table 1 continued

ID	Age at diagnosis	Sex	WHO 2016 integrated diagnosis	Nanopore sequencing performed	Nanopore methylation-based classification	Key alterations identified by nanopore sequencing
5539T FFPE	28	M	Anaplastic astrocytoma, IDH-mutant	Amplicon	N/D	pTERT C228T[¶]
3718T	78	F	Glioblastoma, IDH-wildtype	WGS	N/D	N/D
3719T	74	M	Glioblastoma, IDH-wildtype	WGS	N/D	N/D
2211T	75	F	Glioblastoma, IDH-wildtype	WGS	N/D	N/D
3724T	65	M	Glioblastoma, IDH-wildtype	WGS	N/D	N/D

Age at initial diagnosis, integrated diagnosis and the type of nanopore sequencing performed are reported. Results of methylation-based random forest classification and key genetic alterations identified by WGS or amplicon sequencing are indicated. Samples were considered not classifiable when there was less than 5 percentage points difference of the majority vote to the next best vote

WGS whole genome sequencing, *N/D* not done

[¶] denotes false-positive variant

In contrast, focal deletions, such as *CDKN2A*, were frequently missed by segmentation. Beyond diagnostic needs, we could reconstruct the double minute nature of an *EGFR* amplification (case 3427T), identify the exact genomic breakpoint using algorithmic structural variant discovery [36], and confirm the latter by Sanger sequencing (Fig. S2).

Methylation profiling

A major advantage of nanopore sequencing is the ability to detect base modifications, especially 5-methylation of cytosines, in native DNA without need for bisulfite conversion. Epigenomic changes are functionally important in cancer, but also aid in delineating cancer entities. For example, IDH mutations cause a global hypermethylation of CpG islands [25], a phenotype of utmost prognostic importance in neuro-oncology. We thus aimed to detect the G-CIMP phenotype from nanopore reads.

First, we compared methylation events in CpG sites identified by nanopore sequencing to matched methylome microarrays. Good correlation was observed between single read methylation status of a given CpG site and its corresponding beta value in microarray data (Fig. 2a). Next, we applied random forest (RF) classification to predict IDH mutation.

RF classification is a commonly used machine-learning algorithm based on randomly generated (weak) decision trees [3]. Majority votes then integrate decisions from the entire forest to provide robust classification. The challenge with low-pass WGS data is that it is not known beforehand which CpG sites will be sequenced and the classifier can be built upon. Therefore, we generated random forests ad hoc. With increasing numbers of probed CpG sites, we expect the classifier's error rate to decrease. To test the feasibility of this approach, we simulated multiple random forests for a given number of CpG sites using the low-grade glioma cohort [5] from The Cancer Genome Atlas (TCGA) and determined misclassification rate for this "random taiga" (Fig. 2b). The simulations show that the mean class error rate to predict IDH and 1p/19q status does not improve for more than approximately 500 CpG sites. This amount of data is reliably sampled within 6 h of nanopore sequencing. Thus, information with respect to a cancer's entity is redundantly encoded in the methylome and this fact can be exploited for classification from sparse, randomly sampled CpG sites.

Using the same training set, we then predicted IDH status in our samples from nanopore-based methylation calls. Due to the low read depth (usually $N = 1$), methylation calls from nanopore WGS were binary. To enable classification using microarray-based training data, beta values were dichotomized as described in previous applications of RF in methylation data [5, 7]. All samples were correctly classified (Fig. 2c).

Supervised pan-cancer classification

Next, following the idea of a machine-learning-based molecular classification of tumors to fully recognize molecular entities and rule out interobserver variability [32], we sought to investigate whether nanopore CN and methylation profiles can be used to classify tumor samples on a pan-cancer level. As a training set for all analyses, we used public microarray-based methylation data from primary brain tumors (adult and pediatric glioblastomas, lower grade gliomas, and medulloblastomas) and tumors that frequently metastasize to the brain (melanoma, breast, lung, bladder, prostate, colon, and clear cell renal carcinoma) [1,

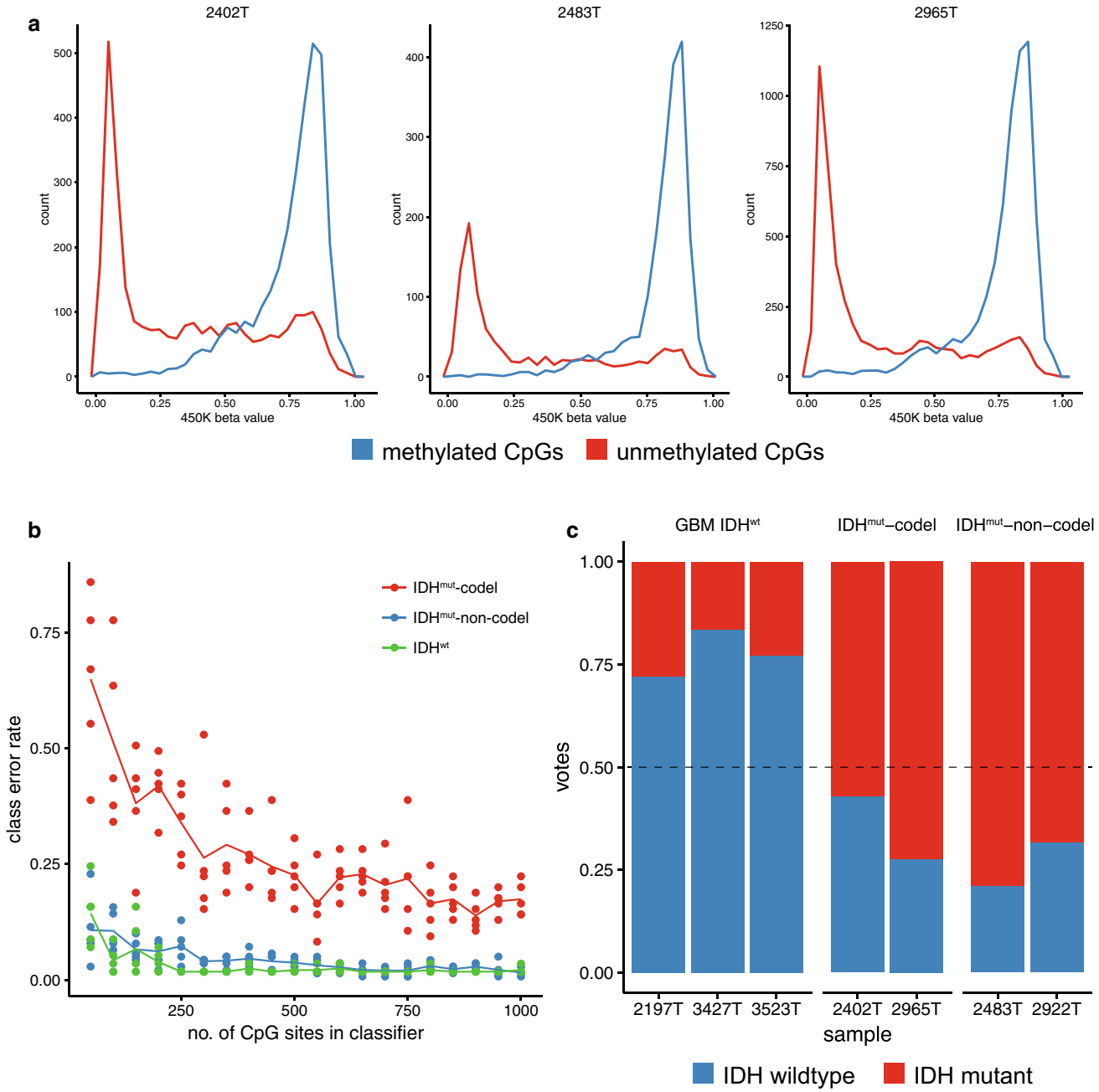
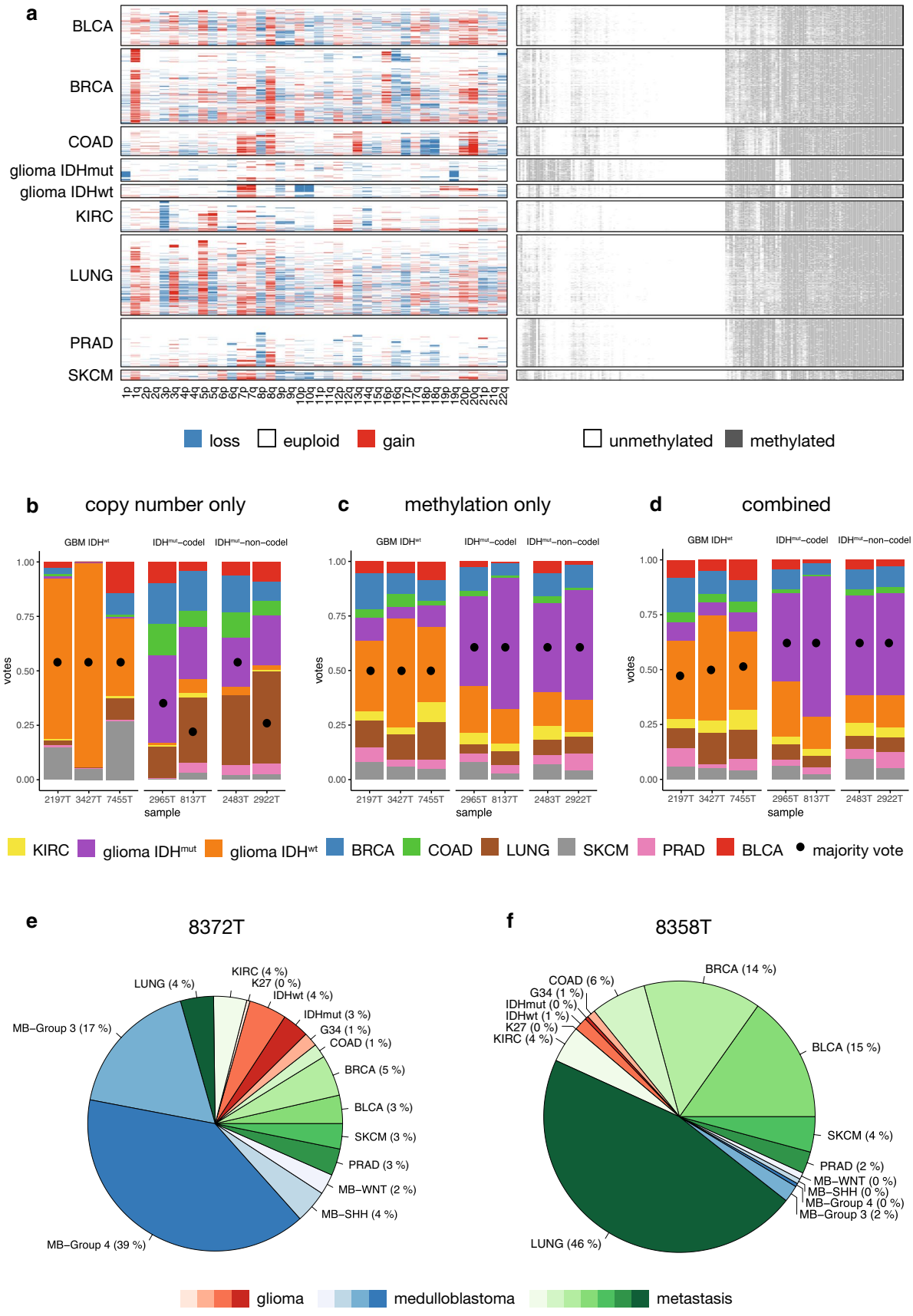


Fig. 2 Methylome profiling by nanopore sequencing of native tumor DNA. **a** Comparison of methylation calls from nanopore sequencing with matched Illumina 450K microarray-based data. Beta value distributions for CpG sites that were identified as unmethylated (*red*) or methylated (*blue*), respectively, by nanopore WGS are shown. **b** "Random taiga" simulation of classification error as a function of the number of randomly sampled CpG sites. Each *dot* represents the class-specific error rate of an ad hoc generated random forest using a random subset of *N* CpG sites (indicated on *X* axis) from the TCGA lower grade glioma Illumina 450K cohort as training set. *Lines* indicate the mean of five independent simulations. **c** Methylation profiles from nanopore sequencing discriminate IDH-mutant and wild-type tumors. *Bar plots* indicate vote distribution from ad hoc random forest classification. The TCGA low-grade glioma cohort was used as a training set. Illumina 450K-based beta values were dichotomized using >0.6 as threshold

2, 4, 5, 12, 23, 24, 37–40]. Where CN data were available, too, SNP array-based CN profiles were aggregated to chromosome arm level and added to the training set (Fig. 3a). The resulting classifiers for any set of CpG sites in our cohort usually yielded an overall out-of-bag classification error rate ≪5%.

We first subjected seven glioma samples with CN and methylation profiles generated by nanopore sequencing to


a
BLCA
BRCA
COAD
glioma IDHmut
glioma IDHwt
KIRC
LUNG
PRAD
SKCM
loss
euploid
gain
unmethylated
methylated
b copy number only
c methylation only
d combined
GBM IDHwt
IDHmut–codel
IDHmut–non–codel
votes
sample
2197T 3427T 7455T 2965T 8137T 2483T 2922T
KIRC
glioma IDHmut
glioma IDHwt
BRCA
COAD
LUNG
SKCM
PRAD
BLCA
majority vote
e 8372T
LUNG (4 %)
KIRC (4 %)
K27 (0 %)
IDHwt (4 %)
IDHmut (3 %)
G34 (1 %)
COAD (1 %)
BRCA (5 %)
BLCA (3 %)
SKCM (3 %)
PRAD (3 %)
MB–WNT (2 %)
MB–SHH (4 %)
MB–Group 3 (17 %)
MB–Group 4 (39 %)
f 8358T
BRCA (14 %)
COAD (6 %)
G34 (1 %)
IDHmut (0 %)
IDHwt (1 %)
K27 (0 %)
KIRC (4 %)
BLCA (15 %)
SKCM (4 %)
PRAD (2 %)
MB–WNT (0 %)
MB–SHH (0 %)
MB–Group 4 (0 %)
MB–Group 3 (2 %)
LUNG (46 %)
glioma
medulloblastoma
metastasis

◂**Fig. 3** Pan-cancer classification using copy number and methylation profiles. **a** Training set composed of TCGA samples from nine cancer entities using arm-level averaged copy number (CN) information (CN loss *blue*, CN gain *red*) and dichotomized methylation data. For illustration purposes, only 200 random CpG sites were sampled, clustered, and plotted. **b**–**d** Classification of samples subjected to WGS using R9.4 flow cells using ad hoc random forests (500 trees per sample). *Bar plots* show vote distributions based on copy number only (**b**), methylation (**c**), or both modalities (**d**). **e**, **f** Methylation-based pan-cancer classification of medulloblastoma (**e**) and a brain metastasis of a lung adenocarcinoma (**f**). *BRCA* breast cancer, *BLCA* bladder urothelial carcinoma, *COAD* colon adenocarcinoma, *KIRC* kidney renal cell carcinoma, *LUNG* lung squamous cell and adenocarcinoma, *SKCM* skin cutaneous melanoma, *PRAD* prostate adenocarcinoma, *MB* medulloblastoma, *K27* diffuse midline glioma H3 K27M mutant, *G34* pediatric glioblastoma, H3 G34R mutant

ad hoc RF classification. When we compared classification using CN alone (Fig. 3b), methylation only (Fig. 3c) or both modalities together (Fig. 3d), using the joint approach improved overall accuracy: all (7/7) samples were correctly classified.

Then, we subjected two medulloblastoma (MB) cases to classification (here, only methylation training data were available). Both samples were identified as MB and also the genetic subtype according to the WHO classification was predicted correctly as WNT-activated (case MB683) or non-SSH-activated/non WNT-activated (i.e., group 4, case 8372T) (Fig. 3e). Next, we attempted classification of brain metastasis and could predict the pulmonary origin in one case (Fig. 3f). We also selected a metastasis of a breast adenocarcinoma in the posterior fossa for study which immunohistochemically showed expression of GFAP and S100, so it was misleading for the diagnosis of carcinoma. Pan-cancer classification based on nanopore WGS correctly identified this sample as breast cancer (Table 1, Fig. S1).

Several cases were not classifiable (requiring a > 5 percentage points' difference of the majority vote to the next best vote) or misclassified (Table 1). These cases had often lower DNA quality with respect to fragment size (Table S1). One GBM sample that was not classifiable had low tumor purity when estimated from matched transcriptomic profiles using the ESTIMATE algorithm [41] (Fig. S3a). This also resulted in false-negative calling of copy number CN alterations using fixed thresholds, even though they were present at visual inspection (Fig. S3b).

Amplicon sequencing

Finally, we explored deep amplicon nanopore sequencing for identification of single nucleotide variants. We designed an amplicon panel covering hotspot exons in *IDH1*, *IDH2*, and *H3F3A*, all coding exons of *TP53* and, additionally, the *TERT* promoter (pTERT) region. Due to the long reads delivered by nanopore sequencing, this could be achieved with only nine PCR reactions (Table S2). Mutations in these genes (with exception of pTERT) inform molecular diagnosis of glioma and medulloblastoma, and are demanded for diagnosis in the 2016 WHO classification of CNS tumors [20]. Sufficient read depth is a critical parameter for variant calling with defined sensitivity and specificity. We thus implemented a real-time analysis pipeline that allowed monitoring of read depth and to stop sequencing when sufficient information to make a diagnosis has been collected (Fig. 4a). In samples run as single samples with real-time monitoring, a sequencing depth of 1000X in all target regions could repeatedly be achieved within 2–20 min of sequencing. Mean overall coverage >1000X could be achieved in single runs, but was lower in runs using barcoding PCR for multiplexing (Fig. 4b).

In all samples, coding mutations were reliably detected as compared to routine diagnostics based on Sanger sequencing, immunohistochemistry or a next-generation sequencing (NGS) panel (Fig. 4c). Nanopore sequencing reads have historically shown high error rates, especially in homopolymer contexts. We, therefore, compared nanopore consensus sequences to matched short-read whole exome data in five cases. Overall concordance was 97.8–98.6% before functional filtering. Even though at low number (<5 per sample) after filtering for coding mutations, false-positive variants were present. Most of these mutations occurred in multiple samples, indicating a context-specific error (Table S3). Improved base calling algorithms are thus needed to reduce the time to manually review mutations for false positives.

Technical aspects

Nanopore sequencing is highly scalable due to low capital cost of the device (use of multiple sequencers) and reuse of flow cells. To exclude carry-over and cross-contamination in sequential sequencing runs and for scalability, we evaluated barcoding and multiplexing for both WGS and amplicon workflows (Table S1, Fig. 4b). For WGS, up to four samples were combined without major protocol changes and permitting convenient overnight runs (e.g., one sample for 6 h and two samples for 12 h). Barcoding of amplicon libraries and multiplexing 12 samples greatly reduces per-assay price at the cost of additional PCR and quality control steps. Finally, we explored use of DNA derived from formalin-fixed paraffin-embedded tissue (FFPE). PCR amplicons were generated from two FFPE samples with identical input amount and protocol. As expected from the usually highly fragmented DNA, PCR yields were lower, especially for large amplicons (>1 kbp). This could only partly be compensated by extending sequencing time. For nanopore WGS, transposase-based library preparation is not compatible with fragment size distribution of FFPE-derived

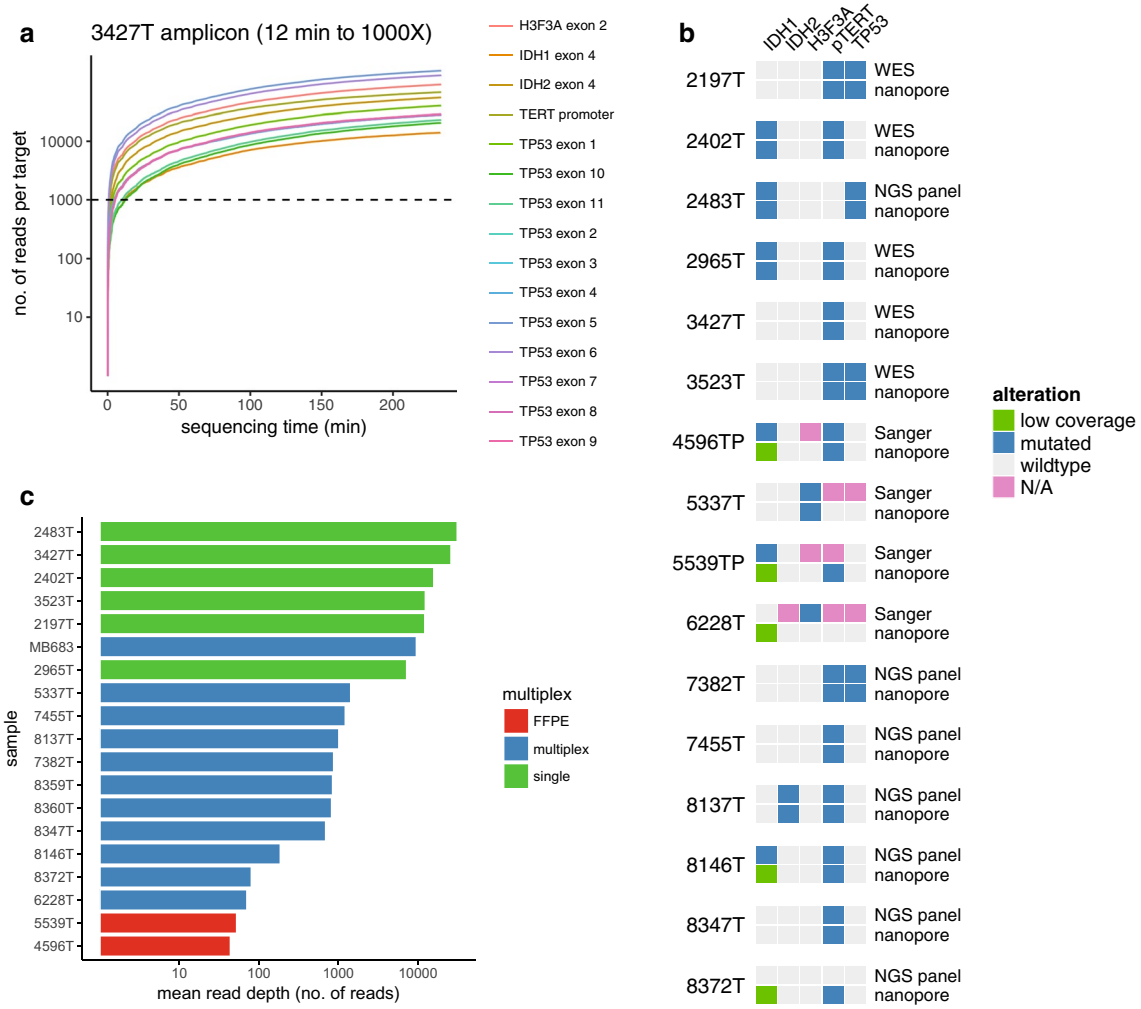


Fig. 4 Real-time amplicon sequencing of single nucleotide variants. **a** Representative coverage plot of target regions in *IDH1*, *IDH2*, *H3F3A*, *TP53*, and *TERT* promoter region over time. The time needed to achieve 1000X depth in all amplicons is indicated. Note log scale on *Y* axis. **b** Mean read depth over all amplicons in samples processed individually or as barcoded multiplex libraries. Of note, FFPE samples were sequenced as part of a multiplex library. **c** Comparison of selected variant calls from nanopore sequencing (filtered for coding or hotspot mutations with minimum allele frequency >0.2) with reference calls from Sanger or Illumina sequencing

DNA samples. We thus performed a different ligation protocol to test WGS in one FFPE sample. While read yield was acceptable (Table S1), the resulting copy number profile was noisy and hard to interpret (Fig. S1). In summary, nanopore sequencing is compatible with FFPE samples, but clearly not recommended due to inferior performance.

Discussion

Histomolecular classification promises to significantly improve diagnosis, prognosis, and treatment decision making of cancer patients by aiding in clearly delineating distinct (molecular) entities and identifying targetable genomic alterations for personalized treatment. It is, therefore, crucial to ensure widespread implementation of appropriate technology in clinical routine for patient benefit. We explored the potential of nanopore sequencing to comprehensively characterize genetic alterations.

CN alterations could be detected in brain tumor samples using ultra low-pass WGS. While overall resolution is lower than current SNP arrays or NGS approaches, arm-level alterations and high-level focal alterations are reliably recapitulated. Most importantly, detection of 1p/19q-codeletion fulfills diagnostic needs for the current WHO 2016 classification of CNS tumors. While WGS using rapid, transposase-based library preparation works very well with high molecular weight DNA, some of the clinical routine fresh-frozen tumor DNA samples were highly fragmented and yielded insufficient results. Quality of input DNA thus seems to be pivotal. For use of FFPE material, changes to the protocol and further optimization are needed.

Methylation data can directly be obtained from the same WGS data set which makes time-consuming bisulfite conversion and specialized methylation assays (sequencing or hybridization-based) expendable. Very recently, it has been shown in the context of meningioma that classification of tumors using methylome data alone is sufficient or superior to make a correct diagnosis [32]. With low genome coverage, we obtained sparse random sampling of CpG sites. We show that this information is sufficient to subtype gliomas into IDH-mutant vs. wild-type samples and that cancer entities from different tissue origins can be distinguished in a few hours. This may aid in the differential diagnosis of primary brain tumors vs. brain metastases and greatly facilitate staging and the search for unknown primary tumors [22]. However, as diagnosis is inferred from relatively sparse data, it precludes inter-patient comparison and reuse of data with currently obtainable coverage in the (relatively short) time frame of 6 h of sequencing.

Finally, we used PCR-based amplicon generation followed by nanopore sequencing to identify point mutations. Using a small, but diagnostically relevant gene panel (covering target regions with a total of 12 kb), high read depth could be routinely obtained in less than 30 min of sequencing when using real-time depth monitoring. However, context-specific base calling errors introduce platform-specific errors and false variant calls that need to be carefully reviewed.

Comparison to existing technologies

Targeted next-generation sequencing panels tailored to detect mutations in brain tumors or, more generally, cancer-related genes have been employed routinely with a turnaround time of several days [8, 31]. Methylation-based classification of brain tumors by microarray allows differentiation of a wealth of different entities within 2 weeks [12, 32]. Intraoperative subtyping of gliomas is possible using allele specific PCR for key alterations (IDH1, pTERT) but remains restricted to hotspot point mutations [34]. Similarly, CN changes and mutations have been detected in cell-free DNA from CSF to allow less invasive diagnostics [10, 26]. A major drawback of all approaches is the high investment cost, need for laboratory space or expertise.

For nanopore sequencing, besides the portable sequencing device and a laptop computer, only a spectrometer for DNA quantification and a thermocycler for library preparation and amplicon generation by PCR are needed. This allows implementation of a complete molecular pathology laboratory even in resource-restricted settings or mobile environments. Per sample cost is ~$200 for WGS and ~$120 for amplicon sequencing without multiplexing. However, being a technology still under development, frequent updates in chemistry and software currently challenge routine use and need to be addressed to allow standardized diagnostics across laboratories. In addition, hybridization microarrays and targeted short-read sequencing both work relatively well with fragmented DNA from FFPE samples, while this currently poses a technical challenge for nanopore sequencing.

Our study has several limitations. First, as this is a proof-of-principle study, sample number is small and precludes accurate quantification of sensitivity or specificity to detect structural alterations and point mutations. Second, a prospective and multi-centric evaluation of the approach presented here is needed to rule out sample selection bias and demonstrate robustness across laboratories. Third, we reused flow cells to reduce per-assay cost, but washing also decreased the number of active pores and thus performance in subsequent runs.

In conclusion, same-day diagnosis of CN alterations, epigenetic modifications, and single nucleotide variants using nanopore sequencing is feasible with minimal capital cost and without need for sophisticated laboratory equipment. For CNS tumors, molecular features demanded for diagnosis by current guidelines can be obtained, which, together with histological data and grading, enable accelerated integrated diagnosis and improve patient care.

Acknowledgements The authors would like to acknowledge Jared T. Simpson for providing methylation training models, Inès Detrait and Amithys Rahimian-Aghda for biobanking and sample management, Ludovic Prevost for excellent IT systems administration, and Mark van Roosmalen for assistance with running NanoSV. We are indebted to the Nanopore Human Genome Sequencing consortium for early release of data at https://github.com/nanopore-wgs-consortium/NA12878. The results published here are in part based upon data generated by The Cancer Genome Atlas (TCGA) project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/.

Compliance with ethical standards

Funding This work has been supported by Deutsche Forschungsgemeinschaft (EU 162/1-1 to PE), the program "Investissements d'avenir" (ANR-10-IAIHU-06 to AI), Institut Universitaire de Cancérologie (to AI), Ligue Nationale Contre le Cancer (to AI), Institut Carnot (to KL), and Fondation ARC pour la recherche sur le cancer (n°PJA 20151203562 to FB).

Conflict of interest The authors declare no conflicts of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD et al (2015) The molecular taxonomy of primary prostate cancer. Cell 163:1011–1025. doi:10.1016/j.cell.2015.10.025
2. Akbani R, Akdemir KC, Aksoy BA, Albert M, Ally A, Amin SB et al (2015) Genomic classification of cutaneous melanoma. Cell 161:1681–1696. doi:10.1016/j.cell.2015.05.044
3. Breiman L (2001) Random forests. Mach Learn 45:5–32. doi:10.1023/A:1010933404324
4. Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR et al (2013) The somatic genomic landscape of glioblastoma. Cell 155:462–477. doi:10.1016/j.cell.2013.09.034
5. Brat DJ, Verhaak RGW, Aldape KD, Yung WKA, Salama SR, Cancer Genome Atlas Research Network et al (2015) Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med 372:2481–2498. doi:10.1056/NEJMoa1402121
6. Cao MD, Ganesamoorthy D, Cooper MA, Coin LJM (2016) Realtime analysis and visualization of MinION sequencing data with npReader. Bioinformatics 32:764–766. doi:10.1093/bioinformatics/btv658
7. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA et al (2016) Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell 164:550–563. doi:10.1016/j.cell.2015.12.028
8. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A et al (2015) Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn 17:251–264. doi:10.1016/j.jmoldx.2014.12.006
9. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L et al (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 6:80–92. doi:10.4161/fly.19695
10. De Mattos-Arruda L, Mayor R, Ng CKY, Weigelt B, Martínez-Ricarte F, Torrejon D et al (2015) Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. Nat Commun 6:8839. doi:10.1038/ncomms9839
11. Goldman M, Craft B, Swatloski T, Ellrott K, Cline M, Diekhans M et al (2013) The UCSC cancer genomics browser: update 2013. Nucl Acids Res 41:D949–D954. doi:10.1093/nar/gks1008
12. Hovestadt V, Remke M, Kool M, Pietsch T, Northcott PA, Fischer R et al (2013) Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. Acta Neuropathol 125:913–916. doi:10.1007/s00401-013-1126-5
13. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS et al (2015) Orchestrating high-throughput genomic analysis with bioconductor. Nat Methods 12:115–121. doi:10.1038/nmeth.3252
14. Kamoun A, Idbaih A, Dehais C, Elarouci N, Carpentier C, Letouzé E et al (2016) Integrated multi-omics analysis of oligodendroglial tumours identifies three subgroups of 1p/19q co-deleted gliomas. Nat Commun 7:11263. doi:10.1038/ncomms11263
15. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L et al (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22:568–576. doi:10.1101/gr.129684.111
16. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291. doi:10.1038/nature19057
17. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324
18. Loman NJ, Quinlan AR (2014) Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics 30:3399–3401. doi:10.1093/bioinformatics/btu555
19. Loose M, Malla S, Stout M (2016) Real-time selective sequencing using nanopore technology. Nat Methods. doi:10.1038/nmeth.3930
20. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK et al (2016) The 2016 World Health Organization classification of tumors of the central nervous system: a summary. Acta Neuropathol 131:803–820. doi:10.1007/s00401-016-1545-1
21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. doi:10.1101/gr.107524.110
22. Moran S, Martínez-Cardús A, Sayols S, Musulén E, Balañá C, Estival-Gonzalez A et al (2016) Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. Lancet Oncol 17:1386–1395. doi:10.1016/S1470-2045(16)30297-2
23. Network TCGA (2012) Comprehensive molecular characterization of human colon and rectal cancer. Nature 487:330–337. doi:10.1038/nature11252
24. Network TCGAR (2012) Comprehensive genomic characterization of squamous cell lung cancers. Nature 489:519–525. doi:10.1038/nature11404
25. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP et al (2010) Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell 17:510–522. doi:10.1016/j.ccr.2010.03.017
26. Pentsova EI, Shah RH, Tang J, Boire A, You D, Briggs S et al (2016) Evaluating cancer of the central nervous system through next-generation sequencing of cerebrospinal fluid. J Clin Oncol. doi:10.1200/JCO.2016.66.6487
27. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L et al (2016) Real-time, portable genome sequencing for Ebola surveillance. Nature 530:228–232. doi:10.1038/nature16996
28. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. doi:10.1093/bioinformatics/btq033
29. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M et al (2017) Mapping DNA methylation with high-throughput nanopore sequencing. Nat Methods. doi:10.1038/nmeth.4189
30. Rosenberg S, Verreault M, Schmitt C, Guegan J, Guehennec J, Levasseur C et al (2017) Multi-omics analysis of primary glioblastoma cell lines shows recapitulation of pivotal molecular features of parental tumors. Neuro-Oncol 19:219–228. doi:10.1093/neuonc/now160
31. Sahm F, Schrimpf D, Jones DTW, Meyer J, Kratz A, Reuss D et al (2016) Next-generation sequencing in routine brain tumor diagnostics enables an integrated diagnosis and identifies actionable targets. Acta Neuropathol 131:903–910. doi:10.1007/s00401-015-1519-8
32. Sahm F, Schrimpf D, Stichel D, Jones DTW, Hielscher T, Schefzyk S et al (2017) DNA methylation-based classification and grading system for meningioma: a multicentre, retrospective analysis. Lancet Oncol. doi:10.1016/S1470-2045(17)30155-9

33. Scheinin I, Sie D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF et al (2014) DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. Genome Res 24:2022–2032. doi:10.1101/gr.175141.114
34. Shankar GM, Francis JM, Rinne ML et al (2015) Rapid intraoperative molecular characterization of glioma. JAMA Oncol. doi:10.1001/jamaoncol.2015.0917
35. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W (2017) Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. doi:10.1038/nmeth.4184
36. Stancu MC, Roosmalen MJ van, Renkens I, Nieboer M, Middelkamp S, Ligt J de, et al. (2017) Mapping and phasing of structural variation in patient genomes using nanopore sequencing. bioRxiv 129379. doi:10.1101/129379
37. Sturm D, Witt H, Hovestadt V, Khuong-Quang D-A, Jones DTW, Konermann C et al (2012) Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. Cancer Cell 22:425–437. doi:10.1016/j.ccr.2012.08.024
38. The Cancer Genome Atlas Research Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499:43–49. doi:10.1038/nature12222
39. The Cancer Genome Atlas Research Network (2014) Comprehensive molecular profiling of lung adenocarcinoma. Nature 511:543–550. doi:10.1038/nature13385
40. The Cancer Genome Atlas Research Network (2014) Comprehensive molecular characterization of urothelial bladder carcinoma. Nature 507:315–322. doi:10.1038/nature12965
41. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W et al (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 4:2612. doi:10.1038/ncomms3612

Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors

Beatrice S Melin[1,41], Jill S Barnholtz-Sloan[2,41], Margaret R Wrensch[3,4,41], Christoffer Johansen[5,41], Dora Il'yasova[6–8,41], Ben Kinnersley[9,41], Quinn T Ostrom[2], Karim Labreche[9,10], Yanwen Chen[2], Georgina Armstrong[11], Yanhong Liu[11], Jeanette E Eckel-Passow[12], Paul A Decker[12], Marianne Labussière[10], Ahmed Idbaih[10,13], Khe Hoang-Xuan[10,13], Anna-Luisa Di Stefano[10,13], Karima Mokhtari[10,13], Jean-Yves Delattre[10,13], Peter Broderick[9], Pilar Galan[14], Konstantinos Gousias[15], Johannes Schramm[15], Minouk J Schoemaker[9], Sarah J Fleming[16], Stefan Herms[16], Stefanie Heilmann[17], Markus M Nöthen[17], Heinz-Erich Wichmann[18–20], Stefan Schreiber[21], Anthony Swerdlow[9,22], Mark Lathrop[23], Matthias Simon[15], Marc Sanson[10,13], Ulrika Andersson[1], Preetha Rajaraman[24], Stephen Chanock[24], Martha Linet[24], Zhaoming Wang[24], Meredith Yeager[24], GliomaScan Consortium[25], John K Wiencke[3,4], Helen Hansen[3], Lucie McCoy[3], Terri Rice[3], Matthew L Kosel[12], Hugues Sicotte[12], Christopher I Amos[26], Jonine L Bernstein[27], Faith Davis[28], Dan Lachance[29], Ching Lau[30], Ryan T Merrell[31], Joellen Shildkraut[7,8], Francis Ali-Osman[7,32], Siegal Sadetzki[33,34], Michael Scheurer[30], Sanjay Shete[35], Rose K Lai[36,42], Elizabeth B Claus[37,38,42], Sara H Olson[27,42], Robert B Jenkins[39,42], Richard S Houlston[9,40,42] & Melissa L Bondy[11,42]

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

Genome-wide association studies (GWAS) have transformed our understanding of glioma susceptibility, but individual studies have had limited power to identify risk loci. We performed a meta-analysis of existing GWAS and two new GWAS, which totaled 12,496 cases and 18,190 controls. We identified five new loci for glioblastoma (GBM) at 1p31.3 (rs12752552; $P = 2.04 \times 10^{-9}$, odds ratio (OR) = 1.22), 11q14.1 (rs11233250; $P = 9.95 \times 10^{-10}$, OR = 1.24), 16p13.3 (rs2562152; $P = 1.93 \times 10^{-8}$, OR = 1.21), 16q12.1 (rs10852606; $P = 1.29 \times 10^{-11}$, OR = 1.18) and 22q13.1 (rs2235573; $P = 1.76 \times 10^{-10}$, OR = 1.15), as well as eight loci for non-GBM tumors at 1q32.1 (rs4252707; $P = 3.34 \times 10^{-9}$, OR = 1.19), 1q44 (rs12076373; $P = 2.63 \times 10^{-10}$, OR = 1.23), 2q33.3 (rs7572263; $P = 2.18 \times 10^{-10}$, OR = 1.20), 3p14.1 (rs11706832; $P = 7.66 \times 10^{-9}$, OR = 1.15), 10q24.33 (rs11598018; $P = 3.39 \times 10^{-8}$, OR = 1.14), 11q21 (rs7107785; $P = 3.87 \times 10^{-10}$, OR = 1.16), 14q12 (rs10131032; $P = 5.07 \times 10^{-11}$, OR = 1.33) and 16p13.3 (rs3751667; $P = 2.61 \times 10^{-9}$, OR = 1.18). These data substantiate that genetic susceptibility to GBM and non-GBM tumors are highly distinct, which likely reflects different etiology.

Glioma accounts for around 27% of all primary brain tumors and is responsible for approximately 13,000 cancer-related deaths in the United States each year[1,2]. Gliomas can be broadly classified into GBM and lower-grade non-GBM tumors[3]. Gliomas typically have a poor prognosis irrespective of medical care, with the most common form, GBM, having a five-year survival rate of only 5% (ref. 4).

So far, no environmental exposures have been robustly linked to the risk of developing glioma, except for moderate to high doses of ionizing radiation, which accounts for a small proportion of cases[5]. Evidence for an inherited predisposition to glioma is provided by a number of rare inherited cancer syndromes, such as Turcot's and Li–Fraumeni syndromes, as well as neurofibromatosis. Even collectively, however, these account for little of the twofold familial risk of glioma[6]. Our understanding of the heritability of glioma has been transformed by recent GWAS, which have identified single-nucleotide polymorphisms (SNPs) at 13 loci influencing risk[7–14].

Previous individual studies have had limited statistical power for the additional discovery of new glioma risk loci[15]. Therefore, to gain more comprehensive insight into glioma etiology, we performed a meta-analysis of previously published GWAS and two new GWAS, which allowed us to identify 13 new risk loci for glioma.

We analyzed GWAS SNP data that passed quality control for 12,496 cases (6,191 classified as GBM and 5,819 classified as non-GBM tumors) and 18,190 controls from eight studies with individuals of European ancestry, a new GWAS of 4,572 cases and 3,286 controls performed by the Glioma International Case Control Consortium (GICC) (**Supplementary Table 1**), a new GWAS of 1,591 cases and 804 controls from the University of California, San Francisco (UCSF)-Mayo,

A full list of affiliations appears at the end of the paper.

Received 20 May 2016; accepted 1 March 2017; published online 27 March 2017; doi:10.1038/ng.3823

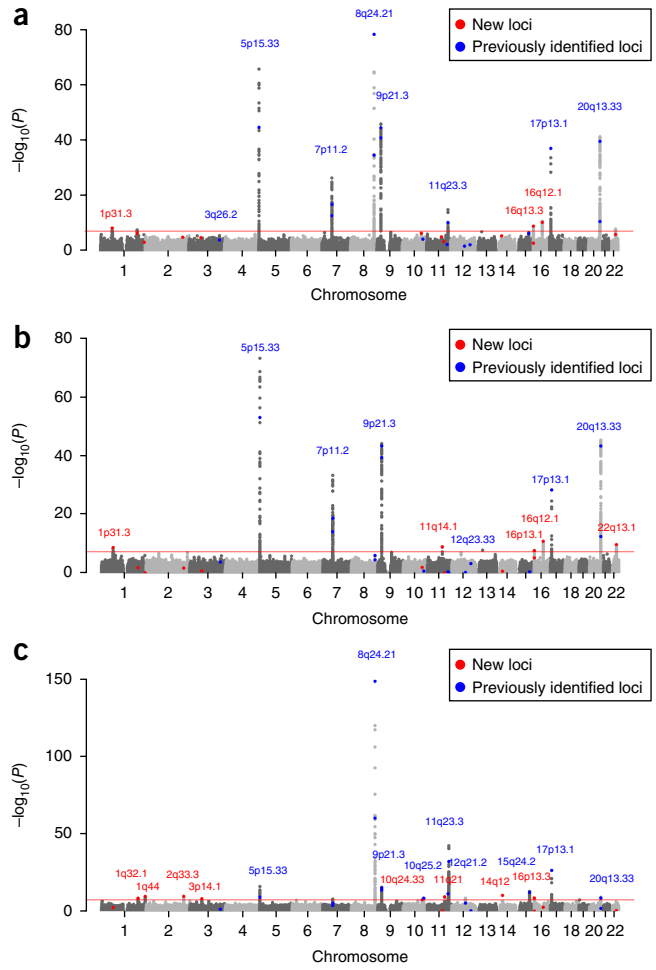


Figure 1 Genome-wide discovery-phase meta-analysis *P*-values ($-\log_{10}P$) plotted against their chromosomal positions. (**a**) All glioma. (**b**) GBM. (**c**) Non-GBM tumors. The red horizontal line corresponds to a significance threshold of $P = 5.0 \times 10^{-8}$. New and known loci are labeled in red and blue, respectively.

and six previously reported GWAS[9,10,13] totaling 6,405 cases and 14,100 controls (**Supplementary Table 2**). To increase genomic resolution, we imputed >10 million SNPs. Quantile–quantile (Q-Q) plots for SNPs with a minor allele frequency (MAF) >1% after imputation did not show evidence of substantive overdispersion ($\lambda = 1.02–1.10$, $\lambda_{90} = 1.02–1.05$; **Supplementary Fig. 1**). We derived joint ORs and 95% confidence intervals (CIs) under a fixed-effects model for each SNP with MAF >1% and associated per-allele principal component (PCA) corrected *P*-values for all glioma, GBM and non-GBM cases versus those for the controls (**Fig. 1**).

In the combined meta-analysis, among previously published glioma risk SNPs, those for all glioma at 17p13.1 (*TP53*), for GBM at 5p15.33 (*TERT*), 7p11.2 (*EGFR*), 9p21.3 (*CDKN2B–AS1*) and 20q13.33 (*RTEL1*), and for non-GBM tumors at 8q24.21 (*CCDC26*), 11q23.2, 11q23.3 (*PHLDB1*) and 15q24.2 (*ETFA*) showed even greater evidence for association (**Supplementary Fig. 2** and **Supplementary Table 3**). SNPs at 10q25.2 and 12q12.1 for non-GBM tumors retained genome-wide significance (i.e., $P < 5.0 \times 10^{-8}$). Associations at the previously reported 3q26.2 (near *TERC*)[11] and 12q23.33 (*POLR3B*)[10] loci for GBM did not retain statistical significance (*P* values for the most associated SNPs are 2.68×10^{-5} and 1.60×10^{-5}, respectively; **Supplementary Table 3**).

In addition to previously reported loci, we identified genome-wide significant associations marking new risk loci (**Table 1**, **Supplementary Fig. 3** and **Supplementary Data 1**) for GBM at 1p31.3 (rs12752552; $P = 2.04 \times 10^{-9}$), 11q14.1 (rs11233250; $P = 9.95 \times 10^{-10}$), 16p13.3 (rs2562152; $P = 1.93 \times 10^{-8}$), 16q12.1 (rs10852606; $P = 1.29 \times 10^{-11}$) and 22q13.1 (rs2235573; $P = 1.76 \times 10^{-10}$) and for non-GBM tumors at 1q32.1 (rs4252707; $P = 3.34 \times 10^{-9}$), 1q44 (rs12076373; $P = 2.63 \times 10^{-10}$), 2q33.3 (rs7572263; $P = 2.18 \times 10^{-10}$), 3p14.1 (rs11706832; $P = 7.66 \times 10^{-9}$), 10q24.33 (rs11598018; $P = 3.39 \times 10^{-8}$), 11q21 (rs7107785; $P = 3.87 \times 10^{-10}$), 14q12 (rs10131032; $P = 5.07 \times 10^{-11}$) and 16p13.3 (rs3751667; $P = 2.61 \times 10^{-9}$). Conditional analysis confirmed the existence of two independent association signals at 7p11.2 (*EGFR*) as previously reported[7] but did not provide evidence for additional signals at any of the other established identified risk loci or at the 13 newly identified loci. Case-only analyses confirmed the specificity of 11q14.1, 16p13.3 and 22q13.1 associations for GBM and of 1q44, 2q33.3, 3p14.1, 11q21 and 14q12 associations for non-GBM tumors (**Fig. 2** and **Supplementary Table 4**). Collectively, our findings provide strong evidence for specific associations for the different glioma subtypes, consistent with their previously described distinctive molecular profiles, presumably resulting from different etiological pathways.

Across the new and known risk loci, we found a significant enrichment of overlap with enhancers in H9-Derived neuronal progenitor cells ($P = 8.2 \times 10^{-5}$; **Supplementary Data 2**). These observations support the assertion that the loci identified in the GWAS influence glioma risk through effects on neural *cis* regulatory networks and that they are strongly involved in transcriptional initiation and enhancement. To gain further insight into the biological basis for associations at the 13 new risk loci, we performed an expression quantitative trait loci (eQTL) analysis using RNA-seq data on ten regions of normal human brain from up to 103 individuals from the Genotype–Tissue Expression (GTEx) project[16] and blood eQTL data on 5,311 individuals from Westra *et al.*[17]. We used summary-level mendelian randomization (SMR)[18] analysis to test for a concordance between signals from GWAS and *cis* eQTL for genes within 1 Mb of the sentinel and correlated SNPs ($r^2 > 0.8$) at each locus (**Supplementary Data 3**) and derived b_{XY} statistics, which estimate the effect of gene expression on glioma risk. Additionally, for each of the risk SNPs at the 13 new loci (as well as the correlated variants), we examined published data[19,20] and made use of the online resources HaploRegv4, RegulomeDB and SeattleSeq for evidence of functional effects (**Supplementary Table 5**).

At 16q12.1, the GBM association signal was significantly associated with *HEATR3* expression in nine of ten regions of the brain ($P_{SMR} = 3.38 \times 10^{-6}$ to 6.55×10^{-10}; $b_{XY} = 0.14–0.24$; **Supplementary Fig. 4** and **Supplementary Data 3**). The risk allele 'C' of rs10852606 that was associated with reduced *HEATR3* expression was consistent with differential expression of *HEATR3* being the functional basis of the 16q12.1 association. The observation that variation at 16q12.1 is associated with risk of testicular[21] (rs8046148; pairwise r^2 and D' with rs10852606 of 0.67 and 1.0, respectively) and esophageal[22] (rs4785204; pairwise r^2 and D' with rs10852606 of 0.16 and 1.0, respectively) cancer suggests that the locus has pleiotropic effects on tumor risk, which are compatible with generic effects as shown by the observation of a *HEATR3* eQTL signal in blood ($P_{SMR} = 5.84 \times 10^{-11}$; $b_{XY} = 0.30$).

Similarly, significant associations between gene expression and glioma risk were observed at the GBM loci 1p31.3 (*JAK1*, brain cortex and cerebellar hemisphere), 16p13.3 (*POLR3K*, whole blood) and 22q13.1 (*CTA-228A9.3*, brain cerebellum; *PICK1*, brain hippocampus) (**Supplementary Fig. 4** and **Supplementary Data 3**). The non-GBM association at 1q32.1 marked by rs4252707 (**Supplementary Fig. 3**)

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

Table 1 Association statistics for the top SNP at each of the newly-reported glioma risk loci

							All glioma		GBM glioma		Non-GBM glioma	
Locus	Subtype	SNP	Position	Alleles	RAF	INFO	*P*	OR (95% CI)	*P*	OR (95% CI)	*P*	OR (95% CI)
1p31.3	GBM	rs12752552	65229299	**T**/C	0.870	0.992	**4.07×10^{-9}**	1.18 (1.11–1.24)	**2.04×10^{-9}**	1.22 (1.15–1.31)	4.78×10^{-3}	1.11 (1.03–1.18)
1q32.1	Non-GBM	rs4252707	204508147	G/**A**	0.220	0.992	2.97×10^{-7}	1.12 (1.07–1.17)	0.015	1.07 (1.01–1.13)	**3.34×10^{-9}**	1.19 (1.12–1.26)
1q44	Non-GBM	rs12076373	243851947	**G**/C	0.837	0.996	4.97×10^{-4}	1.09 (1.04–1.15)	0.846	0.99 (0.94–1.06)	**2.63×10^{-10}**	1.23 (1.16–1.32)
2q33.3	Non-GBM	rs7572263	209051586	**A**/G	0.756	0.997	2.58×10^{-6}	1.11 (1.06–1.15)	0.019	1.06 (1.01–1.12)	**2.18×10^{-10}**	1.20 (1.13–1.26)
3p14.1	Non-GBM	rs11706832	66502981	A/**C**	0.456	0.997	1.06×10^{-5}	1.08 (1.05–1.12)	0.158	1.03 (0.99–1.08)	**7.66×10^{-9}**	1.15 (1.09–1.20)
10q24.33	Non-GBM	rs11598018	105661315	**C**/A	0.462	0.960	3.07×10^{-7}	1.10 (1.06–1.14)	0.0103	1.06 (1.01–1.11)	**3.39×10^{-8}**	1.14 (1.09–1.20)
11q14.1	GBM	rs11233250	82397014	**C**/T	0.868	0.990	5.40×10^{-6}	1.14 (1.08–1.21)	**9.95×10^{-10}**	1.24 (1.16–1.33)	0.592	0.98 (0.91–1.05)
11q21	Non-GBM	rs7107785	95747337	**T**/C	0.479	0.997	2.96×10^{-4}	1.07 (1.03–1.11)	0.844	1.00 (0.95–1.04)	**3.87×10^{-10}**	1.16 (1.11–1.21)
14q12	Non-GBM	rs10131032	33250081	**G**/A	0.916	0.991	2.33×10^{-6}	1.17 (1.09–1.24)	0.247	1.05 (0.97–1.13)	**5.07×10^{-11}**	1.33 (1.22–1.44)
16p13.3	GBM	rs2562152	123896	A/**T**	0.850	0.937	1.18×10^{-3}	1.09 (1.04–1.15)	**1.93×10^{-8}**	1.21 (1.13–1.29)	0.948	1.00 (0.93–1.07)
16p13.3	Non-GBM	rs3751667	1004554	C/**T**	0.208	0.985	**8.75×10^{-10}**	1.14 (1.09–1.19)	5.95×10^{-6}	1.13 (1.07–1.19)	**2.61×10^{-9}**	1.18 (1.12–1.25)
16q12.1	GBM	rs10852606	50128872	T/**C**	0.713	0.990	**3.66×10^{-11}**	1.14 (1.10–1.19)	**1.29×10^{-11}**	1.18 (1.13–1.24)	2.42×10^{-3}	1.08 (1.03–1.14)
22q13.1	GBM	rs2235573	38477930	**G**/A	0.507	0.995	8.64×10^{-7}	1.09 (1.06–1.13)	**1.76×10^{-10}**	1.15 (1.10–1.20)	0.325	1.02 (0.97–1.07)

Associations at $P < 5 \times 10^{-8}$ are highlighted in bold. Odds ratios (ORs) were derived with respect to the risk allele underlined and highlighted in bold. Risk allele frequency (RAF) is according to European samples from the 1000 Genomes project. The INFO column indicates the average imputation info score across all studies, with a score of 1 indicating that the SNP is directly genotyped in all studies. CI, confidence interval.

Figure 2 Relative impact of SNP associations at known and newly identified risk loci for GBM and non-GBM tumors. Odds ratios (ORs) derived with respect to the risk allele. Asterisks denote SNPs showing a significant difference between GBM and non-GBM tumors from the case-only analysis as detailed in **Supplementary Table 4**.

maps to intron 8 of the gene encoding MDM4, a p53-binding protein. The SNP rs4252707 is in strong linkage disequilibrium (LD) with rs12031912 and rs12028476 ($r^2 = 0.92$), both of which map to the *MDM4* promoter. Although no significant eQTL was shown in any brain tissue, an association with *MDM4* was seen in blood ($P_{SMR} = 4.74 \times 10^{-6}$; $b_{XY} = 0.31$; **Supplementary Fig. 4** and **Supplementary Data 3**). Overexpression of *MDM4* is a feature in glioma tumors containing wild-type *TP53* and no amplification of the *MDM2* gene, consistent with *MDM4* amplification being a mechanism by which the p53-dependent growth control is inactivated[23].

The 1q44 association with non-GBM that is marked by rs12076373 maps to intron 8 of *AKT3*, whose encoded product is one of the major downstream effectors of phosphatidylinositol 3-kinase (PI3K) and is highly expressed during active neurogenesis, with haploinsufficiency causing postnatal microcephaly and agenesis of the corpus callosum[24]. Notably, *AKT3* is hyper-expressed in glioma, thus having a role in tumor viability by activating DNA repair[25]. Although rs12076373 does not map to a regulatory element, the correlated SNPs rs12124113 ($r^2 = 0.94$) and rs59953491 ($r^2 = 0.90$) locate within an enhancer element in brain cells and tissues, including H9-derived neuronal progenitor cultured cells, cortex-derived primary cultured neurospheres and NH-A astrocytes.

The 3p14.1 association with non-GBM that is marked by rs11706832 localizes to intron 2 of *LRIG1*. Although we did not identify an eQTL in this gene, *LRIG1* is highly expressed in the brain and is a pan-negative regulator of the epidermal growth factor receptor (EGFR) signaling pathway, which inhibits hypoxia-induced vasculogenic mimicry via EGFR–PI3K–AKT pathway suppression and epithelial-to-mesenchymal transition[26]. Reduced *LRIG1* expression is linked to tumor aggressiveness, temozolomide resistance and radioresistance[27,28]. We have previously shown an association for glioma at *EGFR* (7p11.2)[7], which is well established to be pivotal in both the initiation of primary GBM and the progression of lower-grade glioma to grade IV. Although speculative, our new findings now suggest a more extensive pathway involving variation at *LRIG1* and *AKT3*.

Of particular interest is rs7572263, which maps to 2q33.3, localizes within intron 3 of *C2orf80* and is 50 kb telomeric to *IDH1*. Mutation

of *IDH1* is a driver for gliomagenesis[29,30] and is responsible for the CpG island methylator (G-CIMP) phenotype[31,32]. Mutations in *IDH1* predominate in non-GBM glioma[33,34]; therefore, the association at 2q33.3 is plausible as a basis for susceptibility to non-GBM glioma. In the absence of convincing eQTL or other functional support, this does not preclude *C2orf80* or another gene mapping to the region of LD as being the functional basis for the 2q33.3 association.

The maintenance of telomeres is central to cell immortalization, and it has a central role in gliomagenesis[35]. We have previously shown that the risk of GBM is strongly linked to genetic variation in the telomere-related genes *TERT* (5p15.33) and *RTEL1* (20q13.33), and possibly also *TERC* (3q26.2)[8,9,11]. The 10q24.33 association with non-GBM that is marked by rs11598018 lies intronic to *OBFC1*, which functions in a telomere-associated complex that protects telomeres independently of POT1 (ref. 36). The CST complex, whose components are encoded by *OBFC1*, *CTC1*, and *TEN1*, competes with shelterin for telomeric-DNA-inhibiting telomerase-based telomere extension[37]. The significant association between the risk of non-GBM tumors and *OBFC1* variation is particularly of note in light of our recent exome-sequencing report demonstrating that rare germline loss-of-function mutations in genes that encode components of the shelterin complex are a cause of familial oligodendroglioma[38]. The glioma risk alleles at *TERT*, *TERC* and *OBFC1* are associated with increased leukocyte telomere length, thereby supporting a relationship between genotype and biology (**Supplementary Table 6**)[35,39,40]. However, the *RTEL1* locus is not consistent with such a postulate, and recent data that have not shown a relationship between mutations in the *TERT* promoter and telomere length in glioma[41] raise the possibility of a role for extratelomeric effects.

The deregulation of pathways involved in telomere length and EGFR signaling are thus consistent with glioma risk being governed by pathways that are important in the longevity of glial cells, and they substantiate early observations that genetic susceptibility to GBM and non-GBM tumors is highly distinct, presumably reflecting different etiologies between GBM and non-GBM tumors (**Fig. 2**).

The other associations we identified mark genes with varying degrees of plausibility for having a role in glioma oncogenesis. The GBM association at 16p13.33 marked by rs2562152 localizes 3 kb telomeric to *MPG*, which encodes a *N*-methylpurine DNA glycosylase whose expression is linked to temozolomide resistance in glioma[42]. Although attractive as a candidate, the only genes for which there was found to be a significant association between expression and glioma risk were *POLR3K* and *C16ORF33* in blood (**Supplementary Fig. 4** and **Supplementary Data 3**). At 1p31.3, only *JAK1* provided convincing evidence for a significant eQTL with glioma risk SNPs in brain tissue. The strongest association was shown in the cortex (P_{SMR} = 1.61×10^{-6}; b_{XY} = 0.22; **Supplementary Fig. 4** and **Supplementary Data 3**), with the risk allele 'T' of rs12752552 showing increased *JAK1* expression. The *cis*-eQTL signal for *JAK1* in the cortex maps from 65.3 Mb to 65.35 Mb and shows a consistent direction of effect with the glioma-associated SNPs. JAK1–STAT6 signaling is increasingly being recognized to be relevant in glioma progression[43]. Hence, although *JAK1* remains an attractive candidate mechanistic basis for the glioma association at 1p31.3, we cannot exclude the possibility that the cluster of SNPs between 65.3 Mb and 65.35 Mb contains the true causal variant. In the absence of functional data, potential target genes for associations at 11q14.1 (GBM), 16p13.3 (non-GBM), 11q21 (non-GBM) and 14q12 (non-GBM) remain to be elucidated.

In conclusion, we have performed the largest glioma GWAS to date and have identified 13 new glioma risk loci, thereby providing further evidence for a polygenic basis of genetic susceptibility to glioma. Histological classification of glioma is, in part, being superseded by molecular profiling[34,44]; hence, it is important to understand the biology behind these risk variants in the context of molecularly defined glioma subtypes. Currently identified risk SNPs for glioma account for, at best, ~27% and ~37% of the familial risk of GBM and non-GBM tumors, respectively (**Supplementary Table 7**). Therefore, further GWAS-based analyses in concert with functional analyses should lead to additional insights into the biology and etiological basis of the different glioma histologies. Notably, such information can inform gene discovery initiatives and thus have a measurable effect on the successful development of new therapeutic agents.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to all of the patients and individuals for their participation, and we would also like to thank the clinicians and other hospital staff members, cancer registries and the study staff members in the respective centers who contributed to the blood sample and data collection.

The GICC was supported by grants from the US National Institutes of Health (NIH) (R01CA139020 (M.L.B. and B.S.M.), R01CA52689 (M.R.W.), R01CA52689 (M.L.B.) and P30CA125123 (M. Scheurer). Additional support was provided by the McNair Medical Institute (M. Scheurer) and the Population Sciences Biorepository at Baylor College of Medicine (M. Scheurer).

In Sweden, work was additionally supported by Acta Oncologica through the Royal Swedish Academy of Science (B.S.M.'s salary) and by the Swedish Research Council (B.S.M.) and the Swedish Cancer Foundation (B.S.M.). We are grateful to the National Clinical Brain Tumor Group and to all of the clinicians and research nurses throughout Sweden who identified all of the cases.

In the UK, funding was provided by Cancer Research UK (C1298/A8362 supported by the Bobby Moore Fund (R.S.H., B.K. and P.B.), the Wellcome Trust (R.S.H., B.K. and P.B.) and the DJ Fielding Medical Research Trust (R.S.H., B.K. and P.B.). The National Brain Tumor Study is supported by the National Cancer Research Network, and we acknowledge all clinicians and healthcare professionals who contributed to this initiative. The UK INTERPHONE study was supported by the European Union Fifth Framework Program 'Quality of Life and Management of Living Resources' (QLK4-CT-1999-01563) (A.S., M.J.S. and S.J.F.) and the International Union against Cancer (UICC) (A.S., M.J.S. and S.J.F.). The UICC received funds from the Mobile Manufacturers' Forum and the GSM Association. Provision of funds via the UICC was governed by agreements that guaranteed INTERPHONE's scientific independence (http://www.iarc.fr/ENG/Units/RCAd.html), and the views expressed in the paper are not necessarily those of the funders. The UK centers were also supported by the Mobile Telecommunications and Health Research (MTHR) Programme, and the Northern UK Centre (A.S., M.J.S. and S.J.F.) was supported by the Health and Safety Executive, Department of Health and Safety Executive and the UK Network Operators.

In France, funding was provided by the Ligue Nationale Contre le Cancer (J.-Y.D.), the Fondation ARC (M. Sanson), the Institut National du Cancer (INCa; PL046; (M. Sanson)), the French Ministry of Higher Education and Research and the program "Investissements d'avenir" ANR-10-IAIHU-06 (M. Sanson, J.-Y.D., M. Labussière, A.-L.D.S., P.G., K.M., A.I., K.H.-X. and K.L.). This study was additionally supported by a grant from Génome Québec, le Ministère de l'Enseignement supérieur, de la Recherche, de la Science et de la Technologie (MESRST) Québec (M. Lathrop) and McGill University (M. Lathrop).

In Germany, funding was provided to M. Simon and J. Schramm by the Deutsche Forschungsgemeinschaft (Si552, Schr285), the Deutsche Krebshilfe (70-2385-Wi2, 70-3163-Wi3, 10-6262) and BONFOR. Funding for the WTCCC was provided by the Wellcome Trust (076113 and 085475; M. Simon and J. Schramm). The KORA Ausburg studies are supported by grants from the German Federal Ministry of Education and Research (BMBF) and were mainly financed by the Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg. This work was financed by the German National Genome Research Network (NGFN) (S. Schreiber and H.E.-W.) and supported within the Munich Center of Health Sciences (MC Health) as part of LMUinnovativ (S. Schreiber and H.E.-W.). Generation of the German control data

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

was partially supported by a grant of the German Federal Ministry of Education and Research (BMBF) through the Integrated Network IntegraMent (Integrated Understanding of Causes and Mechanisms in Mental Disorders), under the auspices of the e:Med research and funding concept (01ZX1314A) (M.M.N., S. Herms and S. Heilmann). M.M.N. is a member of the DFG-funded Excellence Cluster ImmunoSensation and received support from the Alfried Krupp von Bohlen und Halbach-Stiftung.

For the UK GWAS, we acknowledge the funders, organizations and individuals who contributed to the blood sample and data collection as listed in Hepworth *et al.*[45]. MD Anderson acknowledges the work of P. Adatto, F. Morice, H. Zhang, V. Levin, A.W.K. Yung, M. Gilbert, R. Sawaya, V. Puduvalli, C. Conrad, F. Lang and J. Weinberg from the Brain and Spine Center for the MDA GWAS. For the French study, we are indebted to A. Rahimian (Onconeurotek), A.M. Lekieffre and M. Brandel for help in collecting data and to Y. Marie for database support. For the German study, we are indebted to B. Harzheim (Bonn), S. Ott and A. Müller-Erkwoh (Bonn) for help with the acquisition of clinical data and to R. Mahlberg (Bonn), who provided technical support. The UK study made use of control genotyping data generated by the Wellcome Trust Case–Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. The MDA GWAS made use of control genotypes from the CGEMS prostate and breast cancer studies. A full list of the investigators who contributed to the generation of the data is available from http://cgems.cancer.gov/. French controls were taken from the SU.VI.MAX study. The German GWAS made use of genotyping data from three population control sources: KORA-gen39, the Heinz-Nixdorf RECALL study and POPGEN. The HNR cohort was established with the support of the Heinz-Nixdorf Foundation. F.D. received support from the BONFOR Programme of the University of Bonn, Germany.

The UCSF Adult Glioma Study was supported by the NIH (grant numbers R01CA52689 (M.R.W. and J.K.W.), P50CA097257 (M.R.W. and J.K.W.), R01CA126831 (J.K.W.) and R01CA139020 (M.R.W.)), the Loglio Collective (M.R.W. and J.K.W.), the National Brain Tumor Foundation (M.R.W.), the Stanley D. Lewis and Virginia S. Lewis Endowed Chair in Brain Tumor Research (M.R.W.), the Robert Magnin Newman Endowed Chair in Neuro-oncology (J.K.W.) and by donations from the families and friends of J. Berardi, H. Glaser, E. Olsen, R.E. Cooper and W. Martinusen. This project also was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, NIH, through UCSF–CTSI grant UL1 RR024131 (UCSF CTSI). The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The collection of cancer incidence data used in this study was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code section 103885, the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201000140C (awarded to the Cancer Prevention Institute of California), contract HHSN261201000035C (awarded to the University of Southern California) and contract HHSN261201000034C (awarded to the Public Health Institute), and the Centers for Disease Control and Prevention's National Program of Cancer Registries under agreement # U58DP003862-01 (awarded to the California Department of Public Health). The ideas and opinions expressed herein are those of the author(s), and endorsement by the State of California Department of Public Health, the National Cancer Institute and the Centers for Disease Control and Prevention, or their contractors and subcontractors, is not intended nor should be inferred. Other significant contributors for the UCSF Adult Glioma Study include M. Berger, P. Bracci, S. Chang, J. Clarke, A. Molinaro, A. Perry, M. Pezmecki, M. Prados, I. Smirnov, T. Tihan, K. Walsh, J. Wiemels and S. Zheng.

At Mayo, the authors wish to acknowledge the study participants and the clinicians and research staff at the participating medical centers, the Mayo Clinic Biobank and Biospecimens Accessioning and Processing Shared Resource (in particular its manager, M. Cicek). Work at the Mayo Clinic beyond the GICC was also supported by the NIH (grants P50CA108961 (B. O'Neill) and P30CA15083 (R. Diasio)), the National Institute of Neurological Disorders and Stroke (grant RC1NS068222Z (R.B.J.)), the Bernie and Edith Waterman Foundation (R.B.J.) and the Ting Tsung and Wei Fong Chao Family Foundation (R.B.J.).

The GliomaScan Consortium comprised (apart from authors listed in the author list): L.E.B. Freeman (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA), S. Koutros (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA), D. Albanes (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA), K. Visvanathan (Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA and Johns Hopkins Sidney Kimmel Comprehensive Cancer Center, Baltimore, Maryland, USA), V.L. Stevens (Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA), R. Henriksson (Department of Radiation Sciences, Oncology, Umea University, Umea, Sweden), D.S. Michaud (Department of Public Health and Community Medicine, Tufts University Medical School, Boston, Massachusetts, USA), M. Feychting (Unit of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden), A. Ahlbom (Unit of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden), G.G. Giles (Cancer Epidemiology Centre, Cancer Council of Victoria, Melbourne, Victoria, Australia and Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne, Melbourne, Victoria, Australia), R. Milne (Cancer Epidemiology Centre, Cancer Council of Victoria, Melbourne, Victoria, Australia and Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne, Melbourne, Victoria, Australia), R. McKean-Cowdin (Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA), L. Le Marchand (Cancer Research Center, University of Hawaii, Honolulu, Hawaii, USA), M. Stampfer (Channing Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA and Departments of Epidemiology and Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA), A.M. Ruder (National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, Cincinnati, Ohio, USA), T. Carreon (National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, Cincinnati, Ohio, USA), G. Hallmans (Department of Public Health and Clinical Medicine/Nutritional Research, Umea University, Umea, Sweden), A. Zeleniuch-Jacquotte (Division of Epidemiology, Department of Environmental Medicine, New York University School of Medicine, New York, New York, USA), J.M. Gaziano (Massachusetts Veteran's Epidemiology, Research and Information Center, Geriatric Research Education and Clinical Center, VA Boston Healthcare System, Boston, Massachusetts, USA), H.D. Sesso (Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA), M.P. Purdue (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA), E. White (Fred Hutchinson Cancer Research Center, Seattle, Washington, USA and Department of Epidemiology, University of Washington, Seattle, Washington, USA) and J. Buring (Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA).

UK10K data generation and access was organized by the UK10K consortium and funded by the Wellcome Trust.

AUTHOR CONTRIBUTIONS

M.L.B., B.S.M., R.S.H. and J.S.B.-S. managed the project; R.S.H., M.L.B., B.S.M., J.S.B.-S., R.B.J., Q.T.O., B.K. and M.R.W. drafted the manuscript; Q.T.O., K.L., B.K., J.E.E.-P. and P.A.D. performed statistical analyses; Y.C., K.L., Y.L. and B.K. performed bioinformatics analyses; B.S.M., J.S.B.-S., M.R.W., J.K.W., C.J., D.I., R.K.L., G.A., P.A.D., U.A., T.R., H.H., L.M., M.L.K., H.S., J.L.B., F.D., D.L, C.I.A, C.L., R.T.M., J. Shildkraut, F.A.-O., S. Sadetski, M. Scheurer, S. Shete, E.B.C., S.H.O., R.B.J., R.S.H. and M.L.B. developed the GICC protocol and performed sample acquisition; and P.R., S.C., M. Linet, Z.W. and M.Y. provided the National Cancer Institute (NCI) data. In the UK, P.B., A.S., M.J.S., S.J.F. and R.S.H. developed patient recruitment, performed sample acquisition and performed sample collection of cases; P.B. oversaw DNA isolation and storage, and performed case and control ascertainment, and supervision of DNA extractions. In Germany, M. Simon, M.M.N., H.-E.W., S. Schreiber and J. Schramm developed patient recruitment, and oversaw performed blood sample collection; M. Simon oversaw DNA isolation and storage and performed case and control ascertainment, and supervision of DNA extractions; and S. Herms, S. Heilmann and K.G. performed experimental work. In France, M. Sanson and J.-Y.D. developed patient recruitment; M. Labussière, A.-L.D.S., P.G., K.M., A.I. and K.H.-X. performed patient ascertainment. M. Lathrop performed laboratory management and oversaw genotyping of the French samples. All authors contributed to the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Bondy, M.L. *et al.* Brain tumor epidemiology: consensus from the Brain Tumor Epidemiology Consortium. *Cancer* **113**, 1953–1968 (2008).
2. Ostrom, Q.T. *et al.* CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2008–2012. *Neuro-oncol.* **17** (Suppl. 4), iv1–iv62 (2015).
3. Louis, D.N. *et al.* The 2007 WHO classification of tumors of the central nervous system. *Acta Neuropathol.* **114**, 97–109 (2007).
4. Ostrom, Q.T. *et al.* CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2007–2011. *Neuro-oncol.* **16**, iv1–iv63 (2014).
5. Ostrom, Q.T. *et al.* The epidemiology of glioma in adults: a 'state of the science' review. *Neuro-oncol.* **16**, 896–913 (2014).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

6. Hemminki, K., Tretli, S., Sundquist, J., Johannesen, T.B. & Granstrom, C. Familial risks in nervous-system tumors: a histology-specific analysis from Sweden and Norway. *Lancet Oncol.* **10**, 481–488 (2009).
7. Sanson, M. *et al.* Chromosome 7p11.2 (*EGFR*) variation influences glioma risk. *Hum. Mol. Genet.* **20**, 2897–2904 (2011).
8. Shete, S. *et al.* Genome-wide association study identifies five susceptibility loci for glioma. *Nat. Genet.* **41**, 899–904 (2009).
9. Wrensch, M. *et al.* Variants in the *CDKN2B* and *RTEL1* regions are associated with high-grade glioma susceptibility. *Nat. Genet.* **41**, 905–908 (2009).
10. Kinnersley, B. *et al.* Genome-wide association study identifies multiple susceptibility loci for glioma. *Nat. Commun.* **6**, 8559 (2015).
11. Walsh, K.M. *et al.* Variants near *TERT* and *TERC* influencing telomere length are associated with high-grade glioma risk. *Nat. Genet.* **46**, 731–735 (2014).
12. Jenkins, R.B. *et al.* A low-frequency variant at 8q24.21 is strongly associated with risk of oligodendroglial tumors and astrocytomas with *IDH1* or *IDH2* mutation. *Nat. Genet.* **44**, 1122–1125 (2012).
13. Rajaraman, P. *et al.* Genome-wide association study of glioma and meta-analysis. *Hum. Genet.* **131**, 1877–1888 (2012).
14. Stacey, S.N. *et al.* A germline variant in the *TP53* polyadenylation signal confers cancer susceptibility. *Nat. Genet.* **43**, 1098–1103 (2011).
15. Kinnersley, B. *et al.* Quantifying the heritability of glioma using genome-wide complex trait analysis. *Sci. Rep.* **5**, 17267 (2015).
16. Lonsdale, J. *et al.* The Genotype–Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
17. Westra, H.J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
18. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex-trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
19. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
20. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
21. Ruark, E. *et al.* Identification of nine new susceptibility loci for testicular cancer, including variants near *DAZL* and *PRDM14*. *Nat. Genet.* **45**, 686–689 (2013).
22. Wu, C. *et al.* Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene–environment interactions. *Nat. Genet.* **44**, 1090–1097 (2012).
23. Riemenschneider, M.J. *et al.* Amplification and overexpression of the *MDM4* (*MDMX*) gene from 1q32 in a subset of malignant gliomas without *TP53* mutation or *MDM2* amplification. *Cancer Res.* **59**, 6091–6096 (1999).
24. Boland, E. *et al.* Mapping of deletion and translocation breakpoints in 1q44 implicates the serine–threonine kinase AKT3 in postnatal microcephaly and agenesis of the corpus callosum. *Am. J. Hum. Genet.* **81**, 292–303 (2007).
25. Turner, K.M. *et al.* Genomically amplified Akt3 activates DNA repair pathway and promotes glioma progression. *Proc. Natl. Acad. Sci. USA* **112**, 3421–3426 (2015).
26. Gur, G. *et al.* LRIG1 restricts growth factor signaling by enhancing receptor ubiquitylation and degradation. *EMBO J.* **23**, 3270–3281 (2004).
27. Yang, J.A. *et al.* LRIG1 enhances the radio-sensitivity of radio-resistant human glioblastoma U251 cells via attenuation of the EGFR–AKT signaling pathway. *Int. J. Clin. Exp. Pathol.* **8**, 3580–3590 (2015).
28. Wei, J. *et al.* miR-20a mediates temozolomide resistance in glioblastoma cells via negatively regulating LRIG1 expression. *Biomed. Pharmacother.* **71**, 112–118 (2015).
29. Watanabe, T., Nobusawa, S., Kleihues, P. & Ohgaki, H. *IDH1* mutations are early events in the development of astrocytomas and oligodendrogliomas. *Am. J. Pathol.* **174**, 1149–1153 (2009).
30. Yan, H. *et al. IDH1* and *IDH2* mutations in gliomas. *N. Engl. J. Med.* **360**, 765–773 (2009).
31. Noushmehr, H. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510–522 (2010).
32. Christensen, B.C. *et al.* DNA methylation, isocitrate dehydrogenase mutation and survival in glioma. *J. Natl. Cancer Inst.* **103**, 143–153 (2011).
33. Sanson, M. *et al.* Isocitrate dehydrogenase 1 codon 132 mutation is an important prognostic biomarker in gliomas. *J. Clin. Oncol.* **27**, 4150–4154 (2009).
34. Eckel-Passow, J.E. *et al.* Glioma groups based on 1p/19q, *IDH* and *TERT* promoter mutations in tumors. *N. Engl. J. Med.* **372**, 2499–2508 (2015).
35. Walsh, K.M. *et al.* Telomere maintenance and the etiology of adult glioma. *Neuro-oncol.* **17**, 1445–1452 (2015).
36. Miyake, Y. *et al.* RPA-like mammalian Ctc1–Stn1–Ten1 complex binds to single-stranded DNA and protects telomeres independently of the Pot1 pathway. *Mol. Cell* **36**, 193–206 (2009).
37. Chen, L.Y., Redon, S. & Lingner, J. The human CST complex is a terminator of telomerase activity. *Nature* **488**, 540–544 (2012).
38. Bainbridge, M.N. *et al.* Germline mutations in shelterin complex genes are associated with familial glioma. *J. Natl. Cancer Inst.* **107**, 384 (2014).
39. Zhang, C. *et al.* Genetic determinants of telomere length and risk of common cancers: a mendelian randomization study. *Hum. Mol. Genet.* **24**, 5356–5366 (2015).
40. Walsh, K.M. *et al.* Longer genotypically estimated leukocyte telomere length is associated with increased adult glioma risk. *Oncotarget* **6**, 42468–42477 (2015).
41. Ceccarelli, M. *et al.* Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164**, 550–563 (2016).
42. Xipell, E. *et al.* Endoplasmic-reticulum-stress-inducing drugs sensitize glioma cells to temozolomide through downregulation of MGMT, MPG and Rad51. *Neuro-oncol.* **18**, 1109–1119 (2016).
43. Nicolas, C.S. *et al.* The role of JAK–STAT signaling within the CNS. *JAK-STAT* **2**, e22925 (2013).
44. Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* **372**, 2481–2498 (2015).
45. Hepworth, S.J. *et al.* Mobile phone use and risk of glioma in adults: case–control study. *BMJ* **332**, 883–887 (2006).

[1]Department of Radiation Sciences, Umeå University, Umeå, Sweden. [2]Case Comprehensive Cancer Center, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA. [3]Department of Neurological Surgery, School of Medicine, University of California, San Francisco, San Francisco, California, USA. [4]Institute of Human Genetics, University of California, San Francisco, San Francisco, California, USA. [5]Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen, Denmark and Rigshospitalet, University of Copenhagen, Copenhagen, Denmark. [6]Department of Epidemiology and Biostatistics, School of Public Health, Georgia State University, Atlanta, Georgia, USA. [7]Duke Cancer Institute, Duke University Medical Center, Durham, North Carolina, USA. [8]Cancer Control and Prevention Program, Department of Community and Family Medicine, Duke University Medical Center, Durham, North Carolina, USA. [9]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK. [10]Sorbonne Universités UPMC Univ Paris 06, INSERM CNRS, U1127, UMR 7225, ICM, Paris, France. [11]Department of Medicine, Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, USA. [12]Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, Minnesota, USA. [13]AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de neurologie 2-Mazarin, Paris, France. [14]Université Paris 13 Sorbonne Paris Cité, INSERM U557, INRA U1125, CNAM, Paris, France. [15]Department of Neurosurgery, University of Bonn Medical Center, Bonn, Germany. [16]Centre for Epidemiology and Biostatistics, Faculty of Medicine and Health, University of Leeds, Leeds, UK. [17]Institute of Human Genetics, University of Bonn, Bonn, Germany. [18]Helmholtz Center Munich, Institute of Epidemiology I, Munich, Germany. [19]Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany. [20]Institute of Medical Statistics and Epidemiology, Technical University Munich, Munich, Germany. [21]1st Medical Department, University Clinic Schleswig–Holstein, Campus Kiel, Kiel, Germany. [22]Division of Breast Cancer Research, Institute of Cancer Research, London, UK. [23]Génome Québec, Department of Human Genetics, McGill University, Montreal, Quebec, Canada. [24]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA. [25]A full list of members and affiliations appears in the Acknowledgments. [26]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA. [27]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [28]School of Public Health, University of Alberta, Edmonton, Alberta, Canada. [29]Department of Neurology, Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, Minnesota, USA. [30]Department of Pediatrics, Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, USA. [31]Department of Neurology, NorthShore University HealthSystem, Evanston, Illinois, USA. [32]Department of Surgery, Duke University Medical Center, Durham, North Carolina, USA. [33]Cancer and Radiation Epidemiology Unit, Gertner Institute, Chaim Sheba Medical Center, Tel Hashomer, Israel. [34]Department of Epidemiology and Preventive Medicine, School of Public Health, Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel. [35]Department of Biostatistics, University of Texas Maryland Anderson Cancer Center, Houston, Texas, USA. [36]Departments of Neurology and Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. [37]School of Public Health, Yale University, New Haven, Connecticut, USA. [38]Department of Neurosurgery, Brigham and Women's Hospital, Boston, Massachusetts, USA. [39]Department of Laboratory Medicine and Pathology, Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, Minnesota, USA. [40]Division of Molecular Pathology, Institute of Cancer Research, London, UK. [41]These authors contributed equally to this work. [42]These authors jointly directed this work. Correspondence should be addressed to B.S.M. (beatrice.melin@umu.se), R.S.H. (richard.houlston@icr.ac.uk) or M.L.B. (mbondy@bcm.edu).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

ONLINE METHODS

Ethics. Collection of patient samples and associated clinico-pathological information was undertaken with written informed consent and relevant ethical review board approval at the respective study centers in accordance with the tenets of the Declaration of Helsinki. Specifically informed consent and ethical board approval was obtained from the South-East Multicentre Research Ethics Committee (MREC) (UK), the Scottish MREC (UK), the APHP ethical committee-CPP (Comité de Protection des Personnes) (France), the Ethics Commission of the Medical Faculty of the University of Bonn (Germany), the University of Texas MD Anderson Cancer Institutional Review Board (USA), the Mayo Clinic Office for Human Research Protection (USA), the UCSF Committee on Human Research (USA), the University Hospitals of Cleveland Institutional Review Board (USA) and the Cleveland Clinic Institutional Review Board (board for the Case Comprehensive Cancer Center) (USA). The diagnosis of glioma (ICDO-3 codes 9380-9480 or equivalent) was established through histology in all cases in accordance with World Health Organization guidelines. Every effort was made to classify tumors as GBM or non-GBM.

GWAS data sets. *GICC, UK, French, German, MDA, SFAGS and GliomaScan.* Studies participating in GICC are described in Amirian *et al.*[46] and in **Supplementary Table 1**. Briefly, they comprise 5,189 glioma cases and 3,827 controls that were ascertained through centers in the USA, Denmark, Sweden and the UK. Cases had newly diagnosed glioma, and controls had no personal history of central nervous system tumor at the time of ascertainment. Detailed information regarding recruitment protocol is given in Amirian *et al.*[46]. Cases and controls were genotyped using the Illumina Oncoarray according to the manufacturer's recommendations (Illumina Inc.). Individuals with a call rate <99%, as well as all individuals evaluated to be of non-European ancestry (<80% estimated European ancestry using the FastPop[47] procedure developed by the GAMEON consortium with HapMap version 2 CEU, JPT/CHB and YRI populations as a reference; **Supplementary Fig. 5**), were excluded. For pairs of apparent first-degree relatives, we removed the control from a case–control pair; otherwise, we excluded the individual with the lower call rate. SNPs with a call rate <95% were excluded as were those with a MAF <0.01 or those displaying significant deviation from the Hardy–Weinberg equilibrium (HWE) (i.e., $P < 10^{-5}$). After performing these quality-control measures, there were 4,572 cases and 3,286 controls remaining for downstream analyses.

The UK, French, German, MDA, SFAGS and GliomaScan GWAS of non-overlapping case–control series of Northern European ancestry have been the subject of previous studies. Briefly, the UK GWAS[7,8,10] was based on 636 cases (401 males; mean age 46 years) who were ascertained through the INTERPHONE study[48]. Individuals from the 1958 Birth Cohort (n = 2,930) served as a source of controls. The French GWAS[7,10] comprised 1,495 patients with glioma who were ascertained through the Service de Neurologie Mazarin, Groupe Hospitalier Pitié-Salpêtrière Paris. The controls (n = 1,213) were ascertained from the SU.VI.MAX (Supplementation en Vitamines et Mineraux Antioxydants) study of 12,735 healthy subjects (women aged 35–60 years; men aged 45–60 years)[49]. The German GWAS[10] comprised 880 patients who had undergone surgery for a glioma at the Department of Neurosurgery, University of Bonn Medical Center, between 1996 and 2008. Control subjects were taken from three population studies: KORA (Cooperative Health Research in the Region of Augsburg; n = 488)[50]; POPGEN (Population Genetic Cohort; n = 678)[51] and the Heinz Nixdorf Recall study (n = 380)[52]. Standard quality-control measures were applied to the UK, French and German GWAS and have previously been reported. The MDA GWAS[8] was based on 1,281 cases (786 males; mean age 47 years) who were ascertained through the MD Anderson Cancer Center, Texas, between 1990 and 2008. Individuals from the Cancer Genetic Markers of Susceptibility (CGEMS, n = 2,245) studies served as controls[53,54]. Quality-control measures were applied as per the primary GWAS. The UCSF adult glioma case–control study (SFAGS–GWAS) included participants of the San Francisco Bay Area Adult Glioma Study (AGS). Details of subject recruitment for AGS have been reported previously[9,12,34,55,56]. Briefly, cases were adults (>18 years of age) with newly diagnosed, histologically confirmed glioma. Population-based cases who were diagnosed between 1991 and 2009 (series 1–4) and who were residing in the six San Francisco Bay area counties were ascertained using the Cancer Prevention Institute of California's early-case ascertainment system. Clinic-based cases who were diagnosed between 2002 and 2012 (series 3–5) were recruited from the UCSF Neuro-oncology Clinic, regardless of the place of residence. From 1991 to 2010, population-based controls from the same residential area as the population-based cases were identified using random digit-dialing and were frequency matched to population-based cases for age, gender and ethnicity. Between 2010 and 2012, all controls were selected from the UCSF general medicine phlebotomy clinic. Clinic-based controls were matched to clinic-based glioma cases for age, gender and ethnicity. Consenting participants provided blood, buccal and/or saliva specimens, and information, during in-person or telephone interviews. A total of 677 cases and 3,940 controls (including 3,347 Illumina iControlDB iControls) were used in the current analysis. For the GliomaScan GWAS[13], in addition to the published analysis, we excluded samples from the ATBC (Finnish study) and controls from NSHDS due to exhibiting outlying population ancestry after manual inspection of PCA plots. In total 1,653 cases and 2,725 controls were used in the current study.

GWAS data from the seven studies were imputed to >10 million SNPs with IMPUTE2 (v2.3)[57] software using a merged reference panel consisting of data from the 1000 Genomes Project (phase 1 integrated release 3, March 2012)[58] and UK10K (ALSPAC, EGAS00001000090 and EGAD00001000195, and TwinsUK EGAS00001000108 and EGAS00001000194 studies). Genotypes were aligned to the positive strand in both imputation and genotyping. Imputation was conducted separately for each study, and in each the data were pruned to a common set of SNPs between cases and controls before imputation. We set thresholds for imputation quality to retain potential risk variants with MAF > 0.01. Poorly imputed SNPs, defined by an information measure <0.40 with IMPUTE2, were excluded, as were SNPs exhibiting a significant deviation from Hardy–Weinberg equilibrium ($P < 1 \times 10^{-8}$) in controls. Test of association between imputed SNPs and glioma was performed using SNPTEST (v2.5)[59] under an additive frequentist model. The adequacy of the case–control matching and the possibility of differential genotyping of cases and controls were formally evaluated using Q-Q plots of test statistics (**Supplementary Fig. 1**). Where appropriate, principal components, generated using common SNPs, were included in the analysis to limit the effects of cryptic population stratification that otherwise might cause inflation of test statistics. Principal components, based on genotyped SNPs, were generated for the GICC, GliomaScan, MDA-GWAS and SFAGS studies using PLINK[60]. Eigenvectors for the German GWAS were inferred using smartpca (part of EIGENSOFTv2.4)[61] by merging cases and controls with Phase II HapMap samples[10]. PCA plots for all studies are provided in **Supplementary Figure 4**.

UCSF-Mayo GWAS. The UCSF-Mayo study comprised Mayo cases (n = 945) and UCSF cases (n = 574) and Mayo Clinic Biobank control (n = 806) data. The Mayo Clinic case–control study has been described previously[9,34,62]. Briefly, adult cases (>18 years of age) were identified at diagnosis (diagnosed at Mayo Clinic) or at pathologic confirmation (diagnosed elsewhere and treated at Mayo Clinic), and the patients had a surgical resection or biopsy between 1973 and 2014. Consenting participants provided blood, buccal and/or saliva specimens, and information, during in-person or telephone interviews. This analysis used 574 non-overlapping cases from the UCSF Adult Glioma Study described above. Mayo Clinic and UCSF cases were genotyped using the Illumina Oncoarray. The Mayo Clinic Biobank controls comprised volunteers who donated biological specimens and provided risk factor data, access to clinical data obtained from the medical record and consent to participate in any study approved by the Access Committee. Recruitment for the Mayo Clinic Biobank took place from April 2009 through December 2015. Although participants could be unselected volunteers, the vast majority of participants were contacted as part of a pre-scheduled medical examination in the Department of Medicine, Divisions of Community Internal Medicine, Family Medicine and General Internal Medicine at Mayo Clinic sites in Rochester (Minnesota), Jacksonville (Florida), and the Mayo Clinic Health System sites in La Crosse and Onalaska (Wisconsin). All individuals were aged 18 years and older at the time of consent. Illumina Omni Express genotyping arrays were run on the 806 Mayo Clinic Biobank participants.

Quality-control analyses were performed on each cohort separately (Mayo cases, UCSF cases and Mayo Clinic Biobank controls). SNPs with call rates <95% were removed, followed by removal of subjects with call rates <95%. Concordance of replicate samples was assessed, and the sample with the higher

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

call rate was retained. Subject's sex was verified using the sex check option in PLINK. Relationship checking was performed by estimating the proportion of alleles shared identical by descent (IBD) for all pairs of subjects in PLINK[60]. STRUCTURE[63] was used to assess population admixture with 1000 Genomes as a reference. Subjects indicated to be non-Caucasian were excluded. Prior to imputation, SNPs were tested for HWE, and SNPs with HWE $P < 10^{-6}$ were removed. Mayo Clinic, UCSF and Mayo Clinic Biobank SNP data were each phased and imputed using the Michigan Imputation Server with the Haplotype Reference Consortium (release 1; http://www.haplotype-reference-consortium.org) as reference. Genotypes were forward-strand-aligned to the 1000 Genomes reference, and for ambiguous SNPs the Browning strand checking utility was used (http://faculty.washington.edu/sguy/beagle/strand_switching/strand_switching.html). PCA was used to correct for population stratification using SNPs common to cases and controls. The first three principal components were significantly ($P < 0.05$) associated with case–control status. An additive logistic regression model was used to assess the association between each SNP and disease status, with genotype being coded as 0, 1 or 2 copies of the minor allele, adjusted for age, sex and the first three principal components.

Meta-analysis and additional statistical analyses. Meta-analyses were performed using the fixed-effects inverse-variance method based on the β-estimates and standard errors from each study using META (v1.6)[64]. Cochran's Q-statistic was used to test for heterogeneity, and the I^2 statistic was used to quantify the proportion of the total variation due to heterogeneity[65], taking I^2 values >75 to indicate significant heterogeneity. Using the meta-analysis summary statistics and LD correlations from a reference panel of the 1000 Genomes Project combined with UK10K, we used GCTA[66,67] to perform conditional association analysis. Association statistics were calculated for all SNPs, conditioning on the top SNP in each locus showing genome-wide significance. This was carried out in a step-wise fashion. We performed a case-only analysis to test for differences in SNP-risk-allele frequency between GBM and non-GBM tumors.

ENCODE and chromatin state dynamics. Risk SNPs and their proxies (i.e., $r^2 > 0.8$ in the 1000 Genomes EUR reference panel) were annotated for putative functional effect using HaploReg (v4)[68], RegulomeDB[69] and SeattleSeq Annotation[70]. These servers make use of data from ENCODE, genomic evolutionary rate profiling (GERP) conservation metrics, combined annotation-dependent depletion (CADD) scores and PolyPhen scores. We searched for overlap of associated SNPs with enhancers defined by the FANTOM5 enhancer atlas[19], annotating by overlap with ubiquitous, permissive and robust enhancers, as well as enhancer–promoter correlations and enhancers specifically expressed in astrocytes, neuronal stem cells and brain tissue. Similarly, we searched for overlap with 'super-enhancer' regions, as defined by Hnisz *et al.*[20], restricting analysis to data from U87 GBM cells, astrocyte cells and brain tissue. We additionally made use of 15-state chromHMM data from H1- and H9-derived neuronal progenitor cells available from the Epigenome Roadmap Project[71]. Enhancer enrichment analysis was carried out using HaploReg (v4.0)[68]. Briefly, from a query list of variants, the overlap with enhancers in each of 107 cell types, as predicted from the Roadmap Epigenomics Project[71] chromatin-state segmentations, was calculated. A binomial test for enrichment was performed against a background set of all (i) 1000 Genomes variants with MAF > 0.05 and (ii) all unique GWAS loci in the European population. We applied a cutoff of $P < 3.94 \times 10^{-4}$ corresponding to a Bonferroni correction for 127 cell lines and tissues.

Expression quantitative trait loci (eQTL) analysis. To examine the relationship between SNP genotype and gene expression, we carried out summary-data-based mendelian randomization (SMR) analysis as per Zhu *et al.*[18] (at http://cnsgenomics.com/software/smr/index.html). We used publicly available brain tissue data from the GTEx[16] (http://www.gtexportal.org) v6p release. Briefly, GWAS summary statistics files were generated from the meta-analysis. Reference files were generated from merging 1000 Genomes phase 3 and UK10K (ALSPAC and TwinsUK) vcfs. Summary eQTL files for GTEx samples were generated from downloaded v6p "all_snpgene_pairs" files. Besd files were generated from these summary eQTL files using the –make-besd command. Additionally, we analyzed downloaded whole-blood eQTL data from Westra *et al.*[17]. Results from the SMR test for each of the 13 new glioma loci are reported in **Supplementary Data 3**. As previously advocated[18], only probes with at least one eQTL P value $<5.0 \times 10^{-8}$ were considered for SMR analysis. We set a threshold for the SMR test of $P_{SMR} < 1.06 \times 10^{-4}$ corresponding to a Bonferroni correction for 473 tests (473 probes with a top eQTL $P < 5.0 \times 10^{-8}$ across the 13 loci, 10 brain regions and Westra data set). For all genes passing this threshold, we generated plots of the eQTL and GWAS associations at the locus, as well as plots of GWAS and eQTL effect sizes (i.e., corresponding to input for the HEIDI heterogeneity test). HEIDI test P values <0.05 were taken to indicate significant heterogeneity. Respective SMR plots for significant eQTLs are shown in **Supplementary Figure 4**.

Additional statistical and bioinformatics analysis. Estimates of individual variance in risk associated with glioma risk SNPs was carried out using the method described in Pharoah *et al.*[72], assuming the familial risk of high-grade and low-grade glioma to be 1.76 and 1.54, respectively, from analysis of the Swedish series in Scheurer *et al.*[73]. Briefly, for a single allele (i) of frequency p, relative risk R and ln risk r, the variance (V_i) of the risk distribution due to that allele is given by:

$$V_i = (1-p)^2 E^2 + 2p(1-p)(r-E)^2 + p^2(2r-E)^2$$

Where E is the expected value of r given by:

$$E = 2p(1-p)r + 2p^2 r$$

For multiple risk alleles, the distribution of risk in the population tends toward the normal with variance:

$$V = \sum V_i$$

The total genetic variance (V) for all susceptibility alleles has been estimated to be $\sqrt{1.77}$. Thus, the fraction of the genetic risk explained by a single allele is given by:

$$V_i / V$$

LD metrics were calculated in vcftools (v0.1.12b)[74] using UK10K data and plotted using visPIG[75]. LD blocks were defined on the basis of HapMap recombination rate (cM/Mb), as defined using the Oxford recombination hotspots and on the basis of distribution of confidence intervals defined by Gabriel *et al.*[76].

Data availability. Genotype data from the GICC GWAS are available from the database of Genotypes and Phenotypes (dbGaP) under accession phs001319.v1.p1. Additionally, genotypes from the GliomaScan GWAS can be accessed through dbGaP accession phs000652.v1.p1. Data from the other studies are available upon request.

46. Amirian, E.S. *et al.* The Glioma International Case–Control Study: a report from the Genetic Epidemiology of Glioma International Consortium. *Am. J. Epidemiol.* **183**, 85–91 (2016).
47. Li, Y. *et al.* FastPop: a rapid principal component–derived method to infer intercontinental ancestry using genetic data. *BMC Bioinformatics* **17**, 122 (2016).
48. Cardis, E. *et al.* The INTERPHONE study: design, epidemiological methods and description of the study population. *Eur. J. Epidemiol.* **22**, 647–664 (2007).
49. Hercberg, S. *et al.* The SU.VI.MAX study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. *Arch. Intern. Med.* **164**, 2335–2342 (2004).
50. Wichmann, H.E., Gieger, C., Illig, T. & MONICA–KORA Study Group. KORA-gen—resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* **67** (Suppl. 1), S26–S30 (2005).
51. Krawczak, M. *et al.* PopGen: population-based recruitment of patients and controls for the analysis of complex genotype–phenotype relationships. *Community Genet.* **9**, 55–61 (2006).
52. Schmermund, A. *et al.* Assessment of clinically silent atherosclerotic disease, and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the Heinz Nixdorf RECALL study. Risk factors, evaluation of coronary calcium and lifestyle. *Am. Heart J.* **144**, 212–218 (2002).
53. Hunter, D.J. *et al.* A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39**, 870–874 (2007).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

doi:10.1038/ng.3823

54. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).
55. Wiemels, J.L. *et al.* History of allergies among adults with glioma and controls. *Int. J. Cancer* **98**, 609–615 (2002).
56. Felini, M.J. *et al.* Reproductive factors and hormone use, and risk of adult gliomas. *Cancer Causes Control* **20**, 87–96 (2009).
57. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
58. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
59. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
60. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
61. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
62. Jenkins, R.B. *et al.* Distinct germline polymorphisms underlie glioma morphologic heterogeneity. *Cancer Genet.* **204**, 13–18 (2011).
63. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
64. Liu, J.Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat. Genet.* **42**, 436–440 (2010).
65. Higgins, J.P., Thompson, S.G., Deeks, J.J. & Altman, D.G. Measuring inconsistency in meta-analyses. *Br. Med. J.* **327**, 557–560 (2003).
66. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex-trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
67. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
68. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
69. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
70. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
71. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
72. Pharoah, P.D., Antoniou, A.C., Easton, D.F. & Ponder, B.A. Polygenes, risk prediction and targeted prevention of breast cancer. *N. Engl. J. Med.* **358**, 2796–2803 (2008).
73. Scheurer, M.E. *et al.* Familial aggregation of glioma: a pooled analysis. *Am. J. Epidemiol.* **172**, 1099–1107 (2010).
74. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
75. Scales, M., Jager, R., Migliorini, G., Houlston, R.S. & Henrion, M.Y. visPIG—a web tool for producing multiregion, multitrack, multiscale plots of genetic data. *PLoS One* **9**, e107497 (2014).
76. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

Published Ahead of Print on July 8, 2016 as 10.1634/theoncologist.2016-0003.

The Oncologist®

Neuro-Oncology

Chromosome 17p Homodisomy Is Associated With Better Outcome in 1p19q Non-Codeleted and *IDH*-Mutated Gliomas

Marianne Labussière,[a,b,c] Amithys Rahimian,[a,b,c,d] Marine Giry,[a,b,c] Blandine Boisselier,[a,b,c,e] Yohann Schmitt,[a,b,c] Marc Polivka,[f] Karima Mokhtari,[a,b,c,d,g] Jean-Yves Delattre,[a,b,c,d,h] Ahmed Idbaih,[a,b,c,h] Karim Labreche,[a,b,c,i] Agusti Alentorn,[a,b,c,h] Marc Sanson[a,b,c,d,h]

[a]Sorbonne Universités, Université Pierre et Marie Curie, Université Paris 06, Centre de Recherche de l'Institut du Cerveau et de la Moelle Épinière, Paris, France; [b]INSERM U1127, Paris, France; [c]Centre National de la Recherche Scientifique, Unité de Recherche Mixte 7225, Paris, France; [d]OncoNeuroTek, Paris, France; [e]Plateforme de Génotypage Séquençage, Institut du Cerveau et de la Moelle Épinière, Paris, France; [f]Laboratoire d'Anatomie Pathologique, Hôpital Lariboisière, [g]Service de Neuropathologie Raymond Escourolle, Groupe Hospitalier Pitié-Salpêtrière, and [h]Service de Neurologie 2, Groupe Hospitalier Pitié-Salpêtrière, Assistance Publique Hôpitaux de Paris, Paris, France; [i]Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey, United Kingdom

Disclosures of potential conflicts of interest may be found at the end of this article.

Key Words. Gliomas • Copy number neutral loss of heterozygosity • *TP53* mutation

Downloaded from http://theoncologist.alphamedpress.org/ by James Studd on October 17, 2016

Abstract

Background. The 1p19q non-codeleted gliomas with *IDH* mutation, defined as "molecular astrocytomas," display frequent *TP53* mutations and have an intermediate prognosis. We investigated the prognostic impact of copy number-neutral loss of heterozygosity (CNLOH) in 17p in this population.

Methods. We analyzed 793 gliomas (206 grade II, 377 grade III, and 210 grade IV) by single nucleotide polymorphism array and for *TP53* mutations.

Results. Homodisomy revealed by CNLOH was observed in 156 cases (19.7%). It was more frequent in astrocytomas and oligoastrocytomas (98/256, 38%) than oligodendrogliomas (28/327, 8.6%; $p < .0001$) or glioblastoma multiforme (30/210, 14.3%; $p < .0001$), tightly associated with *TP53* mutation (69/71 vs. 20/79; $p = 2 \times 10^{-16}$), and mutually exclusive with 1p19q codeletion (1/156 vs. 249/556; $p < .0001$). In the group of *IDH*-mutated 1p19q non-codeleted gliomas, CNLOH 17p was associated with longer survival (86.3 vs. 46.2 months; $p = .004$), particularly in grade III gliomas (overall survival >100 vs. 37.9 months; $p = .007$). These data were confirmed in an independent dataset from the Cancer Genome Atlas.

Conclusion. CNLOH 17p is a prognostic marker and further refines the molecular classification of gliomas. ***The Oncologist*** 2016;21:1–5

Implications for Practice: Homodisomy of chromosome 17p (CNLOH 17p) is a frequent feature in *IDH*-mutated 1p19q non-codeleted gliomas (group 2). It is constantly associated with *TP53* mutation. It was found, within this specific molecular group of gliomas (corresponding to molecular astrocytomas), that CNLOH 17p is associated with a much better outcome and may therefore represent an additional prognostic marker to refine the prognostic classification of gliomas.

Introduction

Independently of histological grading, gliomas can be separated into three distinct prognostic subgroups according to the presence of *IDH* mutation and 1p19q codeletion: group 1, glioma with 1p19q codeletion, has the best survival; group 2, non-codeleted glioma with *IDH* mutation, has an intermediate prognosis; and group 3, *IDH* wild-type glioma, has the poorest outcome [1–3]. Groups 1 and 2 also differ by the occurrence of mutually exclusive mutations: *TERT* promoter (90%), *CIC* (50%–60%), and *FUBP1* (15%–20%) for group 1 and *ATRX* mutation (associated with the alternative lengthening telomeres phenotype) and *TP53* mutation for group 2 [2–4]. Recent single nucleotide polymorphism (SNP) analysis showed several cases of copy neutral loss of heterozygosity (CNLOH) with duplication of the retained allele. The presence of CNLOH in glial tumors has been reported to affect several genomic regions [5–9]. In a recent report on anaplastic oligodendrogliomas, CNLOH frequently affected the short arm of chromosome 17 [5]. Moreover, Yin et al. described eight cases with CNLOH 17p in a series of 55 glioblastomas [9]. To date, the frequency and prognostic significance of this alteration have not been investigated.

Correspondence: Marc Sanson, M.D., Ph.D., Service de Neurologie 2, Groupe Hospitalier Pitié-Salpêtrière, 75651, Paris cedex 13, France. Telephone: 33-1-42-16-03-91; E-Mail: marc.sanson@psl.aphp.fr Received January 3, 2016; accepted for publication April 14, 2016. ©AlphaMed Press 1083-7159/2016/$20.00/0 http://dx.doi.org/10.1634/theoncologist.2016-0003

©AlphaMed Press 2016

Published Ahead of Print on July 8, 2016 as 10.1634/theoncologist.2016-0003.

In this study, we investigated the presence of CNLOH 17p in a large cohort of grade II–IV glial tumors, analyzed the associations with *TP53* mutation and other molecular alterations, and investigated the prognostic impact of CNLOH 17p.

Patients and Methods

Patients and Tissue Samples

Patients were selected according to the following criteria: histologic diagnosis of primary glial tumor, clinical data and follow-up available in the neuro-oncology database (OncoNeurotek, Groupe Hospitalier Pitié Salpêtrière, Paris, France), and written informed consent. Corresponding clinical annotations were collected from the neuro-oncology department database. As a duplication cohort, we used the DNA sequencing, copy number variant (level 1 copy number data), and survival data (level 3) from lower-grade gliomas (LGGs) of the Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov).

DNA Isolation and SNP Array

Tumor DNA from cryopreserved samples was extracted using the QIAmp DNA Midi Kit (Qiagen, Hilden, Germany, http://www.qiagen.com) according to the manufacturer's instructions. DNA was extracted from blood samples by conventional saline method, quantified using a NanoVue spectrophotometer, and qualified by agarose gel electrophoresis. Tumor DNA was run on an Infinium Illumina Human 610-Quad SNP array (Illumina, San Diego, CA, http://www.illumina.com). Array processing, using 250 ng tumor DNA, was outsourced to Integragen, Évry, France. Extracted data using Feature Extraction software were imported and analyzed using Nexus 5.1 (Biodiscovery, El Segundo, CA, http://www/biodiscovery.com), as previously described [10]. The confirmatory cohort from LGG TCGA was analyzed using PennCNV-Affy from the PennCNV algorithm [11] to convert raw CEL files from LGG TCGA into log R ratio and B-allele frequency. Log R ratio and B-allele frequency files were used to perform allele-specific copy number analysis with GC correction using ASCAT (version 2.4) [12]. We considered loss of heterozygosity in a given chromosome region when $\geq$95% of SNP probes in a DNA segment of at least 500 kb exhibited B-allele frequencies $\geq$0.8 and $\leq$0.2. Loss of heterozygosity with a copy number of 2 was considered CNLOH. Only terminal CNLOH on chromosome 17p with a minimum size of 5 Mb was considered. Molecular characterization of glioma samples (*IDH1/2* mutation, *TERT* promoter mutation, and *MGMT* promoter methylation) was performed as previously described [13].

TP53 Pyrosequencing

Coding exons (2–11) of *TP53* gene were first amplified using primers detailed in supplemental online Table 1. Amplification conditions were 94°C for 3 minutes followed by 45 cycles of 94°C for 15 seconds, 60°C for 45 seconds, and 72°C for 1 minute, with a final step at 72°C for 8 minutes. Polymerase chain reaction (PCR) products were purified conforming to the Agencourt AMPure XP PCR purification protocol (Beckman-Coulter, Nyon, Switzerland, http://www.beckmancoulter.com) with the Biomek 3000 Automation Workstation. Universal tailed amplicon resequencing approach (454 Sequencing Technology; Roche, Basel, Switzerland, http://www.roche.com) was used for sequencing of coding exons of *TP53*. This system includes a second PCR, aiming for multiplex identifiers and incorporation of 454 adaptors, an emulsion PCR according to the emPCR Amplification Method Manual Lib-A protocol (GS Junior Titanium Series, Roche), enrichment, and pyrosequencing according to the Sequencing Method Manual (Roche). Sequence analysis was performed using CLC Genomics Workbench software.

Table 1. Frequency of CNLOH 17p according to glioma histologic subtype

		CNLOH 17p	
Subtype and histology	***n***	***n***	**%**
Astrocytoma/oligoastrocytoma	256	98	38
Grade II	104	42	40
Grade III	152	56	37
Oligodendroglioma	327	28	8.6
Grade II	102	8	7.8
Grade III	225	20	8.9
Glioblastoma	210	30	14.3

Abbreviations: CNLOH, copy number-neutral loss of heterozygosity.

TP53 Sanger Sequencing

TP53 mutations identified by pyrosequencing were confirmed by direct Sanger sequencing. Tumor DNA was first amplified and purified using the same primers and conditions described for pyrosequencing. Sequencing reactions were performed in both orientations using Big-Dye Terminator Cycle Sequencing Ready Reaction (PerkinElmer, Waltham, MA, http://www.perkinelmer.com). Extension products were purified with the Agencourt CleanSEQ protocol according to the manufacturer's instructions (Beckman-Coulter). Purified sequences were analyzed on an ABI Prism 3730 DNA Analyzer (Applied Biosystems, Foster City, CA, http://www.appliedbiosystems.com). Forward and reverse sequences were systematically analyzed using Chromas Lite software.

Statistical Analysis

We used chi-square and Fisher exact test to compare genotype distribution. The association with continuous variables was calculated with the Mann-Whitney test. Overall survival (OS) was defined as the time between diagnosis and death or last follow-up. Patients who were alive at last follow-up were considered as a censored event in analysis. Progression-free survival (PFS) was defined as the time between diagnosis and recurrence or last follow-up. Patients who were recurrence-free at last follow-up were considered as a censored event in analysis. To find clinical or genomic factors related to OS or PFS, survival curves were calculated according to the Kaplan-Meier method, and differences between curves were assessed using the log-rank test. Variables with a significant p value were used to build a multivariate Cox model. Two-sided p values $< .05$ were considered significant.

Downloaded from http://theoncologist.alphamedpress.org/ by James Studd on October 17, 2016

©AlphaMed Press 2016

Published Ahead of Print on July 8, 2016 as 10.1634/theoncologist.2016-0003.

Table 2. Association of CNLOH 17p with common molecular alterations in gliomas

	Present		Absent		
CNLOH 17p	**Frequency**	**%**	**Frequency**	**%**	***p* value**
EGFR amplification	6/156	3.8	99/637	15.6	<.0001
CDKN2A deletion	23/156	14.7	165/637	25.9	.0032
IDH mutation	114/141	80.9	309/556	55.6	<.0001
1p19q codeletion	1/156	0.6	249/637	39.1	<.0001
MDM2 amplification	0/154	0.0	14/637	2.2	.0173
CDK4 amplification	8/155	5.2	20/637	3.1	NS
TERT promoter mutation	18/74	24.3	159/248	64.1	<.0001
MGMT promoter methylation	17/23	73.9	78/140	55.7	NS
Chr10q loss	29/156	18.6	212/637	57.8	.0003
TP53 mutation	69/71	97.2	20/79	25.3	<.0001

Abbreviations: Chr, chromosome; CNLOH, copy number-neutral loss of heterozygosity; NS, not significant.

Results

We screened the genomic profiles of 793 gliomas (206 grade II, 377 grade III, and 210 grade IV) for the presence of CNLOH 17p. In the whole cohort, we identified 156 cases with CNLOH 17p (19.7%), affecting the whole chromosome 17 in 14 cases (9.0%), the whole short arm of chromosome 17 in 15 cases (9.6%), and only the telomeric portion of 17p in 127 cases (81.4%), including in all cases the *TP53* locus. The mean size of the affected region was 21.6 ± 1.1 Mb (range 7.7–80.9 Mb) (supplemental online Fig. 1A, 1B). We also screened a series of 96 constitutional DNA samples. We did not find any CNLOH 17p in blood DNA, confirming this as a somatic event.

CNLOH 17p affected 50 of 206 grade II (24.3%), 76 of 377 grade III (20.2%), and 30 of 210 grade IV gliomas (14.3%). CNLOH 17p was more frequent in astrocytomas and oligoastrocytomas (98/256, 38%) than oligodendrogliomas (28/327, 8.6%; $p < .0001$) or glioblastoma multiforme (30/210, 14.3%; $p < .0001$) (Table 1).

We investigated the presence of *TP53* mutation by pyrosequencing. Each nonsilent variation was then validated by Sanger sequencing. Of the 71 tumors with CNLOH 17p and available DNA, 97.2% (69/71) were mutated on the *TP53* gene. Electropherograms showed a pattern of homozygous mutation (supplemental online Fig. 2A) in all cases. Missense mutations were the most frequent (58/71, 81.7%), compared with nonsense mutations (8/71, 11.3%) and frameshifts (5/71, 7.0%). Strikingly, one of the two nonmutated tumors had a focal homozygous deletion of *TP53* locus (supplemental online Fig. 3). In all, the *TP53* gene was altered in all but one tumor with CNLOH 17p (70/71, 98.6%). Interestingly, P53 was overexpressed by immunohistochemistry in the remaining nonaltered case, suggesting abnormal P53 sequestration (data not shown).

In non-CNLOH 17p gliomas, *TP53* mutational status was available in 79 tumors. We identified 24 *TP53* mutations (25.3%; $p < .0001$) on 20 tumors, with four tumors having a double variant consisting of 21 (80.8%) missense mutations, four (15.5%) nonsense mutations, and one (3.8%) frameshift. In all these non-CNLOH 17p gliomas, electropherograms showed a heterozygous pattern of *TP53* mutation (supplemental online Fig. 2B). Based on the *TP53* database reported by Edlund et al. [14], we found that 86 of 97 (89%) of these mutations affected the *TP53* DNA binding domain (65/71 in the CNLOH 17p group and 21/26 in the control group; not significant). All mutations are predicted to be transcriptionally inactive.

Table 3. Relative frequency of CNLOH 17p in molecular groups 1, 2, and 3 of grade II–III gliomas

	Present		
CNLOH 17p	**Frequency**	**%**	***p* value**
Group 1 (1p19q codeletion)	1/225	0.44	<.0001
Group 2 (*IDH* mutation without 1p19q codeletion)	85/152	55.92	—
Group 3 (*IDH* wild-type)	7/98	7.14	<.0001

p value determined by Fisher's exact test with group 2.
Abbreviations: —, no data; CNLOH, copy number-neutral loss of heterozygosity.

We next investigated the association of CNLOH 17p with other molecular alterations commonly found in gliomas (Table 2). CNLOH 17p was mutually exclusive with 1p19q codeletion (1/156 vs. 249/556; $p < .0001$) and was associated with *IDH* mutation (114/141 vs. 309/556; $p < .0001$). In grade II and III gliomas, CNLOH 17p was associated with the 1p19q noncodeleted *IDH*-mutated gliomas (group 2) (55.9% of group 2 tumors compared with groups 1 and 3) (Table 3).

We then evaluated the prognostic impact of CNLOH 17p. We did not find any impact on PFS or OS for grade II–IV gliomas with available clinical data (supplemental online Fig. 4). This is not surprising, because CHLOH 17p is strongly associated with the *TP53* mutation, which itself is associated with group 2 gliomas, which have an intermediate prognosis (Fig. 1A). We therefore considered specifically the prognostic impact of CNLOH 17p in group 2 and found an association with a much better outcome (OS 86.3 vs. 46.2 months; $p = .004$) (Fig. 1B). The difference was particularly clear in grade III gliomas (OS >100 vs. 37.9 months; $p = .007$) (Fig. 2) but was not found in grade II and IV gliomas.

We then entered into the Cox model the major histological and biological prognostic markers, i.e., the grading and the molecular subgroup (1p19q codeletion, *IDH* mutation, *IDH*

Downloaded from http://theoncologist.alphamedpress.org/ by James Studd on October 17, 2016

©AlphaMed Press 2016

Published Ahead of Print on July 8, 2016 as 10.1634/theoncologist.2016-0003.

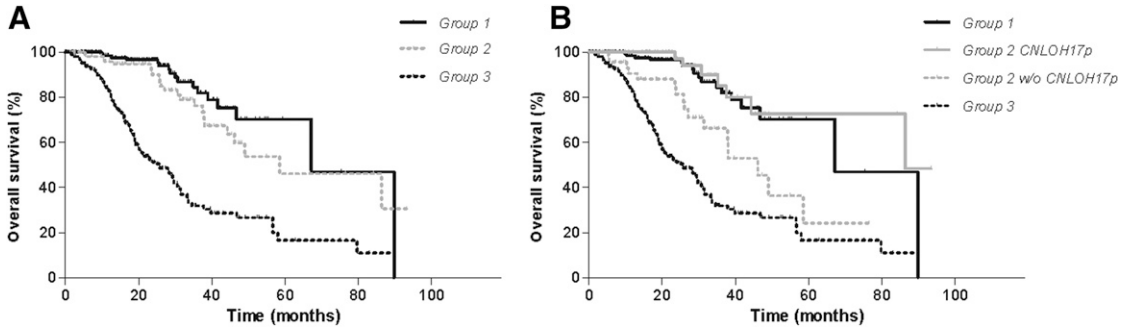


Figure 1. **(A)** Prognostic classification of grade II–IV gliomas according to 1p19q and *IDH* status (groups 1, 2, and 3). **(B)** Prognostic impact of CNLOH 17p in group 2. Survival times were compared using log-rank test (Mantel-Cox). The presence of CNLOH 17 p in group 2 was associated with better outcome (OS 86.3 vs. 46.2 months for group 2 with and without CNLOH 17p, respectively; $p = .004$).

Abbreviations: CNLOH, copy number-neutral loss of heterozygosity; OS, overall survival; w/o, without.

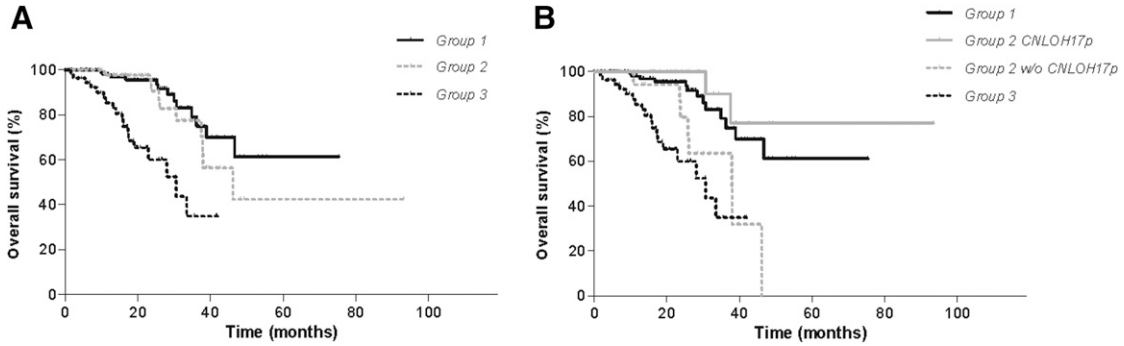


Figure 2. **(A)** Prognostic classification of grade III gliomas according to 1p19q and *IDH* status (groups 1, 2, and 3). **(B)** Prognostic impact of CNLOH 17p in group 2. Survival times were compared using log-rank test (Mantel-Cox). The presence of CNLOH 17 p in group 2 was associated with better outcome (OS >100 vs. 37.9 months for group 2 with and without CNLOH 17p, respectively; $p = .007$).

Abbreviations: CNLOH, copy number-neutral loss of heterozygosity; OS, overall survival; w/o, without.

wild-type): both were strongly predictive of outcome (hazard ratios 2.094 and 1.840, $p = 7 \times 10^{-7}$ and 2×10^{-5}, respectively), but the negative prognostic impact of CNLOH 17p remained significant (hazard ratio 1.641; $p = .04$). Because CNLOH 17p is specifically found in group 2 (IDH-mutated non-codeletion gliomas), we performed multivariate analysis specifically in this group, entering CNLOH 17p, grade, *EGFR* amplification, *CDKN2A* deletion, and *TP53* mutation. We found that CNLOH 17p was the strongest (odds ratio [OR] for non-CNLOH p17 = 3.58) and the most significant ($p = .014$) prognostic marker.

To confirm this result, we analyzed survival data from 142 LGGs from TCGA with *IDH1*/*IDH2* mutations and no 1p19q codeletion. Despite the high rate of censured data, we found that CNLOH 17p, including the *TP53* locus, was associated with better outcome (OR = 0.27; $p = .026$) (supplemental online Fig. 5) [11].

Discussion

Using SNP array, we found that CNLOH 17p is a frequent alteration in gliomas. A similar mechanism has also been reported in other malignancies [15]. Strikingly, CNLOH affects selectively 17p and not (or only marginally) the other chromosome segments, as shown by a recent whole-exome sequencing analysis [2, 16]. We found CNLOH 17p to be almost systematically associated with *TP53* mutation or deletion (70 of 71 samples). The sequence analysis showed a homozygous mutation in all cases, suggesting that during the mechanism of tumorigenesis, the normal arm of chromosome 17p is lost and the altered chromosome arm is duplicated, leading to a homozygous mutation of *TP53* [9, 17–19].

In our series, CNLOH 17p is mutually exclusive with 1p19q codeletion and is associated with *IDH* mutation. Regarding the three molecular subgroups [1–3], CNLOH 17p samples were mostly found in group 2, the 1p19q non-codeleted *IDH*-mutated group, which is associated with *TP53* mutation (85/152 vs. 1/225 in the 1p19q codeleted group and 7/98 in the non-1p19q codeleted, non-*IDH* mutated group).

We therefore analyzed the prognostic impact of CNLOH 17p in this particular subgroup (*IDH* mutated, non-1p19q codeleted). We found that tumors harboring CNLOH 17p had a better OS than tumors without CNLOH 17p and similar to that of 1p19q codeleted tumors (Fig. 2B). The upcoming World Health Organization classification of gliomas will integrate molecular markers; in this setting, the replication of this finding in the independent TCGA series allows generalization of our conclusion; thus we propose CNLOH 17p as a stratification marker in this subgroup defined as molecular astrocytomas [20].

Acknowledgments

Supported by grants from the Ligue Nationale contre le Cancer and the Association pour la Recherche sur les Tumeurs

Downloaded from http://theoncologist.alphamedpress.org/ by James Studd on October 17, 2016

©AlphaMed Press 2016

Published Ahead of Print on July 8, 2016 as 10.1634/theoncologist.2016-0003.

Cérébrales. This work is part of the national program Cartes d'Identité des Tumeurs (http://cit.ligue-cancer.net/) funded and developed by the Ligue Nationale Contre le Cancer. The research leading to these results has received funding from the program "Investissements d'Avenir" (ANR-10-IAIHU-06). The results published here are based in part on data generated by the Cancer Genome Atlas Research Network (http://cancergenome.nih.gov/).

Author Contributions

Conception/Design: Marianne Labussière, Ahmed Idbaih, Marc Sanson

Provision of study material or patients: Marianne Labussière, Amithys Rahimian, Marine Giry, Blandine Boisselier, Marc Polivka, Karima Mokhtari, Agusti Alentorn, Marc Sanson

Collection and/or assembly of data: Marianne Labussière, Amithys Rahimian, Marine Giry, Blandine Boisselier, Yohann Schmitt, Agusti Alentorn, Marc Sanson

Data analysis and interpretation: Marianne Labussière, Amithys Rahimian, Marc Polivka, Karima Mokhtari, Jean-Yves Delattre, Ahmed Idbaih, Karim Labreche, Marc Sanson

Manuscript writing: Marianne Labussière, Jean-Yves Delattre, Ahmed Idbaih, Karim Labreche, Agusti Alentorn, Marc Sanson

Final approval of manuscript: Marianne Labussière, Amithys Rahimian, Marine Giry, Blandine Boisselier, Yohann Schmitt, Marc Polivka, Karima Mokhtari, Jean-Yves Delattre, Ahmed Idbaih, Karim Labreche, Marc Sanson

Disclosures

Marc Sanson: Roche (C/A). The other authors indicated no financial relationships.

(C/A) Consulting/advisory relationship; (RF) Research funding; (E) Employment; (ET) Expert testimony; (H) Honoraria received; (OI) Ownership interests; (IP) Intellectual property rights/inventor/patent holder; (SAB) Scientific advisory board

References

1. Labussière M, Idbaih A, Wang XW et al. All the 1p19q codeleted gliomas are mutated on IDH1 or IDH2. Neurology 2010;74:1886–1890.

2. Suzuki H, Aoki K, Chiba K et al. Mutational landscape and clonal architecture in grade II and III gliomas. Nat Genet 2015;47:458–468.

3. Brat DJ, Verhaak RG, Aldape KD et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med 2015;372:2481–2498.

4. Liu XY, Gerges N, Korshunov A et al. Frequent ATRX mutations and loss of expression in adult diffuse astrocytic tumors carrying IDH1/IDH2 and TP53 mutations. Acta Neuropathol 2012;124:615–625.

5. Idbaih A, Ducray F, Dehais C et al. SNP array analysis reveals novel genomic abnormalities including copy neutral loss of heterozygosity in anaplastic oligodendrogliomas. PLoS One 2012;7:e45950.

6. Kotliarov Y, Kotliarova S, Charong N et al. Correlation analysis between single-nucleotide polymorphism and expression arrays in gliomas identifies potentially relevant target genes. Cancer Res 2009;69:1596–1603.

7. Kuga D, Mizoguchi M, Guan Y et al. Prevalence of copy-number neutral LOH in glioblastomas revealed by genomewide analysis of laser-microdissected tissues. Neuro-oncol 2008;10:995–1003.

8. Lo KC, Bailey D, Burkhardt T et al. Comprehensive analysis of loss of heterozygosity events in glioblastoma using the 100K SNP mapping arrays and comparison with copy number abnormalities defined by BAC array comparative genomic hybridization. Genes Chromosomes Cancer 2008;47:221–237.

9. Yin D, Ogawa S, Kawamata N et al. High-resolution genomic copy number profiling of glioblastoma multiforme by single nucleotide polymorphism DNA microarray. Mol Cancer Res 2009;7:665–677.

10. Gonzalez-Aguilar A, Idbaih A, Boisselier B et al. Recurrent mutations of MYD88 and TBL1XR1 in primary central nervous system lymphomas. Clin Cancer Res 2012;18:5203–5211.

11. Wang K, Li M, Hadley D et al. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 2007;17:1665–1674.

12. Van Loo P, Nordgard SH, Lingjærde OC et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci USA 2010;107:16910–16915.

13. Labussière M, Di Stefano AL, Gleize V et al. TERT promoter mutations in gliomas, genetic associations and clinico-pathological correlations. Br J Cancer 2014;111:2024–2032.

14. Edlund K, Larsson O, Ameur A et al. Data-driven unbiased curation of the TP53 tumor suppressor gene mutation database and validation by ultradeep sequencing of human tumors. Proc Natl Acad Sci USA 2012;109:9551–9556.

15. O'Keefe C, McDevitt MA, Maciejewski JP. Copy neutral loss of heterozygosity: A novel chromosomal lesion in myeloid malignancies. Blood 2010;115:2731–2739.

16. Bai H, Harmanci AS, Erson-Omay EZ et al. Integrated genomic characterization of IDH1-mutant glioma malignant progression. Nat Genet 2016;48:59–66.

17. Heinrichs S, Li C, Look AT. SNP array analysis in hematologic malignancies: Avoiding false discoveries. Blood 2010;115:4157–4161.

18. Jasek M, Gondek LP, Bejanyan N et al. TP53 mutations in myeloid malignancies are either homozygous or hemizygous due to copy number-neutral loss of heterozygosity or deletion of 17p. Leukemia 2010;24:216–219.

19. Tuna M, Knuutila S, Mills GB. Uniparental disomy in cancer. Trends Mol Med 2009;15:120–128.

20. Sahm F, Reuss D, Koelsche C et al. Farewell to oligoastrocytoma: In situ molecular genetics favor classification as either oligodendroglioma or astrocytoma. Acta Neuropathol 2014;128:551–559.

See http://www.TheOncologist.com for supplemental material available online.

Downloaded from http://theoncologist.alphamedpress.org/ by James Studd on October 17, 2016

©AlphaMed Press 2016

nature COMMUNICATIONS

ARTICLE

Received 22 Feb 2015 | Accepted 17 Apr 2015 | Published 12 Jun 2015

DOI: 10.1038/ncomms8207 OPEN

TCF12 is mutated in anaplastic oligodendroglioma

Karim Labreche[1,2,3,4,5,*], Iva Simeonova[2,3,4,5,*], Aurélie Kamoun[6,*], Vincent Gleize[2,3,4,5,*], Daniel Chubb[1], Eric Letouzé[6], Yasser Riazalhosseini[7,8], Sara E. Dobbins[1], Nabila Elarouci[6], Francois Ducray[9], Aurélien de Reyniès[6], Diana Zelenika[10], Christopher P. Wardell[11], Mathew Frampton[1], Olivier Saulnier[2,3,4,5], Tomi Pastinen[7,8], Sabrina Hallout[2,3,4], Dominique Figarella-Branger[12,13], Caroline Dehais[14], Ahmed Idbaih[2,3,4,5,14], Karima Mokhtari[2,3,4,15], Jean-Yves Delattre[2,3,4,5,14,**], Emmanuelle Huillard[2,3,4,5,**], G. Mark Lathrop[7,8,**], Marc Sanson[2,3,4,5,14,**], Richard S. Houlston[1,**] & POLA Network[†]

Anaplastic oligodendroglioma (AO) are rare primary brain tumours that are generally incurable, with heterogeneous prognosis and few treatment targets identified. Most oligodendrogliomas have chromosomes 1p/19q co-deletion and an *IDH* mutation. Here we analysed 51 AO by whole-exome sequencing, identifying previously reported frequent somatic mutations in *CIC* and *FUBP1*. We also identified recurrent mutations in *TCF12* and in an additional series of 83 AO. Overall, 7.5% of AO are mutated for *TCF12*, which encodes an oligodendrocyte-related transcription factor. Eighty percent of *TCF12* mutations identified were in either the bHLH domain, which is important for TCF12 function as a transcription factor, or were frameshift mutations leading to TCF12 truncated for this domain. We show that these mutations compromise *TCF12* transcriptional activity and are associated with a more aggressive tumour type. Our analysis provides further insights into the unique and shared pathways driving AO.

[1] Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK. [2] Inserm, U 1127, ICM, F-75013 Paris, France. [3] CNRS, UMR 7225, ICM, F-75013 Paris, France. [4] Institut du Cerveau et de la Moelle épinière ICM, Paris 75013, France. [5] Sorbonne Universités, UPMC Université Paris 06, UMR S 1127, F-75013 Paris, France. [6] Programme Cartes d'Identité des Tumeurs (CIT), Ligue Nationale Contre Le Cancer, 75013 Paris, France. [7] Department of Human Genetics, McGill University, Montreal, Quebec, Canada H3A 0G1. [8] McGill University and Genome Quebec Innovation Centre, Montreal, Quebec, Canada H3A 0G1. [9] INSERM U1028, CNRS UMR5292, Service de Neuro-oncologie, Hopital neurologique, Hospices civils de Lyon, Lyon Neuroscience Research Center, Neuro-Oncology and Neuro-Inflammation Team, 69677 Lyon, France. [10] Centre National de Génotypage, IG/CEA, 2 rue Gaston Crémieux, CP 5721, Evry 91057, France. [11] Division of Molecular Pathology, The Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK. [12] AP-HM, Hôpital de la Timone, Service d'anatomie pathologique et de neuropathologie, 13385 Marseille, France. [13] Université de la Méditerranée, Aix-Marseille, Faculté de Médecine La Timone, CRO2, UMR 911 Marseille, France. [14] AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de neurologie 2-Mazarin, 75013 Paris, France. [15] AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Laboratoire de Neuropathologie R. Escourolle, 75013 Paris, France. * These authors contributed equally to this work. ** These authors jointly supervised this work. † A full list of consortium members appears at the end of the paper. Correspondence and requests for materials should be addressed to R.S.H. (email: richard.houlston@icr.ac.uk).

© 2015 Macmillan Publishers Limited. All rights reserved.

Anaplastic oligodendrogliomas (AO; World Health Organization grade III oligodendrogliomas) are rare primary malignant brain tumours with a highly variable overall prognosis. The emblematic molecular alteration in oligodendrogliomas is 1p/19q co-deletion, which is associated with a better prognosis and response to early chemotherapy with procarbazine, lomustine and vincristine[1–3]. Recent high-throughput sequencing approaches have identified *IDH* (*IDH1* and *IDH2*), *CIC*, *FUBP1* and *TERT* promoter mutations in oligodendroglioma (75, 50, 10 and 75%, respectively)[2,4,5], *IDH* mutation status typically being associated with a better clinical outcome[6]. Identifying additional driver genes and altered pathways in oligodendroglioma offers the prospect of developing more effective therapies and biomarkers to predict individual patient outcome.

Here we perform whole-exome and transcriptome sequencing of AO to search for additional tumour driver mutations and pathways disrupted. In addition to previously reported recurrently mutated genes, we report the identification of somatic mutations in *TCF12* in AO. These mutations compromise *TCF12* transcriptional activity and confer a more aggressive AO phenotype.

Results

In accordance with conventional clinical practice, we considered three molecular subtypes for our analyses: (i) *IDH*-mutated 1p/19q co-deleted (*IDH*mut-codel); (ii) *IDH*-mutated 1p/19q non-co-deleted (*IDH*mut-non-codel) and (iii) *IDH*-wild type (*IDH*wt)[7]. Assignment of *IDH*-mutated (defined by *IDH1* R132 or *IDH2* R172 mutations), 1p/19q and *TERT* promoter mutation (defined by C228T or C250T) status in tumours was determined using conventional sequencing and single-nucleotide polymorphism (SNP) array methods.

Mutational landscape. We performed whole-exome sequencing of 51 AO tumours (Supplementary Data 1) and matched germline DNA, targeting 318,362 exons from 18,901 genes. The mean sequencing coverage across targeted bases was 57 ×, with 80% of target bases above 20 × coverage (Supplementary Fig. 1). We identified a total of 4,733 mutations (with a mean of 37 non-silent mutations per sample) equating to a mean somatic mutation rate of 1.62 mutations per megabase (Mb) (Fig. 1). Although the tumours of two patients (3,063 and 3,149) had high rates of mutation (9.1 and 12.4, respectively), this was not reflective of tumour site (both frontal lesions as were 68% of the whole series) or treatment. Excluding these two cases the mean rate of non-silent mutations per tumour was 33 ± 14, which is similar to the number found in most common adult brain tumours. The mutation spectrum in AO tumours was characterized by a predominance of C > T transitions, as observed in most solid cancers (Fig. 1)[8,9]. While few of the tumours were *IDHwt*, these did not harbour a significantly higher number of mutations compared with *IDHmut-1p/19q* co-deleted and *IDHmut-non-1p/19q* co-deleted tumours (Fig. 1). Intriguingly, one tumour (2,688) was co-mutated for *IDH1* (R132H) and *IDH2* (P162S), but exhibited no distinguishing phenotype in terms of clinicopathology or mutation rate.

We used MutSigCV version 1.4 (ref. 8) to identify genes harbouring more non-synonymous mutations than expected by chance given gene size, sequence context and mutation rate of each tumour for the three molecular subtypes, respectively. As expected, we observed frequent mutations of the tumour suppressors *FUBP1* (22%) located on 1p, and *CIC* (32%) located on 19q, which have been reported in the context of 1p/19q co-deletion (Fig. 1; Supplementary Fig. 2); these were not mutually exclusive events (Fig. 1). Also within the *IDH*mut-codel group, 37 of tumours tested carried *TERT* C228T or C250T promoter mutations (72%), none of which also carried an *ATRX* mutation, concordant with the previously reported finding that these are mutually exclusive events[2].

In addition to the mutation of *IDH1* (78%), *IDH2* (17%), *CIC* (32%) and *FUBP1* (22%), *TCF12* was also significantly mutated (Q-value < 0.1; Fig. 1; Supplementary Table 2). Heterozygous somatic mutations in *TCF12*, which encodes the basic helix–loop–helix (bHLH) transcription factor 12 (aliases *HEB*, *HTF4* and *ALF1*) were identified in five (1 missense, R602M; 2 splice-site, c.825 + 5G > T, c.1979-3_1979-delTA and 2 frameshift, E548fs*13, S682fs*14) of the 46 *IDH*-mutated 1p/19q co-deleted. Intriguingly, germline mutations of residues E548 and R602 have been previously shown to cause coronal craniosynostosis[10].

The availability of high-quality tumour material allowed us to generate SNP array and expression data on 31 of the cases exome sequenced. In addition to co-deletion of chromosome arms 1p/19q, we identified several other recurrent genomic alterations—mainly loses of chromosomes 4 (29%), 9p (28%) and 14q (19%); Supplementary Fig. 3; Supplementary Table 1). Notably, tumours featuring mutation of Notch-pathway genes showed significant chromosome 4 loss ($P = 0.02$, χ^2-test). To identify fusion transcripts, we analysed RNA-sequencing (RNA-seq) data, which was available for 36 of the 51 tumours. After filtering, the only chimeric transcript identified was the predicted driver *FGFR3–TACC3* fusion, previously described in *IDH* wild-type gliomas[11–13], which was seen in two of the *IDH*wt-non-1p/19q co-deleted tumours—patients 2463 and 2441; Of note was that patient 2463 carried an *IDH2* intron-5 mutation (c.679-28C > T).

Incorporation of TCGA mutation data. To explore the mutational spectra of AO in an independent series, we made use of data generated by The Cancer Genome Atlas (TCGA) study of low-grade glioma, which provides exome sequencing data on a further 43 AO tumours. Two of these 43 tumours harboured frameshift mutations in *TCF12* (E548R and D171fs) (Supplementary Table 2). As with our series, these *TCF12* mutations were exclusive to *IDH*-1p/19q co-deleted tumours. In a combined analysis, mutations in *PI3KCA*, *NOTCH1* and *TP53* were significantly overrepresented when analysed using MutSigCV (Q-value < 0.1; Supplementary Table 2). In addition, mutation of *ATRX* and *RBPJ* were of borderline significance.

A bias towards variants with functional impact (FM) is a feature of cancer drivers[14]. To increase our ability to identify cancer drivers and delineate associated oncogenic pathways for AO, we incorporated mutation data from multiple tumour types using Oncodrive-fm[14] implemented within the IntOGen-mutations platform[15] (Fig. 2). The most recurrently mutated genes according to MutSig were also detected by Oncodrive-fm as significantly mutated (Q-value < 0.05). Oncodrive-fm also identified a number of other important mutated genes (that is, displaying high FM bias) including *SETD2*, *NOTCH2*, *RBPJ*, *ARID1A*, *ARID1B*, *HDAC2* and *SMARCA4* (Fig. 2).

Using all mutation results, we performed an analysis to identify pathways or gene ontologies that were significantly enriched in mutated genes. As expected, the most significantly altered pathways were linked to the tricarboxylic acid cycle and isocitrate metabolic process as a consequence of *IDH* mutation. Consistent with the other genes that were found significantly mutated by MutSigCV and Oncodrive-fm analysis, the Notch signalling pathway ($P = 1.0 \times 10^{-5}$, binomial test), genes involved in neuron differentiation ($P = 2.0 \times 10^{-5}$, binomial test) and genes involved in chromatin organization ($P = 0.02$, binomial test) were also significantly enriched for mutations (Supplementary Data 3).

© 2015 Macmillan Publishers Limited. All rights reserved.

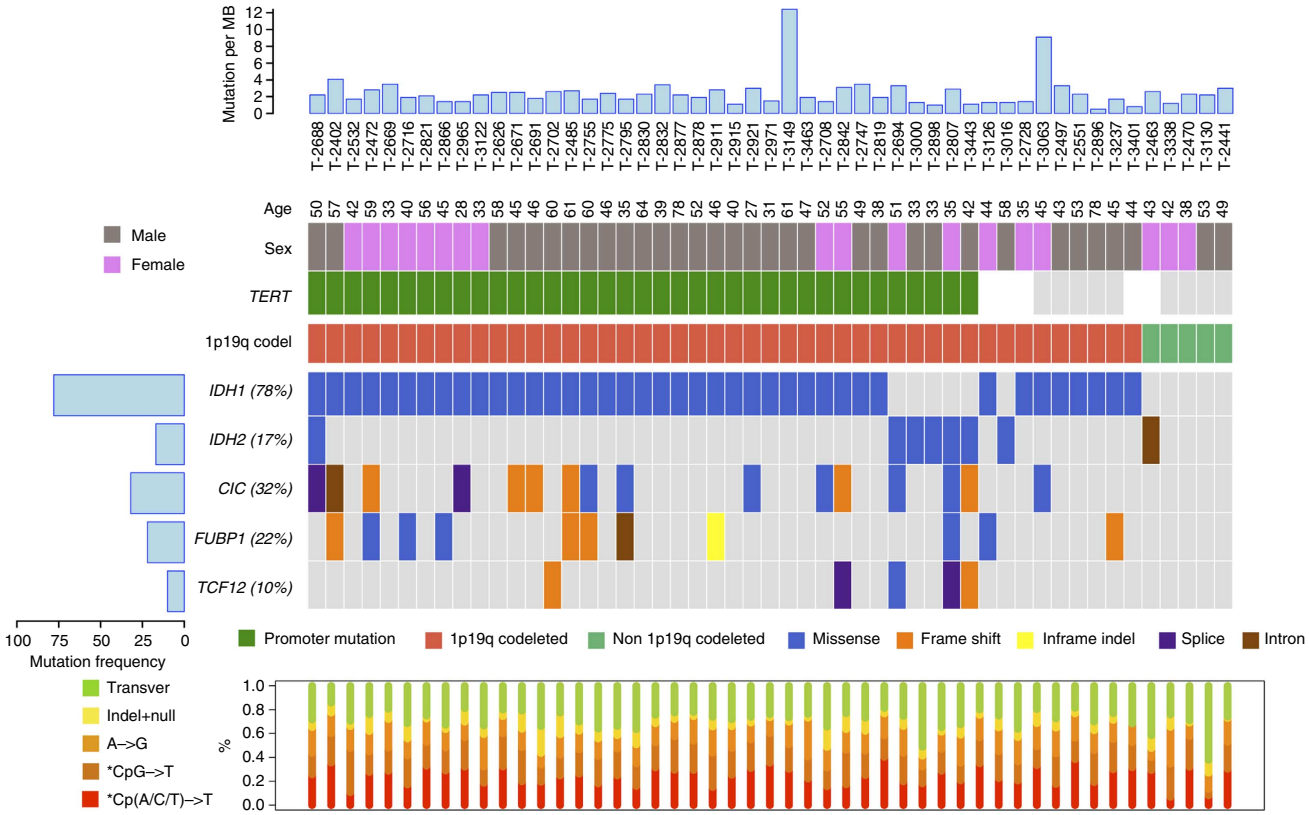


Figure 1 | Significantly mutated genes in anaplastic oligodendroglioma by molecular subtype. Significantly mutated genes (Q-value < 0.1) identified by exome sequencing are listed by Q-value. The percentage of AO samples with mutation detected by automated calling is detailed on the left. Samples are displayed as columns, with the mutation rate plotted at the top. Samples are arranged to emphasize mutual exclusivity. Mutation types are indicated in different colours (see legend). White colour indicates no information available. Also shown is the relative proportion of base-pair substitutions within mutation categories for each tumour.

Validation of *TCF12* in an additional series of AO. To identify additional *TCF12*-mutated AO tumours, we conducted targeted sequencing of a further 83 AO. Five tumours harboured *TCF12* mutations—G48fs*38, M260fs*5, R326S, D455fs*59 and delN606 (Supplementary Data 1). On the basis of our combined sample of 134 tumours, the mutation frequency of *TCF12* in AO is 7.5% (95% confidence interval 3.6–13.2%). No significant difference in patient survival in 1p/19q co-deleted AO was associated with *TCF12* mutation in 69 patients (Supplementary Fig. 4). While our power to demonstrate a statistically significant relationship was limited (that is, ~40% for a hazard ratio of 2.0, stipulating $P = 0.05$), we noted that patients having either *TCF12* mutated or *TCF12* loss of heterozygosity (LOH) tended to be associated with shorter survival (Supplementary Fig. 4). To gain further insight into the role of *TCF12* mutation in oligodendroglioma, we sequenced 75 grade II tumours identifying one mutation carrier (P212fs*31; Supplementary Data 1). The observation that the frequency of *TCF12* mutations is higher in AO as compared with grade II tumours ($P = 0.049$, χ^2-test) is compatible with *TCF12* participating in the generation of a more aggressive phenotype.

***TCF12* bHLH mutants compromised transactivation**. To explore the functional consequences of *TCF12* mutation, we tested the transcriptional activity of several mutants (Fig. 3). We tested the frameshift mutations M260fs*5 and E548fs*13, which in the germline cause coronal craniosynostosis[10] and S682fs*14, since introduction of a C-terminal premature stop codon may result in escape from non-sense-mediated decay. We also tested the missense mutation R602M, which is predicted to destabilize the bHLH domain required for DNA binding and dimerization (Fig. 3) and whose adjacent residue (R603) has been found recurrently mutated in colon cancer[16]. Finally, we tested the missense mutation R326S, since mutations of adjacent G327 have been reported in lung adenocarcinoma[17]. The frameshift mutants M260fs*5 and E548fs*13 completely abolished TCF12 transactivation, consistent with the lack of bHLH DNA-binding domain (Fig. 3). R602M retained only 34% of WT transcriptional activity ($P = 0.0018$, Student's t-test; Fig. 3). We did not observe significant modulation of transactivation for the R326S and S682fs*14 mutants, although the latter consistently showed decreased activity (Fig. 3).

Downregulation of pathways in *TCF12* bHLH mutants. We profiled gene expression in 8 *TCF12*-mutated and 45 wild-type tumours within 1p/19q co-deleted samples (Supplementary Table 1). *TCF12* mutation was associated with significant enrichment of immune response pathways (Supplementary Data 4). Restricting the analysis to tumours with the *TCF12*-altered bHLH domain ($n = 6$), we found downregulation of pathways featuring known partners of TCF12, such as TCF21, EZH2 and BMI1 (ref. 18) (Supplementary Table 2). Interestingly, we found decreased activity of genes sets related to E-cadherin (*CDH1*), which is a TCF12 target gene associated with tumour phenotype[18]. Since the promotor sequences of *CDH1* and *BMI1* feature E-box motifs and are modulated by the bHLH binding[19,20], this provides a mechanistic basis for change in gene expression associated with mutant *TCF12*.

© 2015 Macmillan Publishers Limited. All rights reserved.

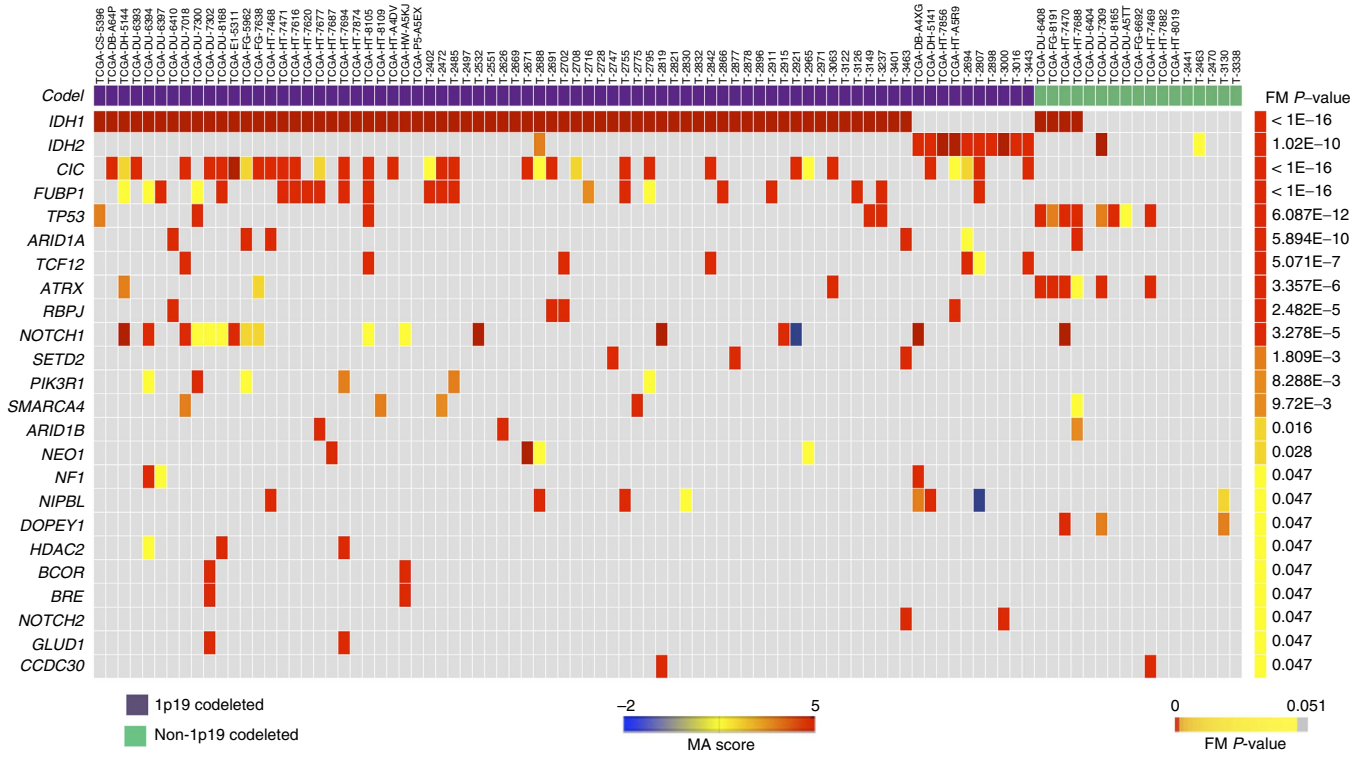


Figure 2 | FM-biased genes and gene modules in AO identified by Oncodrive-fm using data from this study and tumours profiled by TCGA. Heatmap shows tumours in columns and genes in rows, the colour reflecting the MutationAssessor (MA) scores of somatic mutations. FM ext. qv, corrected *P* values of the FM bias analysis using the external null distribution.

Mutant TCF12 proteins show subcellular localization changes. We evaluated TCF12 expression and subcellular localization for all of our 11 *TCF12*-mutated tumours (10 AO and 1 oligodendroglioma grade II) and 11 *TCF12* wild-type tumours by immunohistochemistry. All *TCF12* wild-type tumours showed nuclear expression in a heterogeneous cell population (Fig. 4; Supplementary Fig. 5), whereas several *TCF12*-mutated tumours showed nuclear and cytoplasmic staining (Fig. 4; Supplementary Fig. 5). Interestingly, mutations abolishing transcriptional activity were associated with increased staining, suggesting inactive mutant protein accumulation.

TCF12 mutations associate with aggressive tumour phenotype. We profiled the extent of necrosis, microvascular proliferation and the mitotic index available for *TCF12* wild-type or mutated tumours. A significant increase in palisading necrosis (Fig. 5) as well as a trend towards a higher mitotic index was associated with *TCF12* mutation, consistent with a more aggressive phenotype (Fig. 5). Intriguingly, tumours harbouring disruptive bHLH domain mutations exhibited the highest proportion of palisading necrosis and mitotic figures.

Discussion

Our genome sequencing of AO has confirmed the mutually exclusive mutational profile in *IDH*mut-1p/19q co-deleted and *IDH*mut non-1p/19q co-deleted tumour subtypes, which reflect distinct molecular mechanisms of oncogenesis—consistent with the requirement for either 1p/19q co-deletion or *TP53* mutation post *IDH* mutation. Moreover, as previously proposed, the genomic abnormalities in *IDH*mut-1p/19p co-deleted tumours are consistent with one common mechanism of tumour initiation being through 1p/19q loss, mutation of *IDH1* or *IDH2* and *TERT* activation through promoter mutation[2], which in turn predisposes to deactivation of *CIC*, *FUBP1*, *NOTCH* and activating mutations/amplifications in the PI3K pathway.

We identified and replicated mutations in *TCF12*, a bHLH transcription factor that mediates transcription by forming homo- or heterodimers with other bHLH transcription factors. Tcf12 is highly expressed in neural progenitor cells during neural development[21] and in cells of the oligodendrocyte lineage[22].

We found that mutations generating truncated TCF12 lacking the bHLH DNA-binding domain abrogate the transcriptional activity of TCF12. In addition, single residue substitutions such as R602M within the bHLH domain also dramatically reduce TCF12 transcriptional ability. Finally, we found that the loss of TCF12 transcriptional activity was associated with a more aggressive tumour phenotype. Although speculative, our expression data provides evidence that the effects of *TCF12* mutation on AO development may be mediated in part through E-cadherin related pathway. Indeed, this was one of the pathways down-regulated in mutated tumours and intriguingly *CDH1* has been implicated in metastatic behaviour in a number of cancers[18,23]. It is likely that some *TCF12* mutations may have subtle effects on bHLH function or act through independent pathways. Irrespective of the downstream effects of *TCF12* mutation on glioma, our data are compatible with *TCF12* having haploinsufficient tumour suppressor function. *TCF12* haploinsufficiency has previously been reported in patients with coronal craniosynostosis and in their unaffected relatives[10]. Strikingly, 3 of the 11 mutations we identified in AO, which concern residues M260, E548 and R602, cause coronal craniosynostosis[10,24]. Although speculative, collectively these data raise the possibility that carriers of germline *TCF12* mutations may be at an increased risk of developing AO.

To our knowledge, this study represents the largest sequencing study of AO conducted to date. However, given the number of

© 2015 Macmillan Publishers Limited. All rights reserved.

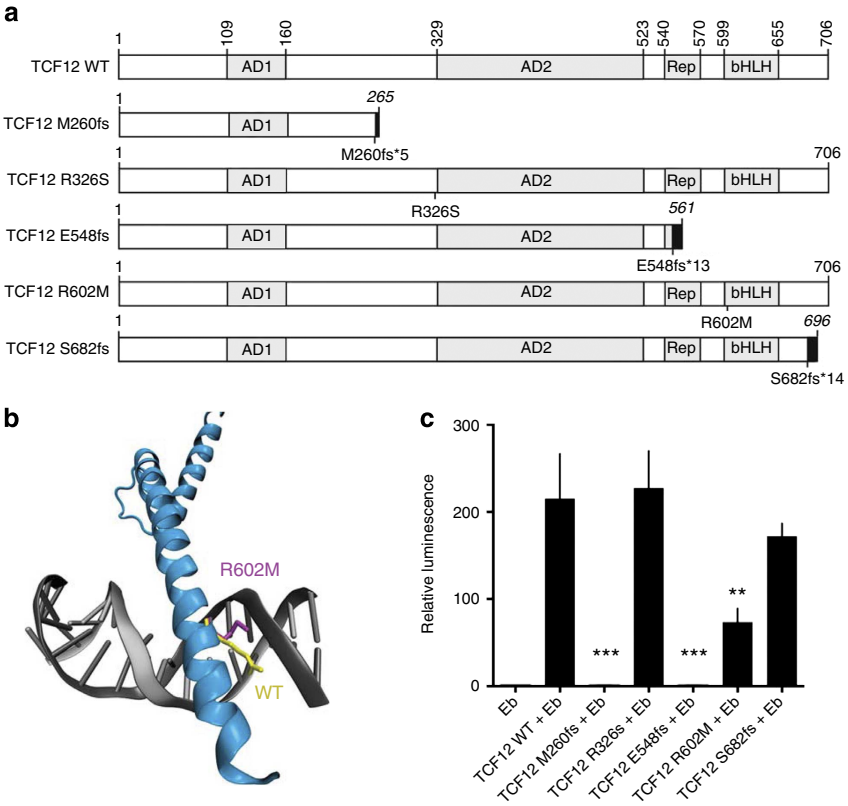


Figure 3 | *TCF12* mutations altering the bHLH domain result in impaired transactivation. (**a**) Schematic view of the wild-type and mutant TCF12 proteins for which the transactivation capacity has been assessed. Upper panel: wild-type human TCF12, functional domains in grey—activation domain 1 (AD1), activation domain 2 (AD2), repressor domain (Rep) and bHLH domain (bHLH). Lower panel: resulting truncated proteins. Black boxes indicate non-related amino-acid sequences resulting from frameshift mutations (fs), and truncated proteins size is in italic. (**b**) Schematic structure of the bHLH domain of TCF12 (blue) bound to DNA (grey). WT R602 (yellow) and mutant M602 (purple) residues are indicated. (**c**) E-box-luciferase reporter plasmid (Eb) was transfected alone or in combination with TCF12 wild-type or mutant expression plasmids. Both frameshift mutants that lack the bHLH DNA binding domain completely abolish TCF12 transcriptional activity. All samples were run in triplicate in four independent experiments. Data were normalized to control renilla luciferase. Values are mean ± s.d. $^{***}P = 0.0002$, $^{**}P = 0.0018$ (Student's *t*-test).

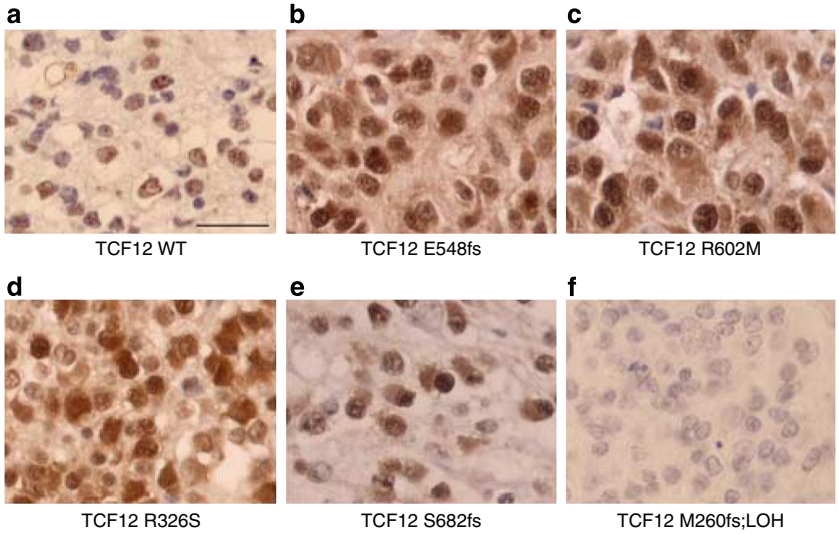


Figure 4 | TCF12 is highly expressed in a subset of anaplastic oligodendroglioma. Representative TCF12 immunostainings are shown: (**a**) wild-type TCF12 tumours show nuclear staining in a heterogeneous cell population. (**b**–**e**) Mutant TCF12 tumours show strong nuclear and cytoplasmic staining. (**f**) Mutant M260fs (resulting in a truncated protein) is associated with 15q21.3 LOH and shows no staining. Scale bar, 50 μm.

tumour-normal pairs we have analysed and the mutational frequency in AO, we were only well powered to identify genes that have a high-frequency mutations (that is, >10%). Hence further insights into the biology of AO should be forthcoming through additional sequencing initiatives and meta-analyses of these data.

© 2015 Macmillan Publishers Limited. All rights reserved.

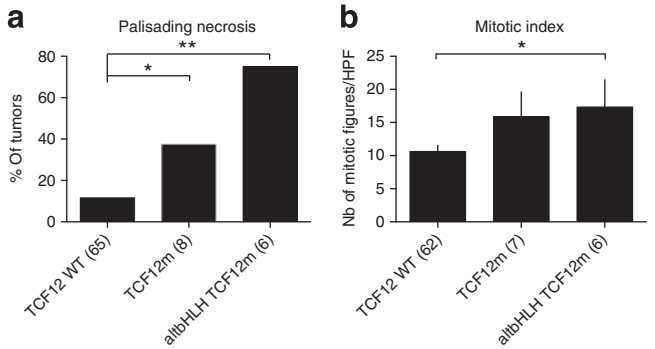


Figure 5 | *TCF12* mutation correlates with a higher necrotic and mitotic index. (**a**) Percentage of palisading necrosis in tumours with wild-type *TCF12*, all tumours mutated for *TCF12* or only altered bHLH *TCF12* mutants; $^{*}P=0.02$, $^{**}P=0.004$. (**b**) Mitotic index in *TCF12* wild-type, *TCF12*-mutated and altered bHLH *TCF12* mutants; $^{*}P=0.039$, mean ± s.e.m. CN, copy number; LOH, loss of heterozygosity; HPF, high-power field. The number of samples is indicated in parenthesis.

Methods

Patient samples and consent. Samples were obtained with informed and written consent and the study was approved by Comité de Protection des Personnes Ile de France-VI (October 2008) of respective hospitals participating in the Prise en charge des oligodendrogliomes anaplasiques (POLA) network. All patients were aged 18 years or older at diagnosis, and tumour histology was centrally reviewed and validated according to World Health Organization (WHO) guidelines[25]. Exome sequencing was conducted on samples from 51 AO patients (33 male; median age 49 years at diagnosis, range 27–81). For targeted follow-up analyses, we studied the tumours from an additional 83 AO patients and 75 patients with grade II tumours. A summary of each of the tumour cohorts and respective pathological information on the patients is provided in Supplementary Table 1.

DNA and RNA extraction. Germline DNA was extracted from EDTA-venous blood samples using QIAquick PCR Purification Kits (Qiagen Ltd). Tumour DNA was extracted from snap-frozen tumour samples using the iPrep ChargeSwitchH Forensic Kit, according to manufacturer's recommendations. DNAs were quantified and qualified using a NanoVue Plus spectrophotometer (GE Healthcare Life Sciences) and gel electrophoresis. RNA was extracted from tumours lysed by Lysing Matrix D tube and FastPrep instrument (MP Biomedicals) using the iPrep Trizol Plus RNA Kit (Life Technologies). Stringent criteria for RNA quality were applied to rule out degradation, specifically a 28S/18S ratio >1.8.

SNP array analysis. In total, 115 samples from tumours were genotyped using Illumina SNP microarrays: 32 samples with Illumina 370-Duo 1.0 BeadChips, 31 with Human610-Quad, 46 with HumanOmniexpress-12V1 and 6 with HumanCore-12v1. Raw fluorescent signals were imported into BeadStudio software (Illumina) and normalized to obtain log R ratio and B-allele frequency (BAF) values. The tQN normalization procedure was then applied to correct for asymmetry in BAF signals due to bias between the two dyes used in Illumina assays. Genomic profiles were divided into homogeneous segments by applying the circular binary segmentation algorithm to both log R ratio and BAF values. We then used the Genome Alteration Print method to determine the ploidy of each sample, the level of contamination with normal cells and the allele-specific copy number of each segment. Chromosome aberrations were defined using empirically determined thresholds as follows: gain, copy number $\geq$ploidy $+1$; loss, copy number $\leq$ploidy -1; high-level amplification, copy number $>$ploidy $+2$; homozygous deletion, copy number $=0$. Finally, we considered a segment to have undergone LOH when the copy number of the minor allele was equal to 0. Lists of homozygous deletions and focal amplifications, defined by at least five consecutive probes, were generated and verified manually to remove doubtful events. Significantly recurrent copy number changes were identified using the GISTIC2.0 algorithm[26].

***TERT* promoter mutation sequencing.** Characterized mutations in the *TERT* promoter, C228T and C250T variants with G>A nucleotide substitutions at genomic positions 1,295,228 bp and 1,295,250 bp (hg19), respectively, were obtained by Sanger sequencing. Primer sequences were: TERT-F—5′-GGCCGA TTCGACCTCTCT-3′ and TERT-R 5′-AGCACCTCGCGGTAGTGG-3′.

Whole-exome sequencing. DNA was quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies). Libraries were generated robotically using the SureSelectXT Automated Human All Exon Target Enrichment for Illumina Paired-End Multiplexed Sequencing (Agilent) as per the manufacturer's recommendations. Libraries were quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies) and the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (D-Mark). Average size of the fragment was determined using a LaChip GX (PerkinElmer) instrument. Sequencing was performed by pooling four libraries per lane at a 9-pM dilution on an Illumina HiSeq 2,000 instrument for 2×100 cycles using the recommended manufacturer's conditions. PhiX control was added at 1% on each lane. BCL2FASTQ (Illumina) was used to convert bcl files to fastqs (v 1.8.4). Coverage statistics are summarized in Supplementary Fig. 1. Paired-end fastq files were extracted using Illumina CASAVA software (v.1.8.1, Illumina) and aligned to build 37 (hg19) of the human reference genome using Stampy and Burrows–Wheeler Aligner[27], and PCR duplicates were removed with PicardTools 1.5. We assessed coverage of consensus coding sequence bases using Genome Analysis Toolkit[28] v2.4-9. Somatic single-nucleotide variants were called using MuTect[29] and the Genome Analysis Toolkit v2.4-9, and indels using IndelGenotyper. We excluded potential Covaris-induced mutations as per Costello *et al.*[30] using in-house scripts. Confirmation of selected single-nucleotide variants including *TCF12*, *CIC*, *FUBP1*, *SYNE1*, *FAT1*, *SETD2*, *RBPJ*, *NOTCH1*, *IDH1* and *IDH2* was performed by Sanger sequencing implemented on ABI 3,300 ×l platforms (Applied Biosystems, Foster City, USA). Primer sequences are detailed in Supplementary Data 5. In all cases, Sanger sequencing was 100% concordant with next-generation sequencing.

We used MutSigCV[8] version 1.4 to identify genes harbouring more non-synonymous mutations than expected by chance, given gene size, sequence context and the mutation rate. We used as genomic covariates the mean expression level of each gene in our AO expression data set, the DNA replication time and the HiC statistic of chromatin state available in MutSig reference files. To increase our ability to identify cancer drivers and delineate associated oncogenic pathways for AO, we incorporated mutation data from multiple tumour types using Oncodrive-fm[14] implemented within the IntOGen-mutations platform[15].

Transcriptome sequencing. Extracted RNA was cleaned using the RNeasy MinElute Cleanup Kit (Qiagen) and the RNA integrity assessed using an Agilent 2,100 Bioanalyzer and quantified using a Nanodrop 1,000. Libraries for stranded total RNA-seq were prepared using the Illumina Stranded Total RNA protocol (RS-122-2301). Libraries were assessed by the Agilent 2,100 Bioanalyzer. Sequencing was performed by pooling four libraries per lane at a 9-pM dilution on an Illumina HiSeq 2,000 instrument for 2×100 cycles using the recommended manufacturer's conditions. PhiX control was added at 1% on each lane. BCL2FASTQ was used to convert bcl files to fastqs (v 1.8.4). Paired-end reads from RNA-seq were aligned to the following database files using Burrows–Wheeler Aligner 0.5.5: (i) the human GRCh37-lite reference sequence, (ii) RefSeq, (iii) a sequence file representing all possible combinations of non-sequential pairs in RefSeq exons and (iv) the AceView database flat file downloaded from UCSC, representing transcripts constructed from human expressed sequence tag (ESTs). The mapping results from databases (ii)-(iv) were aligned to human reference genome coordinates. The final BAM file was constructed by selecting the best alignment. To identify fusion transcripts, we analysed RNA-seq data using Chimerascan software[31] (version 0.4.5). As advocated, algorithmic output was analysed for high-confidence fusion transcripts imposing filters: (i) spanning reads >2 (ii) total supported reads ≥10 (ref. 32). In absence of corresponding paired normal tissue samples, we made use of data from the human body map project data to identify fusions seen in normal tissue.

***TCF12* sequencing in the validation series.** PCR amplification of 21 amplicons covering each exon of *TCF12* on DNA extracted from fresh-frozen tumours were performed using Fluidigm technology according to the manufacturer's recommendations. The 21 PCR products from one tumour sample were then equimolarly pooled and submitted to the MiSeq (Illumina) sequencing as per the manufacturer's protocol. All mutations were validated by Sanger sequencing. Somatic mutations were confirmed using paired constitutional DNA.

mRNA expression profiling. Gene expression profiles of 71 samples were analysed using Affymetrix Human Genome U133 Plus 2.0 arrays. All samples were normalized in batches using the RMA algorithm (Bioconductor *affy* package), and probe set intensities were then averaged per gene symbol.

Identification of significantly mutated pathways. Gene set member lists were retrieved online from MSigDB[33], GO[34] and SMD[35] databases. We searched for gene sets harbouring more damaging mutations than expected by chance. Given the set G of all the genes sequenced with sufficient coverage, the set S of tumour samples (of size n) and any gene set P, we calculated the probability of observing a number of mutations equal or greater to that observed in P across the n samples according to a binomial law $B(k, p)$, with $k=n\times L(\mathrm{P})$ and the mutation rate $p=A(\mathrm{G,S})/(n\times L(G))$, where $L(\mathrm{X})$ is the sum of the lengths (in bp) of all genes/exons from a gene set X, and $A(\mathrm{G,S})$ is the total number of mutations observed in all the targeted sequences across all the samples from S.

© 2015 Macmillan Publishers Limited. All rights reserved.

Deregulated gene sets in TCF12 mutant samples. We performed a moderate *t*-test using LIMMA R package to identify significantly differentially expressed genes between *TCF12* mutant samples and *TCF12* wild-type samples ($P<0.05$ and absolute log fold change >0.6). Biological pathways and gene set member lists were retrieved online from MSigDB[33], GO[34] and SMD[35] databases. Enrichment *P* values were computed from a hypergeometric test between those gene sets and the initial list of differentially expressed genes. To visualize gene set activity, for each gene set defined as target genes of either *CDH1*, *TCF21*, *BMI1*, *EZH2* and found to be significantly deregulated in *TCF12* bHLH-altered samples compared with *TCF12* wild-type samples in O3 samples with co-deletion, we retrieved the complete member list from MSigDB[33] and computed a global mean gene expression value in each sample. We then ranked the samples according to the later global mean expression value for each of these gene sets.

Structure modelling. The Swiss Model[36] server was used to model mutated TCF12 and VMD software[37] used to align the structures of wild-type and mutated TCF12 proteins with STAMP (STructural Alignment of Multiple Proteins)[38]. Prediction of the functional effect of the R602M mutation on TCF12 was made using Project HOPE[39].

Statistical analysis. Statistical analysis was carried out using R3.0.1 software. A *P* value ≤ 0.05 was considered to be significant. Continuous variables were analysed using the Student's *t*-test or Mann–Whitney test. Categorical data were compared using Fisher's exact test or the χ^2-test. Overall survival of patients was the end point of the analysis. Survival time was calculated from the date of tumour diagnosis to the date of death. Patients who were not deceased were censored at the date of last contact. Mean follow-up time was computed among censored observations only. Kaplan–Meier survival curves according to genotype were generated and the homogeneity of the survival curves between genotypes was evaluated using the log-rank test. Power to demonstrate a relationship between mutation status and overall survival was estimated using sample size formulae for comparative binomial trials[40].

Cell culture. Human embryonic kidney HEK293T cell line (American Type Culture Collection) was maintained in a 5% CO_2-regulated incubator in DMEM Glutamax (Life Technologies), completed with 10% fetal bovine serum and penicillin/streptomycin (Life Technologies).

Plasmid construction. To construct the TCF12 wild-type plasmid, we cloned, by Gateway recombination (Life Technologies), a pENTR221 TCF12 Ultimate ORF Clone (Life Technologies) into a pDEST12 lentiviral vector (kind gift from P. Ravassard), under the control of hCMV promoter. The M260fs*5 and R326S mutations were generated by PCR mutagenesis using the Q5 Site-directed Mutagenesis kit (New England Biolabs) on pENTR221 TCF12 plasmid (primer sequences are detailed in Supplementary Data 5) and then cloned into the pDEST12 vector by LR Gateway cloning. Synthetic NdeI/MfeI fragments (encompassing sequences from exon 16 to the TAG stop codon of the ENST00000438423 isoform), containing the mutations E548fs*13, R602M and S683fs*14, were obtained from GeneCust, then substituted into pENTR221 and finally cloned by Gateway recombination into the pDEST12 plasmid. All expression plasmids were sequenced before use.

Luciferase expression assays. For each experiment, 10^5 exponentially growing HEK293T cells were seeded in 12-well plates and transfected 24 h later using Fugene6 (Promega), according to manufacturer's instructions, with 0.3 µg of a reporter plasmid encoding firefly luciferase under the control of an E-box-responsive element (Eb, kind gift from A. Lasorella), or 0.3 µg of Eb plasmid and 0.7 µg of a *TCF12* wild-type expression plasmid, or 0.3 µg of Eb plasmid and 0.7 µg of either *TCF12* mutant (M260fs*5, R326S, E548fs*13, R602M or S628fs*14) expression plasmid. For all points, data were normalized by adding 30 ng of renilla luciferase expression plasmid (pGL4.73, Promega, gift from F. Toledo). Cells were harvested 24 h after transfection, and luminescence was monitored using the Dual-Glo Luciferase assay system (Promega), according to the manufacturer's instructions, on a Spectramax M4 instrument and SoftMax Pro 6.2.2 software. All samples were run in triplicate, in four independent experiments.

Immunohistochemistry. Paraffin-embedded tumour sections were deparaffinized using standard protocols. Heat-mediated antigen retrieval was achieved by boiling sections in a pressure cooker with Citrate buffer at pH 6. Sections were blocked in 10% goat serum in PBS + 0.5% Triton X-100 for 30 min prior to incubation with an anti-TCF12 antibody (Proteintech Cat no.: 14419-1-AP) and then revealed using the Polink-2 HRP Plus Rabbit DAB Detection System (GBI Labs:D39-6). Photographs were taken at $\times 400$ magnification and processed using AxioVision software (Zeiss). The mitotic index in tumours was recorded as the number of mitotic figures in 10 high-power fields.

TCGA data. To complement our analysis, we made use of exome sequencing data on AO tumours generated by the TCGA (Supplementary Data 2).

References

1. Cairncross, G. *et al.* Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: long-term results of RTOG 9402. *J. Clin. Oncol.* **31,** 337–343 (2013).
2. Killela, P. J. *et al.* TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc. Natl Acad. Sci. USA* **110,** 6021–6026 (2013).
3. Riemenschneider, M. J., Koy, T. H. & Reifenberger, G. Expression of oligodendrocyte lineage genes in oligodendroglial and astrocytic gliomas. *Acta Neuropathol.* **107,** 277–282 (2004).
4. Bettegowda, C. *et al.* Mutations in CIC and FUBP1 contribute to human oligodendroglioma. *Science* **333,** 1453–1455 (2011).
5. Yip, S. *et al.* Concurrent CIC mutations, IDH mutations, and 1p/19q loss distinguish oligodendrogliomas from other cancers. *J. Pathol.* **226,** 7–16 (2012).
6. Yan, H. *et al.* IDH1 and IDH2 mutations in gliomas. *N Engl. J. Med.* **360,** 765–773 (2009).
7. Labussiere, M. *et al.* All the 1p19q codeleted gliomas are mutated on IDH1 or IDH2. *Neurology* **74,** 1886–1890 (2010).
8. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
9. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
10. Sharma, V. P. *et al.* Mutations in TCF12, encoding a basic helix-loop-helix partner of TWIST1, are a frequent cause of coronal craniosynostosis. *Nat. Genet.* **45,** 304–307 (2013).
11. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155,** 462–477 (2013).
12. Singh, D. *et al.* Transforming fusions of FGFR and TACC genes in human glioblastoma. *Science* **337,** 1231–1235 (2012).
13. Di Stefano, A. L. *et al.* Detection, characterization and inhibition of FGFR-TACC fusions in IDH wild type glioma. *Clin. Cancer Res.* doi: 10.1158/1078-0432.CCR-14-2199 (2015).
14. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40,** e169 (2012).
15. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10,** 1081–1082 (2013).
16. Seshagiri, S. *et al.* Recurrent R-spondin fusions in colon cancer. *Nature* **488,** 660–664 (2012).
17. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150,** 1107–1120 (2012).
18. Lee, C. C. *et al.* TCF12 protein functions as transcriptional repressor of E-cadherin, and its overexpression is correlated with metastasis of colorectal cancer. *J. Biol. Chem.* **287,** 2798–2809 (2012).
19. Sideridou, M. *et al.* Cdc6 expression represses E-cadherin transcription and activates adjacent replication origins. *J. Cell Biol.* **195,** 1123–1140 (2011).
20. Yang, M. H. *et al.* Bmi1 is essential in Twist1-induced epithelial-mesenchymal transition. *Nat. Cell Biol.* **12,** 982–992 (2010).
21. Uittenbogaard, M. & Chiaramello, A. Expression of the bHLH transcription factor Tcf12 (ME1) gene is linked to the expansion of precursor cell populations during neurogenesis. *Brain Res. Gene Expr. Patterns* **1,** 115–121 (2002).
22. Fu, H. *et al.* A genome-wide screen for spatially restricted expression patterns identifies transcription factors that regulate glial development. *J. Neurosci.* **29,** 11399–11408 (2009).
23. Paredes, J. *et al.* Epithelial E- and P-cadherins: role and clinical significance in cancer. *Biochim. Biophys. Acta* **1826,** 297–311 (2012).
24. Paumard-Hernandez, B. *et al.* Expanding the mutation spectrum in 182 Spanish probands with craniosynostosis: identification and characterization of novel TCF12 variants. *Eur. J. Hum. Genet.* doi: 10.1038/ejhg.2014.205 (2014).
25. Kleihues, P. & Cavenee, W. E. *World Health Organisation Classification of Tumours of the Central Nervous System* (WHO/IARC, 2000).
26. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).
27. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
28. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
29. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).
30. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41,** e67 (2013).

© 2015 Macmillan Publishers Limited. All rights reserved.

31. Iyer, M. K., Chinnaiyan, A. M. & Maher, C. A. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* **27,** 2903–2904 (2011).
32. Maher, C. A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA* **106,** 12353–12358 (2009).
33. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).
34. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).
35. Hubble, J. *et al.* Implementation of GenePattern within the Stanford Microarray Database. *Nucleic Acids Res.* **37,** D898–D901 (2009).
36. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22,** 195–201 (2006).
37. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38 27-8 (1996).
38. Russell, R. B. & Barton, G. J. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* **14,** 309–323 (1992).
39. Venselaar, H., Te Beek, T. A., Kuipers, R. K., Hekkelman, M. L. & Vriend, G. Protein structure analysis of mutations causing inheritable diseases. an e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* **11,** 548 (2010).
40. Farrington, C. P. & Manning, G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat. Med.* **9,** 1447–1454 (1990).

Acknowledgements

This work is part of the national program Cartes d'Identité des Tumeurs (CIT) (http://cit.ligue-cancer.net), Prise en charge des oligodendrogiomes anaplasiques (POLA) Network, POLA Tumor Bank, OncoNeuroTek tumorothèque du système nerveux central ICM APHP and the Institut National du Cancer (INCa) (http://www.e-cancer.fr). Research in Huillard and Sanson labs has received funding from the program 'Investissements d'avenir' ANR-10-IAIHU-06. Grant support from Génome Québec, le Ministère de l'Enseignement supérieur, de la Recherche, de la Science et de la Technologie (MESRST) Québec and McGill University is also acknowledged. At The Institute of Cancer Research, work was primarily supported by Cancer Research UK (C1298/A8362 Bobby Moore Fund for Cancer Research UK). D.C. is supported by Leukaemia Lymphoma Research. C.P.W. is funded by Myeloma UK. We are indebted to A. Lasorella and A. Iavarone for helpful discussion, technical advices and for providing the E-box-responsive reporter plasmid. We thank P. Ravassard, S Rozenberg and V. Lejour for discussion and technical advice, and A. Nadaradjane for the TCF12 structure modelling. I.S. is supported by a fellowship from the Ligue Nationale Contre le cancer. V.G. is supported by a fellowship from the Fondation ARC pour la Recherche sur le Cancer. Research in Huillard lab is supported by the Ligue Nationale Contre le Cancer, Fondation ARC pour la Recherche sur le Cancer, Institut National de la Santé et de la Recherche Médicale (INSERM) and European Union (FP7-PEOPLE-CIG-2012). Research in Sanson lab has been supported by grants from the Ligue Nationale Contre le Cancer, Fondation ARC pour la Recherche sur le Cancer and the Institut National du Cancer. The results published here are in whole or part based upon data generated by The Cancer Genome Atlas (TCGA) pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions that constitute the TCGA research network can be found at http://cancergenome.nih.gov/.

Author contributions

M.S., A.I., J.-Y.D., G.M.L., E.H. and R.S.H conceived the study. R.S.H., K.L., I.S., A.K., E.H. and M.S. wrote the manuscript. K.L., A.K., D.C., E.L. and A.d.R. designed and reviewed statistical and bioinformatic analyses. I.S., V.G., D.Z., T.P., Y.R., O.S. and S.H. performed experiments. K.L., D.C., S.E.D., C.W., M.F., A.K. and E.L. performed bioinformatic analyses. D.F.-B., F.D. and C.D. performed sample preparation. N.E. reviewed samples annotations and performed data management. All authors reviewed and contributed to the manuscript.

Additional information

Accession codes: All whole-exome sequencing and transcriptome data have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI), under the accession code EGAS00001001209. mRNA expression and SNP data can be accessed through ArrayExpress under accession numbers E-MTAB-2768 for mRNA expression data, and E-MTAB-3457, E-MTAB-3458, E-MTAB-2772 and E-MTAB-2771 for SNP data.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Labreche, K. *et al. TCF12* is mutated in anaplastic oligodendroglioma. *Nat. Commun.* 6:7207 doi: 10.1038/ncomms8207 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

POLA Network

Clovis Adam[16], Marie Andraud[17], Marie-Hélène Aubriot-Lorton[18], Luc Bauchet[19], Patrick Beauchesne[20], Claire Blechet[21], Mario Campone[22], Antoine Carpentier[23], Catherine Carpentier[24], Ioana Carpiuc[25], Marie-Pierre Chenard[26], Danchristian Chiforeanu[27], Olivier Chinot[28], Elisabeth Cohen-Moyal[29], Philippe Colin[30], Phong Dam-Hieu[31], Christine Desenclos[32], Nicolas Desse[33], Frederic Dhermain[34], Marie-Danièle Diebold[35], Sandrine Eimer[36], Thierry Faillot[37], Mélanie Fesneau[38], Denys Fontaine[39], Stéphane Gaillard[40], Guillaume Gauchotte[41], Claude Gaultier[42], Francois Ghiringhelli[43], Joel Godard[44], Edouard Marcel Gueye[45], Jean Sebastien Guillamo[46], Selma Hamdi-Elouadhani[47], Jerome Honnorat[48], Jean Louis Kemeny[49], Toufik Khallil[50], Anne Jouvet[51], Francois Labrousse[52], Olivier Langlois[53], Annie Laquerriere[54], Emmanuelle Lechapt-Zalcman[55], Caroline Le Guérinel[56], Pierre-Marie Levillain[57], Hugues Loiseau[58], Delphine Loussouarn[59], Claude-Alain Maurage[60], Philippe Menei[61], Marie Janette Motsuo Fotso[62], Georges Noel[63], Fabrice Parker[64], Michel Peoc'h[65], Marc Polivka[66], Isabelle Quintin-Roué[67], Carole Ramirez[68], Damien Ricard[69], Pomone Richard[70], Valérie Rigau[71], Audrey Rousseau[72], Gwenaelle Runavot[73], Henri Sevestre[74], Marie Christine Tortel[75], Emmanuelle Uro-Coste[76], Fanny Burel-Vandenbos[77], Elodie Vauleon[78], Gabriel Viennet[79], Chiara Villa[80], Michel Wager[57]

© 2015 Macmillan Publishers Limited. All rights reserved.

[16]Hôpital Bicêtre, Pathology Department, 94275 Le Kremlin-Bicêtre, France. [17]CHU Saint-Pierre de la Réunion, Pathology Department, Saint-Pierre de la Réunion, 97410 France. [18]CHU Dijon, Pathology Department, 21000 Dijon, France. [19]CHU de Montpellier, Neurosurgery Department, 34295 Montpellier, France. [20]CHU Nancy, Neuro-oncology Department, 54035 Nancy, France. [21]CHR Orléans, Pathology Department, 45000 Orléans, France. [22]Centre René Gauducheau, Medical Oncology Department, 44805 Saint-Herblain, France. [23]Hôpital Avicenne, Neurology Department, 93009 Bobigny, France. [24]Universite Pierre et Marie Curie, Centre de Recherche de l'institut du Cerveau et de la Moelle Epiniere and INSERM UMRS 975/CNR, 75013 Paris, France. [25]Clinique des Cèdres, Medical Oncology Department, 31700 Cornebarrieu, France. [26]CHU Strasbourg, Pathology Department, 67098 Strasbourg, France. [27]CHU Rennes, Pathology Department, 35033 Rennes, France. [28]Hôpital de la Timone, Assistance Publique—Hôpitaux de Marseille, Neuro-oncology Department, 13385 Marseille, France. [29]Institut Claudius Regaud, Radiotherapy Department, 31059 Toulouse, France. [30]Clinique de Courlancy, Radiotherapy Department, 51100 Reims, France. [31]Hôpital de la cavale blanche, CHU Brest, Neurosurgery Department, 29609 Brest, France. [32]Hôpital Nord, CHU Amiens, Neurosurgery Department, 80054 Amiens, France. [33]HIA Sainte-Anne, Neurosurgery Department, 83800 Toulon, France. [34]Institut Gustave Roussy, Radiotherapy Department, 94805 Villejuif, France. [35]CHU Reims, Pathology Department, 51092 Reims, France. [36]CHU de Bordeaux-GH Pellegrin, Pathology Department, 33000 Bordeaux, France. [37]Hôpital Beaujon, Neurosurgery Department, 92110 Clichy, France. [38]CHR Orléans, Radiotherapy Department, 45000 Orléans, France. [39]CHU Nice, Neurosurgery Department, 06002 Nice, France. [40]Hôpital Foch, Neurosurgery Department, 92151 Suresnes, France. [41]CHU Nancy, Pathology Department, 54035 Nancy, France. [42]CH Colmar, Neurology Department, 68024 Colmar, France. [43]Centre Georges-François Leclerc, Medical Oncology, 21079 Dijon, France. [44]Hôpital Jean Minjoz, CHU Besançon, Neurosurgery Department, 25030 Besançon, France. [45]Hôpital Dupuytren, CHU de Limoges, Neurosurgery Department, 87042 Limoges, France. [46]CHU de Caen, Neurology Department, 14033 Caen, France. [47]Hôpital Lariboisière, Neurosurgery Department, 75475 Paris, France. [48]Hospices Civils de Lyon, Hôpital Neurologique, Neuro-oncology Department, 69677 Bron, France. [49]CHU Clermont-Ferrand, Pathology Department, 63003 Clermont-Ferrand, France. [50]CHU Clermont-Ferrand, Neurosurgery Department, 63003 Clermont-Ferrand, France. [51]Hospices Civils de Lyon, Hôpital Neurologique, Pathology and Neuropathology Department, 69677 Bron, France. [52]Hôpital Dupuytren, CHU de Limoges, Pathology Department, 87042 Limoges, France. [53]CHU Charles Nicolle, Neurosurgery Department, 76000 Rouen, France. [54]CHU Charles Nicolle, Pathology Department, 76031 Rouen, France. [55]CHU de Caen, Pathology Department, Caen, 14033 France. [56]Hôpital Henri Mondor, Neurosurgery Department, 94010 Henri Mondor, France. [57]CHU Poitiers, Neurosurgery Department, 86000 Poitiers, France. [58]CHU de Bordeaux-GH Pellegrin, Neurosurgery Department, 33000 Bordeaux, France. [59]CHU Nantes, Pathology Department, 44093 Nantes, France. [60]CHU de Lille, Pathology Department, 59037 Lille, France. [61]CHU Angers, Neurosurgery Department, 49933 Angers, France. [62]Hôpital Nord, CHU Saint-Étienne, Neurosurgery Department, 42270 Saint-Priest en Jarez, France. [63]Centre Paul Strauss, Radiotherapy Department, 67065 Strasbourg, France. [64]Hôpital Bicêtre, Neurosurgery Department, 94275 Le Kremlin-Bicêtre, France. [65]Hôpital Nord, CHU Saint-Étienne, Pathology Department, 42270 Saint-Priest en Jarez, France. [66]Hôpital Lariboisière, Pathology Department, 75475 Paris, France. [67]Hôpital de la cavale blanche, CHU Brest, Pathology Department, 29609 Brest, France. [68]CHU de Lille, Neurosurgery Department, Lille, 59037 France. [69]HIA du Val de Grâce, Neurology Department, 75230 Paris, France. [70]Laboratoire les Feuillants, Pathology Department, 31023 Toulouse, France. [71]CHU de Montpellier, Pathology Department, 34295 Montpellier, France. [72]CHU Angers, Pathology Department, 49933 Angers, France. [73]CHU Saint-Pierre de la Réunion, Neurology Department, 97410 Saint-Pierre de la Réunion, France. [74]Hôpital Nord, CHU Amiens, Pathology Department, 80054 Amiens, France. [75]CH Colmar, Pathology Department, 68024 Colmar, France. [76]Hôpital Rangueil, CHU Toulouse, Pathology Department, 31059 Toulouse, France. [77]CHU Nice, Pathology Department, 06002 Nice, France. [78]Centre Eugène Marquis, Medical Oncology, 35042 Rennes, France. [79]Hôpital Jean Minjoz, CHU Besançon, Pathology Department, 25030 Besançon, France. [80]Hôpital Foch, Pathology Department, 92151 Suresnes, France.

© 2015 Macmillan Publishers Limited. All rights reserved.

nature COMMUNICATIONS

ARTICLE

Received 26 Sep 2014 | Accepted 25 Nov 2014 | Published 22 Jan 2015

DOI: 10.1038/ncomms6973 OPEN

Whole-exome sequencing reveals the mutational spectrum of testicular germ cell tumours

Kevin Litchfield[1], Brenda Summersgill[2], Shawn Yost[1], Razvan Sultana[1], Karim Labreche[1,3], Darshna Dudakia[1], Anthony Renwick[1], Sheila Seal[1], Reem Al-Saadi[2], Peter Broderick[1], Nicholas C. Turner[4], Richard S. Houlston[1], Robert Huddart[5], Janet Shipley[2] & Clare Turnbull[1,6]

Testicular germ cell tumours (TGCTs) are the most common cancer in young men. Here we perform whole-exome sequencing (WES) of 42 TGCTs to comprehensively study the cancer's mutational profile. The mutation rate is uniformly low in all of the tumours (mean 0.5 mutations per Mb) as compared with common cancers, consistent with the embryological origin of TGCT. In addition to expected copy number gain of chromosome 12p and mutation of *KIT*, we identify recurrent mutations in the tumour suppressor gene *CDC27* (11.9%). Copy number analysis reveals recurring amplification of the spermatocyte development gene *FSIP2* (15.3%) and a 0.4 Mb region at Xq28 (15.3%). Two treatment-refractory patients are shown to harbour *XRCC2* mutations, a gene strongly implicated in defining cisplatin resistance. Our findings provide further insights into genes involved in the development and progression of TGCT.

[1] Division of Genetics and Epidemiology, The Institute of Cancer Research, Fulham Road, London SW3 6JB, UK. [2] Divisions of Molecular Pathology and Cancer Therapeutics, The Institute of Cancer Research, Fulham Road, London SW3 6JB, UK. [3] Inserm U 1127, CNRS UMR 7225, Sorbonne Universités, UPMC Univ Paris 06 UMR S 1127, Institut du Cerveau et de la Moelle épinière, ICM, F-75019, Paris, France. [4] The Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, Fulham Road, London SW3 6JB, UK. [5] Academic Radiotherapy Unit, The Institute of Cancer Research, Fulham Road, London SW3 6JB, UK. [6] William Harvey Research Institute, Queen Mary University London, Charterhouse Square, London EC1M 6BQ, UK. Correspondence and requests for materials should be addressed to C.T. (email: clare.turnbull@icr.ac.uk).

© 2015 Macmillan Publishers Limited. All rights reserved.

TGCTs are the most common cancer affecting young men, with a mean age at diagnosis of 36 years[1,2]. The main TGCT histologies are seminomas, which resemble undifferentiated primary germ cells, and non-seminomas, which show differing degrees of differentiation. Cure rates for TGCTS are generally high, due to the sensitivity of malignant testicular germ cells to platinum-based chemotherapies, however this is at the cost of an increased risk of metabolic syndrome, infertility and secondary cancer[3–5]. Furthermore, there are limited options for the patients who are platinum resistant, a group for whom the long-term survival rate is poor[6].

Overall, TGCTs are markedly aneuploid with recurring gain of chromosomes 7, 8, 21, 22 and X[7–13]. In addition, gain of chromosomal material from 12p is noted in virtually all cases[7–9], with genomic amplification and overexpression of genes in the 12p11.2-p12.1 region reported in ~10% of TGCTs[14]. *KRAS* is located in this region and has been proposed as the candidate driver[14]. Focused studies of TGCTs have identified somatic missense mutations and amplifications of the oncogene *KIT*, present in ~25% of seminomas[15,16]. These reported mutations are clustered in the juxta membrane and kinase encoding domains of KIT[15,16]. However, a study of 518 other protein kinase encoding genes failed to conclusively identify any new driver mutations[17]. Beyond these focused interrogations of specific genes, no systematic mutational analysis across all genes in a large series of TGCT samples has been reported to our knowledge.

Here we perform WES of a series of 42 TGCTs to characterize the mutational signature of these tumours and to search for additional driver mutations and pathways disrupted. Our analyses demonstrate these tumours to be relatively homogeneous in profile with a markedly low rate of non-synonymous mutations and provide some novel insights into the genomic architecture of this biologically interesting tumour type.

Results

Overview of TGCT mutational landscape. The 42 TGCT cases comprised 16 seminomas, 18 non-seminomas, 4 mixed seminoma/non-seminoma histology and 4 tumours of indeterminant classification. Fresh frozen tumour tissue and matched germline blood samples were obtained from each patient and WES was performed on extracted DNA, achieving mean coverage of 72 × across targeted bases with 86% of targeted bases being covered at ≥20. Sequencing was conducted using Ilumina technology, with subsequent alignment, mapping and variant calling performed using Burrows–Wheeler Aligner (BWA)/Stampy/GATK/MuTect software. Across all 42 cases a total of 1,168 somatic single nucleotide variants (SNVs), and 111 small scale somatic insertion –deletions (indels) were identified, resulting in a combined total of 795 non-synonymous mutations, equating to a mean rate of 0.51 somatic mutations per Mb. By comparison, recent large-scale analysis across 27 cancer types recorded mean rates as high as 11.0 Mb^{-1} in melanoma and 8.0 Mb^{-1} in lung cancers with a mean rate across all tumour types of 4.0 Mb^{-1}, some eight times higher than that seen here in TGCT (ref. 18). Indeed the mutation rate in TGCT is within the second lowest decile, only marginally greater than paediatric cancers such as Ewing sarcoma (0.3 Mb^{-1}) and Rhabdoid tumour (0.15 Mb^{-1}). This observation is entirely consistent with oncogenic origins of TGCT arising during embryonic development[19]. Of additional note is the high intra-patient homogeneity in mutation rate present in our data, with a s.d. of just 0.24 across the 42 tumours and the extreme lowest to extreme highest mutation rate varying by only 1 order of magnitude. This variation is low compared with the 3 orders of magnitude inter-sample variation observed for acute myeloid leukaemia, which has a comparable mutation rate[18]. Of note, there were no genes that were recurrently mutated or structural variants shared between the tumours in which the mutational rate was >2 s.d. above the mean (two tumours). The mutational spectrum of SNVs in the TGCTs was typified by an excess of CG>TA transitions (27% of SNVs), as observed in most solid tumours[18,20] (Fig. 1). In addition, TA>CG transitions (23%) as well as CG>AT transversions (31%, of which the majority were C>A) were also over-represented. While C>A transversions are observed at higher proportion in lung cancers postulated to be due to exposure to tobacco carcinogens[18], this pattern is also has also been reported in melanoma, neuroblastoma and chronic lymphocytic leukaemia[21].

Driver genes. We used MutSigCV version 1.4 to identify genes harbouring more non-synonymous mutations than expected by chance given gene size, sequence context and gene-specific background mutation rates[18]. *KIT* was identified as the most significantly mutated gene (Fig. 2), with mutations seen in 14.3% across all TGC tumours, but predominantly found in seminomas (31.3%); a result consistent with previously reported observations[16,22]. All of the six *KIT* mutations we identified were in hotspot domains—five non-synonymous SNVs in exon 17 (kinase encoding domain) and one in exon 11 (juxta membrane domain). The absence of another gene ranked above *KIT* is a notable result, given our study assesses an exome-wide compliment of genes. In addition to *KIT*, a non-synonymous SNV was also observed in previously proposed TGCT driver gene *KRAS*. While *p53* mutations have been suggested to be a feature of TGCT[23], none were observed in our data set, consistent with most recent studies[17,24,25]. We validated all *KIT/KRAS* mutations called by next generation sequencing (NGS) using Sanger sequencing of the respective exons across all samples and to ensure no additional mutations were missed. In all cases, Sanger sequencing was 100% concordant with NGS.

In addition to *KIT* and *KRAS*, there was an over-representation of mutations in cell division cycle 27 (*CDC27*) (11.9%;

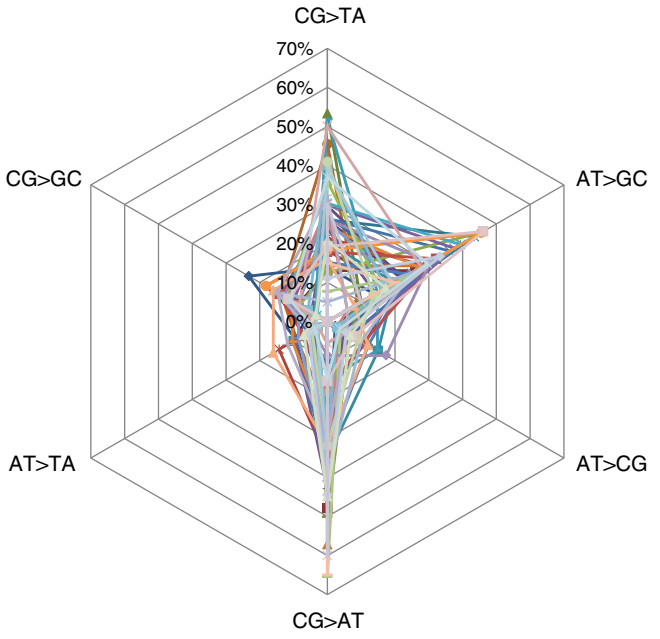


Figure 1 | TGCT somatic SNV spectrum exome wide. Proportions are displayed for all 12 possible SNV alterations, collapsed by strand complementarity. Each line represents one of the 42 tumours.

© 2015 Macmillan Publishers Limited. All rights reserved.

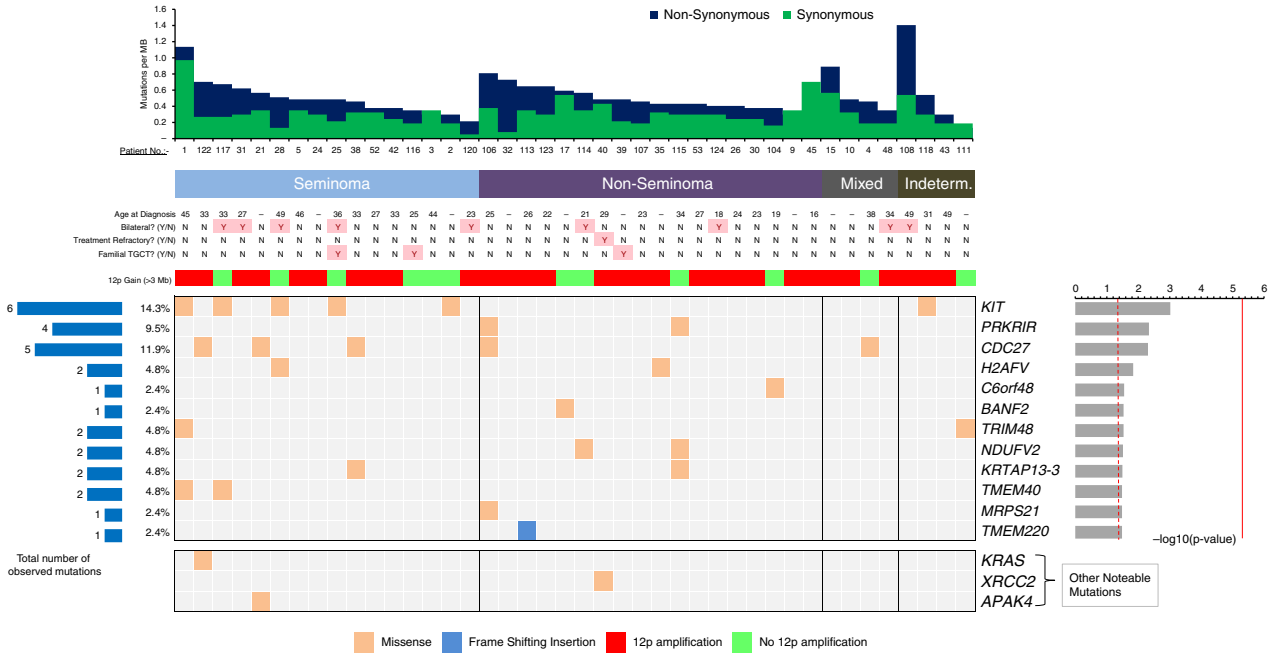


Figure 2 | Mutated genes in testicular germ cell tumour by histological subtype. The top bars represent somatic mutation rate per sample for the 42 samples (synonymous and non-synonymous (including small-scale indels)). The genes listed on the right are mutated genes as prioritized by MutSigCV, ranked by $-\log_{10}(P \text{ value})$ (far right), with the dotted red line denoting a significance threshold of $P = 0.05$ and the solid red line a genome-wide significance threshold of 5×10^{-6} (see Methods). Below the top ranked genes in a separate box are other notable but non-significant mutations. Mutations by sample are depicted in the central box, with colour indicating mutation type as per the legend. The far left bars represent the absolute number of mutations observed per gene across all samples and adjacent to this is the % of samples this represents.

5 mutations, 5 tumours) and *PRKRIR* (4 mutations, 2 tumours), neither of which have been previously reported as TGCT drivers. *CDC27* is a core component of the anaphase-promoting complex/cyclosome, a multi-subunit E3 ubiquitin ligase that governs cell cycle progression, through ubiquitination and degradation of G1/mitotic checkpoint regulators[26]. Anaphase-promoting complex/cyclosome recruits its substrates via one of the two adaptor proteins CDC20 or CDH1, overexpression of which have been linked to multiple tumours[27–29]. *CDC27* is downregulated in breast cancer and *CDC27* is postulated to be a tumour suppressor[30]. All of the *CDC27* mutations we identified were missense variants, characterized by a consistently low frequency of mutant allelic reads (8–14%), consistent with *CDC27* mutation being present only in a subclone of each tumour sample. Intriguingly subclonal low frequency of *CDC27* mutation has also recently been demonstrated in a colonic adenocarcinoma[31].

Pathway analysis. To increase our ability to identify cancer drivers and delineate associated oncogenic pathways for TGCT, we incorporated mutation data from multiple tumour types using Oncodrive-fm[32] as implemented within the IntOGen-mutations platform[33]. The most frequently mutated pathways were those involved in metabolism (mutated in 93%), pathways in cancer (54%), endocytosis (54%) and PI3K–Akt signalling (54%). The most significantly mutated pathway was RNA degradation (14.6%), with a biased accumulation of functional mutations (fm-bias, $P = 3.8 \times 10^{-3}$), observed across six different genes (see methods and Supplementary Table 2).

Copy number variation. The 42 tumours were analyzed for copy number variation (CNV) using software package ExomeCNV[34]. Focal CNVs (up to 3 Mb) were identified in all tumours and large-scale CNVs (≥ 3 Mb) were detected in 35 (83%) tumours, (Fig. 3). Across all 42 cases the proportion of the tumour genome showing CNV ranged from 0.1 to 48.4% per genome (mean 10.8%). The most frequent large-scale chromosome abnormality was 12p copy number gain, present in 30 of the 42 tumours (71%), of which 25 were 12p isochromosomes, a result consistent with previous experimental observations[7–9]. The remaining 12 cases without large-scale 12p gain all showed evidence of focal copy number amplification of 12p, however, detailed analysis of these sub-regions did not reveal any recurring hotspots. Other recurring large-scale copy number changes included gain of chromosome X (16 cases, 38%) as well as gains of chromosomes 7 ($n = 15$; 36%), 21 ($n = 12$; 29%) and 22 ($n = 11$; 26%), findings again consistent with previous studies[7–13]. In addition, we observed large-scale copy number deletion of chromosome Y (10 cases, 24%). We used previously generated chromosomal comparative genomic hybridization (CGH) data for 24 of the tumours[12,35,36] to validate our large-scale CNVs for the known mutational event at 12p; concordance between NGS/CGH was 92%.

In terms of focal events three tumours (patients 115, 53 and 43) exhibited a high degree of chromosomal instability, with a 19-fold increase in focal alterations compared with the others. We assessed these cases for evidence of chromothripsis, which we defined as >20 CNVs on a chromosome single arm. While this technical definition was met for several loci, the majority of events were spread uniformly across the genome with no common hotspots across the three tumours. Excluding these three tumours we undertook an analysis of the focal alterations seen in the remaining 39 tumours to identify any recurrent patterns. Mapping the coordinates of all focal copy number events to genes, all possible gene alterations were assessed, quality filtered and ranked by frequency (Table 1 and methods). The highest ranking gene from this analysis was fibrous sheath interacting protein 2 (*FSIP2*) at 2q32.1, with seven recurring amplifications observed across six (15.3%) tumours. *FSIP2*

© 2015 Macmillan Publishers Limited. All rights reserved.

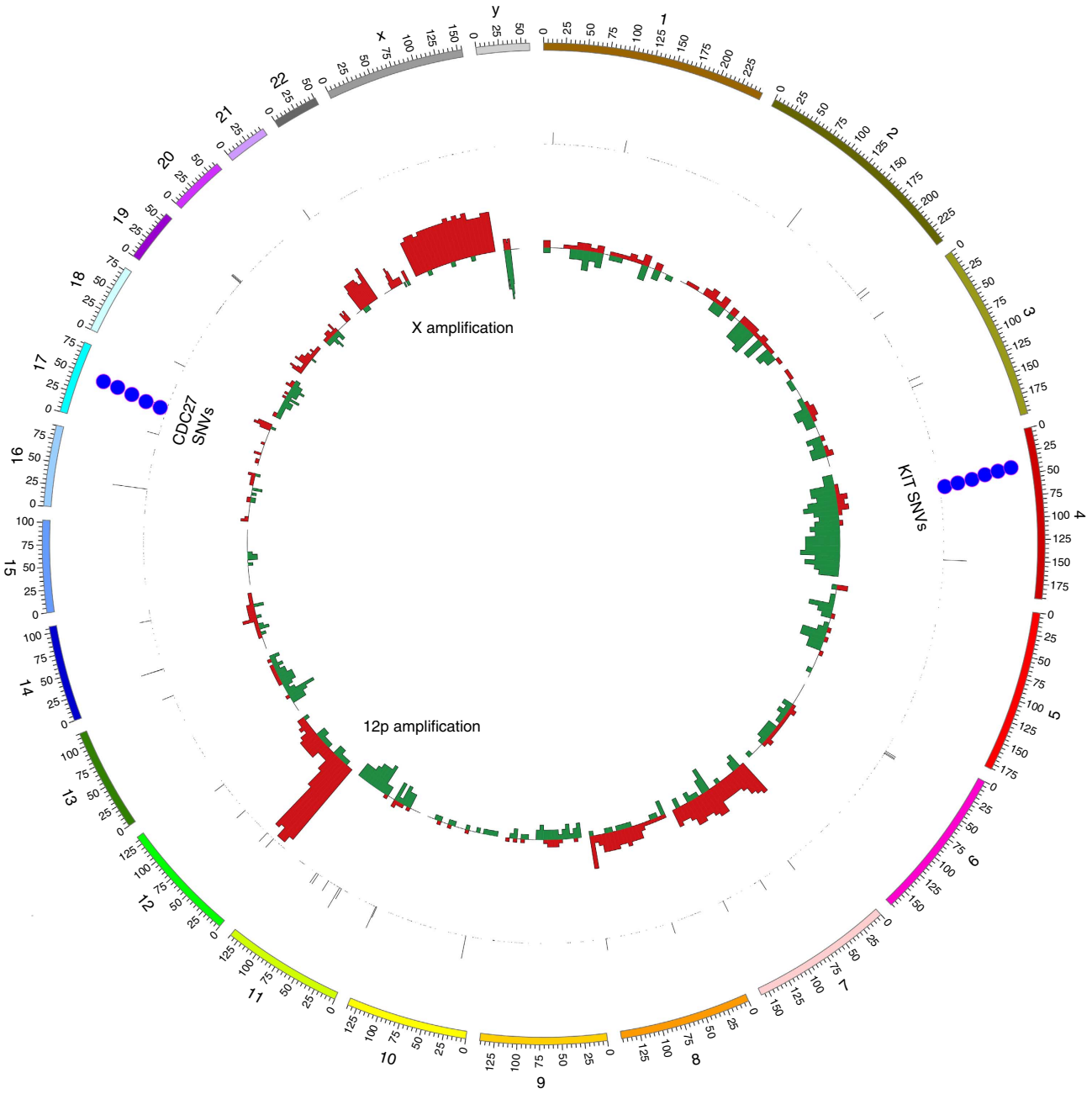


Figure 3 | Circos Plot showing the count of SNV variants and copy number changes in the 42 tumours. Outer ring marks the count of SNV variants across all 42 samples with proposed driver SNVs as blue dots and other SNVs as black lines; inner ring marks large-scale copy number gains (red) and losses (green).

amplifications were all 8–9 kb in length spanning a sub-region of the gene coding sequence, encompassing exons 16–17. Recent functional evidence has demonstrated that part-gene amplifications do affect gene expression levels, with an effect size comparable to that of full-gene amplification[37]. Our finding of recurrent *FSIP2* amplification is corroborated by recent high resolution SNP array data on an independent series of seminomas[38], which documented *FSIP2* amplification in 22% of tumours. Across both studies *FSIP2* is the only gene consistently observed with focal amplification in >10% of cases. There is a strong biological basis for abnormalities of *FSIP2* being a feature of TGCTs *a priori*. The fibrous sheath is a cytoskeletal structure located in the principle piece region of the sperm flagellum. Transcription of *FSIP2* begins in late spermatocyte development with mouse model data demonstrating it to be expressed exclusively in the testis[39]. Furthermore, FSIP2 also binds to another fibrous sheath enzyme A kinase (PRKA) anchor protein 4 (AKAP4), which has been linked to male infertility[40]. Interestingly the tumour from patient 21, which harboured a *FSIP2* amplification, also carried a missense mutation in *AKAP4*.

Other focal events observed included a 0.4 Mb region at Xq28, with amplification in six cases. This region contains 18 genes, including testis expressed 28 (*TEX28*) and transketolase like gene 1 (*TKTL1*), both of which are overexpressed in the human testis[41]. *TKTL1* is hypothesized to play a role in tumour response to hypoxia with increased *TKTL1* expression correlating with poor patient outcome in many solid tumours[42].

© 2015 Macmillan Publishers Limited. All rights reserved.

Table 1 | Genes with five or more recurrent copy number gains/losses.

Gene (s)	Region	Losses	Gains	Total CNVs
FSIP2	2q32.1	2	7	9
AK2	1p35.1	0	7	7
ZNF644	1p22.2	0	7	7
ENPP3	6q23.2	0	7	7
MUC12	7q22.1	0	7	7
AHNAK2	14q32.33	0	7	7
TSPEAR	21q22.3	0	7	7
FLG	1q21.3	1	6	7
AK056431	1q21.3	1	6	7
HCFC1, TMEM187, MIR3202-1, IRAK1, MIR718, MECP2, OPN1LW, TEX28, OPN1MW, TKTL1, FLNA, EMD, AK307233, RPL10, SNORA70, DQ570720, DNASE1L1, TAZ	Xq28	0	6	6
CHRND	2q37.1	0	6	6
CTAGE9	6q23.2	0	6	6
MUC5B	11p15.5	0	6	6

CNV, copy number variation.
Focal CNVs included are defined as <3 Mb in length. See methods for further details on quality filters applied.

Clinicopathological-molecular associations. SNV/indel somatic mutation rates between seminoma and non-seminoma cases were almost identical; 0.50 mutations per Mb and 0.49 mutations per Mb respectively. *KIT* mutations were observed predominantly in seminoma cases, as previously reported. The proportion of the genome showing CNV was elevated (+47%) in non-seminona tumours. A correlation between somatic mutational rate and patient age was seen ($r=0.36$), with the mean rate for patients aged >40 years being 0.69 compared with 0.48 for cases <40 ($P=0.05$, two-sided Student's *t*-test). This is consistent with a model in which the majority of mutations are passenger mutations that accumulate with patient age following the early *in utero* oncogenic transformation of germ cells. Of particular clinical interest is the mutational profile of treatment-refractory TGCT, a rare subset of ~3% of patients in whom there is disease progression despite platinum-based chemotherapy. Within our cohort only one such patient, 40, had this profile of therapeutic response, so any conclusions are speculative. Accepting this caveat the mutational rate for this tumour was 0.49 Mb^{-1}, a rate comparable to the overall cohort, and of the 18 SNVs identified in this patient (see Supplementary Table 1), a mutation in gene *XRCC2* (c.6T>Gp.Cys2Trp) is of particular note. *XRCC2* encodes a member of the RecA/Rad51-related protein family, which participates in homologous recombination maintaining chromosome stability and repair of DNA damage. Importantly *XRCC2* mutant animal clones show increased resistance to cisplatin through enhanced DNA repair activity[43], and *XRCC2* germline variants have been shown to significantly associate with cytotoxic resistance in breast cancer[44]. In addition to the treatment-refractory patient in our main cohort, we also performed exome sequencing of tumour DNA from one additional platinum refractory case (germline DNA was not available, patient 109), identifying a further mutation in *XRCC2* (c.2T>Gp.Met1Arg). This additional variant had alternative allele frequency of only 4%, making it difficult to validate by Sanger. Both *XRCC2* mutations are predicted to be pathogenic on the basis of *in silico* analysis using the CONDEL algorithm (CONsensus DELeteriousness (CONDEL) score of non-synonymous SNVs, http://bg.upf.edu/fannsdb/help)[45,46].

Discussion

Our exome analysis has confirmed mutation of *KIT* and recurrent copy number gain of 12p as archetypical features of TGCT. We have also characterized the mutational signature of TGCTs, demonstrating a homogeneous profile with a markedly low SNV mutation rate, consistent with the embryonic origins of the disease. This low rate of point mutations (that is, SNVs) is contrasted, however, by frequent large-scale copy number gains, of not only 12p but also chromosomes 7, 21, 22 and X. Since our study was empowered to identify recurrent mutations having frequency of >15% (84% power), we can conclude that it is unlikely that additional high frequency driver mutations will exist.

We did, however, identify novel mutations in the probable tumour suppressor gene *CDC27*, implicating *CDC27* mutation as a potential oncogenic factor in a subset of TGCTs. Functionally *CDC27* interacts with spindle checkpoint proteins encoded by *MAD2* (ref. 47) and *TEX14* (ref. 48) genes, the latter of which resides in a linkage disequilibrium block associated through recent genome-wide association study (GWAS) with germline TGCT predisposition[49]. Interestingly three of the other TGCT GWAS risk loci contain genes also related to mitotic spindle assembly—*MAD1L1, CENPE* and *PMF1* (refs 49,50). Collectively, such observations provide further evidence of commonality between germline and somatic TGCT pathways, a notable result given the previous precedent that *KITLG*, the ligand which binds *KIT*, is the only gene within the linkage disequilibrium block at the strongest existing TGCT GWAS risk locus (odds ratio ~2.5)[51]. Aside from *CDC27*, we also observed mutations in several other genes at a frequency of <10%; at this lower frequency our study was not sufficiently powered to comprehensively evaluate the genetic mutational profile (our power to detect mutations with frequencies of 10% and 5% was only 14%).

Previous CGH studies have characterized the aneuploidy nature of TGCTs, and our findings are consistent with these analyses. We hypothesized that NGS exome data, with average probe lengths of ~200 bp, would allow identification of novel small-scale CNVs below the level detectable by CGH. We performed this analysis and identified recurring focal copy number alterations in the spermatocyte development gene *FSIP2*, a finding corroborated by previous independent orthologous study. Meta-analysis of the two experiments shows this to be significant at $P=6.8\times10^{-9}$. *FSIP2* is shown to be unique to spermatogenic cells and is hypothesized to act as a linker protein, binding *AKAP4* to the fibrous sheath[39]. Dysplasia of the fibrous sheath and mutations in *AKAP4* have both been linked to male infertility[40,52], an established risk factor for TGCT[53]. The additional observation of an *AKAP4* missense mutation further implicates this pathway, although the exact mechanisms facilitating tumorigenesis remain to be elucidated. Furthermore, we observed recurrent deletion of chromosome Y, a finding that also has interesting resonance with the germline as chromosome Y 'gr/gr' germline deletions are linked to both TGCT predisposition and male infertility[54,55]. In addition, we identified a recurring focal amplification of 0.4 Mb in length at Xq28, a region encompassing 18 genes, several of which may plausibly link to TGCT. Several observations implicate chromosome X in germ cell oncogenesis, with family studies suggesting a possible X-linked model of inheritance for TGCT genetic susceptibility[56]. In addition, patients with Klinefelter syndrome (47XXY constitutional karyotype) have a 67-fold elevated risk of developing mediastinal germ cell tumours[57].

We found no significant difference observed in the mutational rate between seminoma and non-seminoma cases. This is consistent with findings from germline genetic studies of TGCT,

© 2015 Macmillan Publishers Limited. All rights reserved.

where no differential genotype risk has been observed between histological sub-groups[49,51,58]. This supports a hypothesis of commonality in the oncogenic pathways activated, with differentiation occurring later in the tumour formation. This hypothesis is further supported by the observation of TGCT cases with mixed pathology[59], as well as bilateral and familial cases displaying tumours with inconsistent histological types[60,61]. Descriptive analysis of a single treatment-refractory patient in our cohort revealed a *XRCC2* mutation, a DNA repair gene which has been demonstrated to promote cisplatin resistance in animal studies[43]. Further analysis of one additional treatment-refractory tumour sample revealed some evidence for a second *XRCC2* mutation. Cell line studies suggest that the exceptional sensitivity of TGCTs to cisplatin is due to their inability to repair treatment-induced DNA damage, due to the low expression of DNA repair genes such as *ERCC1* (ref. 62). In addition, cisplatin-resistant embryonal carcinoma cell lines show sensitivity to poly(ADP-ribose) polymerase (PARP) inhibition, through blocking their acquired ability to repair DNA[63]. The observation of *XRCC2* mutations in our patient tumour data expands on these previous animal and cell line studies, further supporting an important role for this pathway.

To our knowledge this study represents the largest comprehensive sequencing study of TGCT conducted to date. While we have implemented strategies to accurately identify the mutational landscape of this tumour, we were only well powered to identify genes with high mutational frequency. Hence further insights into the biology of TGCT should be forthcoming through additional sequencing initiatives and meta-analyses of such data. This is likely to be especially important given the importance of probable histological subtype-specific changes, the subclonal architecture of TGCT and differences that are likely to be seen in platinum-resistant tumours.

Methods

Sample description. Samples were collected from TGCT patients at the Royal Marsden Hospital NHS Trust, UK. Informed consent was obtained from all participants and the study was approved by the Institute of Cancer Research/Royal Marsden Hospital Committee for Clinical Research (study number CCR2014). The samples have been previously reported in other studies[10,12,36,61,64]. Surgical specimens were snap frozen within 30 min of surgery and matched blood samples were collected at the time of surgery. Tumour samples were trimmed to remove surrounding normal tissue, and tumour cells were confirmed by histological assessment. Tumour and matched lymphocyte DNA were extracted by standard techniques[65,66]. Tumour samples from patients 26 and 9 were obtained post chemotherapy. Clinical characteristics of our sample cohort were representative of the broader patient population, in terms of histological sub-types, patient age, familial TGCT and response to treatment. Our series was, however, enriched for cases with bilateral disease (9/42 cases in our series compared with a frequency of ~5% in the broader patient population).

Whole-exome sequencing. Samples were quantified using Qubit technology (Invitrogen, Carlsbad, CA, USA) and sequencing libraries constructed from 50 ng of respective normal/tumour DNA. Library preparation was performed using 37 Mb Nextera Rapid Capture Exome kits (Ilumina, San Diego, CA, USA), with enzymatic tagmentation, indexing PCR, clean-up, pooling, target enrichment and post-capture PCR amplification/quality control performed in-house, following standardized protocols as per manufacturer guidelines. Samples underwent paired-end sequencing using the Ilumina HiSeq2500 platform with a 100-bp read length. Mean coverage of 73.6× and 69.0× were achieved across targeted bases for tumour and normal samples, respectively. FASTQ files were generated using Illumina CASAVA software (v.1.8.1, Illumina) and aligned to the human reference genome (b37/hg19) using BWA (v. 0.5.10, http://bio-bwa.sourceforge.net/)/Stampy (v.1.0.23) packages. PCR duplicates were removed and coverage metrics were calculated using Picard-tools (v.1.48, http://picard.sourceforge.net/). Coverage metrics demonstrated a mean of 95% of target bases achieved >10× coverage and 86% >20×. The Genome Analysis Toolkit (GATK, v. 3.1-1, http://www.broadinstitute.org/gatk/) was used for local indel realignment/base quality score recalibration and SNVs were called using MuTect (v. 1.1.4). Data was quality filtered using in-house FoxoG software to remove potential artefactual variants introduced through DNA oxidation[21]. FoxoG ensured variants were supported by a minimum of one alternative read in each strand direction, a mean Phred base quality score of >26, mean mapping quality ≥50 and an alignability site score of 1.0. Small-scale insertion/deletions (indels) were called using GATK.

We used MutSigCV (v.1.4) to identify genes that somatically mutated more often than would be expected by chance[18], after first excluding common germline SNPs with minor allele frequency >25% as recorded in either dbSNP (http://www.ncbi.nlm.nih.gov/SNP/), 1000 genomes (http://www.1000genomes.org) or in our in-house data from exome sequencing of the UK 1958 birth cohort (Houlston *et al.*, personal communication). In total, 33 common germline SNP variants were removed across all samples. MutSigCV was run using the standard genomic covariates of (i) global gene expression data, (ii) DNA replication time and (iii) HiC statistic of open versus closed chromatin states. We used Oncodrive-fm[32] as implemented within the IntOGen-mutations platform[67], using data mutation data from multiple tumour studies (http://bg.upf.edu/group/projects/oncodrive-fm.php; http://www.intogen.org/analysis/mutations/)

Confirmation sequencing. Confirmation sequencing was performed with bidirectional Sanger sequencing of *KIT* (exons 11 and 17) and *KRAS* (exon 2) across all 84 tumour/normal samples. Primer sequences are shown in Supplementary Table 3. Mutational analysis was conducted using Mutation Surveyor (v.3.97, SoftGenetics, State College, PA, USA).

CNV analysis. CNV analysis was conducted using the CRAN package ExomeCNV[34], a statistical algorithm designed to detect CNV, and loss of heterozygosity (LOH) events using depth-of-coverage and B-allele frequencies (https://secure.genome.ucla.edu/index.php/ExomeCNV_User_Guide). ExomeCNV is calibrated to achieve high levels of sensitivity and specificity, with a power to detect 95% for CNVs down to 500 bp in length[34]. When recently tested using a matched tumour/normal exome data set with ~40× coverage, ExomeCNV achieved 97% specificity and 86% sensitivity compared with results from Illumina Omni-1 SNP array[34]. To calculate CNVs, we first generated coverage files using GATK, and then used ExomeCNV to calculate log coverage ratios between matched tumour/normal samples and make CNV calls per exon. Exonic CNV calls were combined into segments using circular binary segmentation. LOH calls were made by first identifying all heterozygous germline positions per case, using Platypus (v.0.5.2) for germline variant calling. GATK was then used to create BAF files per case and ExomeCNV used to call LOH at heterozygous positions individually and at combined LOH segments.

CNV results were classified as large-scale (>3 Mb in length) or focal (<3 Mb) and filtered by coverage ratio selecting copy number gain >1.3 or loss <0.7, retaining calls with a specificity confidence score of 1.0. Focal events were analyzed by gene, mapping the coordinates of all events to gene coding start and end points to assess all possible gene alterations. Small-scale regions showing susceptibility to variable levels of coverage, that is, exact same probes frequently altered and with both copy number gain and loss, were removed to avoid false-positive associations.

Pathway analysis. Pathway analysis was performed using Oncodrive-fm[32] as implemented within the IntOGen-mutations platform[67], using the 1,168 SNVs and 111 indel mutations called across the 42 tumours.

Statistical analyses. Statistical significance of mutations were determined by testing whether the observed mutation counts in a gene significantly exceeded the expected counts based on a gene-specific background mutation rate, as implemented in MutSigCV (v.1.4). Plotted in the far section of Fig. 2 are the resulting $-\log10$ (P values), with the dotted red line denoting a significance threshold of $P=0.05$ and the solid red line a genome-wide significance threshold of $P=5\times10^{-6}$. Due to the overall low frequency of mutations observed in our data set, and the way such tumour types are treated by MutSigCV, no genes were significant at the genome-wide level, not even previously known TGCT driver gene *KIT*. Power analysis was conducted using a binomial power model, based on recent methods published by the Cancer Genome Analysis group at the Broad Institute[68], incorporating the average background somatic mutation rate specifically observed for TGCT, sample size and assuming a genome-wide significance level of $P\leq5\times10^{-6}$. Significance of focal copy number events by gene was calculated under a binomial distribution. Meta-analysis was conducted using the Fisher method of combining P values from independent tests. Statistical analysis were carried out using R3.0.2 (http://www.r-project.org/) and Stata12 (StataCorp, Lakeway Drive College Station, TX, USA) software. Continuous variables were analyzed using Student's t-tests. We considered a P value of 0.05 (two sided) as being statistically significant.

References

1. Bray, F., Ferlay, J., Devesa, S. S., McGlynn, K. A. & Moller, H. Interpreting the international trends in testicular seminoma and nonseminoma incidence. *Nat. Clin. Pract. Urol.* **3,** 532–543 (2006).
2. Ruf, C. G. *et al.* Changes in epidemiologic features of testicular germ cell cancer: age at diagnosis and relative frequency of seminoma are constantly and significantly increasing. *Urol. Oncol.* **32,** 33.e31–33.e36 (2014).

© 2015 Macmillan Publishers Limited. All rights reserved.

3. de Haas, E. C. *et al.* Early development of the metabolic syndrome after chemotherapy for testicular cancer. *Ann. Oncol.* **24,** 749–755 (2013).
4. Bujan, L. *et al.* Impact of chemotherapy and radiotherapy for testicular germ cell tumors on spermatogenesis and sperm DNA: a multicenter prospective study from the CECOS network. *Fertil. Steril.* **100,** 673–680 (2013).
5. Rusner, C. *et al.* Risk of second primary cancers after testicular cancer in East and West Germany: a focus on contralateral testicular cancers. *Asian J. Androl.* **16,** 285–289 (2014).
6. Nitzsche, B. *et al.* Anti-tumour activity of two novel compounds in cisplatin-resistant testicular germ cell cancer. *Br. J. Cancer* **107,** 1853–1863 (2012).
7. Sandberg, A. A., Meloni, A. M. & Suijkerbuijk, R. F. Reviews of chromosome studies in urological tumors.3. Cytogenetics and genes in testicular tumors. *J. Urol.* **155,** 1531–1556 (1996).
8. Atkin, N. B. & Baker, M. C. Specific chromosome change, I(12p), in testicular-tumors. *Lancet* **2,** 1349–1349 (1982).
9. Atkin, N. B. & Baker, M. C. I(12p)—specific chromosomal marker in seminoma and malignant teratoma of the testis. *Cancer. Genet. Cytogenet.* **10,** 199–204 (1983).
10. Henegariu, O., Vance, G. H., Heiber, D., Pera, M. & Heerema, N. A. Triple-color FISH analysis of 12p amplification in testicular germ-cell tumors using 12p band-specific painting probes. *J. Mol. Med.* **76,** 648–655 (1998).
11. Roelofs, H. *et al.* Restricted 12p amplification and RAS mutation in human germ cell tumors of the adult testis. *Am. J. Pathol.* **157,** 1155–1166 (2000).
12. Summersgill, B. *et al.* Molecular cytogenetic analysis of adult testicular germ cell tumours and identification of regions of consensus copy number change. *Br. J. Cancer* **77,** 305–313 (1998).
13. Zafarana, G. *et al.* 12p-amplicon structure analysis in testicular germ cell tumors of adolescents and adults by array CGH. *Oncogene* **22,** 7695–7701 (2003).
14. Rodriguez, S. *et al.* Expression profile of genes from 12p in testicular germ cell tumors of adolescents and adults associated with i(12p) and amplification at 12p11.2-p12.1. *Oncogene* **22,** 1880–1891 (2003).
15. McIntyre, A. *et al.* Amplification and overexpression of the KIT gene is associated with progression in the seminoma subtype of testicular germ cell tumors of adolescents and adults. *Cancer Res.* **65,** 8085–8089 (2005).
16. Kemmer, K. *et al.* KIT mutations are common in testicular seminomas. *Am. J. Pathol.* **164,** 305–313 (2004).
17. Bignell, G. *et al.* Sequence analysis of the protein kinase gene family in human testicular germ-cell tumors of adolescents and adults. *Genes Chromosomes Cancer* **45,** 42–46 (2006).
18. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
19. Kristensen, D. M. *et al.* Origin of pluripotent germ cell tumours: the role of microenvironment during embryonic development. *Mol. Cell Endocrinol.* **288,** 111–118 (2008).
20. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
21. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41,** e67 (2013).
22. Nakai, Y. *et al.* KIT (c-kit oncogene product) pathway is constitutively activated in human testicular germ cell tumors. *Biochem. Biophys. Res. Commun.* **337,** 289–296 (2005).
23. Ye, D. W. *et al.* P53 gene-mutations in chinese human testicular seminoma. *J. Urol.* **150,** 884–886 (1993).
24. Heidenreich, A. *et al.* Immunohistochemical and mutational analysis of the p53 tumour suppressor gene aml the bcl-2 oncogene in primary testicular germ cell, tumours. *APMIS* **106,** 90–99 (1998).
25. Laumann, R., Jucker, M. & Tesch, H. Point mutations in the conserved regions of the P53 tumor suppressor gene do not account for the transforming process in the jurkat acute lymphoblastic-leukemia T-cells. *Leukemia* **6,** 227–228 (1992).
26. Zhang, J. F., Wan, L. X., Dai, X. P., Sun, Y. & Wei, W. Y. Functional characterization of Anaphase Promoting Complex/Cyclosome (APC/C) E3 ubiquitin ligases in tumorigenesis. *Biochim. Biophys. Acta* **1845,** 277–293 (2014).
27. Kato, T. *et al.* Overexpression of CDC20 predicts poor prognosis in primary non-small cell lung cancer patients. *J. Surg. Oncol.* **106,** 423–430 (2012).
28. Chang, D. Z. *et al.* Increased CDC20 expression is associated with pancreatic ductal adenocarcinoma differentiation and progression. *J. Hematol. Oncol.* **5,** 15 (2012).
29. Marucci, G. *et al.* Gene expression profiling in glioblastoma and immunohistochemical evaluation of IGFBP-2 and CDC20. *Virchows Arch.* **453,** 599–609 (2008).
30. Pawar, S. A. *et al.* C/EBP delta targets cyclin D1 for proteasome-mediated degradation via induction of CDC27/APC3 expression. *Proc. Natl Acad. Sci. USA* **107,** 9210–9215 (2010).
31. Yu, C. *et al.* Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. *Cell Res.* **24,** 701–712 (2014).
32. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40,** e169 (2012).
33. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10,** 1081–1082 (2013).
34. Sathirapongsasuti, J. F. *et al.* Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **27,** 2648–2654 (2011).
35. Gilbert, D. C. *et al.* Minimum regions of genomic imbalance in stage I testicular embryonal carcinoma and association of 22q loss with relapse. *Genes Chromosomes Cancer* **50,** 186–195 (2011).
36. Summersgill, B., Osin, P. S., Lu, Y. J., Huddart, R. & Shipley, J. Chromosomal imbalances associated with carcinoma in situ and associated testicular germ cell tumours of adolescents and adults. *Br. J. Cancer* **85,** 213–219 (2001).
37. Chao, H. H., He, X., Parker, J. S., Zhao, W. & Perou, C. M. Micro-scale genomic DNA copy number aberrations as another means of mutagenesis in breast cancer. *PLoS ONE* **7,** e51719 (2012).
38. LeBron, C. *et al.* Genome-wide analysis of genetic alterations in testicular primary seminoma using high resolution single nucleotide polymorphism arrays. *Genomics* **97,** 341–349 (2011).
39. Brown, P. R., Miki, K., Harper, D. B. & Eddy, E. M. A-kinase anchoring protein 4 binding proteins in the fibrous sheath of the sperm flagellum. *Biol. Reprod.* **68,** 2241–2248 (2003).
40. Miki, K. *et al.* Targeted disruption of the Akap4 gene causes defects in sperm flagellum and motility. *Dev. Biol.* **248,** 331–342 (2002).
41. Coy, J. F., Dressler, D., Wilde, J. & Schubert, P. Mutations in the transketolase-like gene TKTL1: clinical implications for neurodegenerative diseases, diabetes and cancer. *Clin. Lab.* **51,** 257–273 (2005).
42. Schwaab, J. *et al.* Expression of Transketolase like gene 1 (TKTL1) predicts disease-free survival in patients with locally advanced rectal cancer receiving neoadjuvant chemoradiotherapy. *BMC Cancer* **11,** 363 (2011).
43. Danoy, P., Sonoda, E., Lathrop, M., Takeda, S. & Matsuda, F. A naturally occurring genetic variant of human XRCC2 (R188H) confers increased resistance to cisplatin-induced DNA damage. *Biochem. Biophys. Res. Commun.* **352,** 763–768 (2007).
44. Lin, W. Y. *et al.* A role for XRCC2 gene polymorphisms in breast cancer risk and survival. *J. Med. Genet.* **48,** 477–484 (2011).
45. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11,** 863–874 (2001).
46. Sunyaev, S. *et al.* Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10,** 591–597 (2001).
47. Poddar, A., Stukenberg, P. T. & Burke, D. J. Two complexes of spindle checkpoint proteins containing Cdc20 and Mad2 assemble during mitosis independently of the kinetochore in Saccharomyces cerevisiae. *Eukaryot. Cell* **4,** 867–878 (2005).
48. Monda, G., Ohashi, A., Yang, L., Rowley, M. & Couch, F. J. Tex14, a Plk1-regulated protein, is required for kinetochore-microtubule attachment and regulation of the spindle assembly checkpoint. *Mol. Cell* **45,** 680–695 (2012).
49. Ruark, E. *et al.* Identification of nine new susceptibility loci for testicular cancer, including variants near DAZL and PRDM14. *Nat. Genet.* **45,** 686–689 (2013).
50. Chung, C. C. *et al.* Meta-analysis identifies four new loci associated with testicular germ cell tumor. *Nat. Genet.* **45,** 680–685 (2013).
51. Rapley, E. A. *et al.* A genome-wide association study of testicular germ cell tumor. *Nat. Genet.* **41,** 807–810 (2009).
52. Chemes, H. E., Brugo, S., Zanchetti, F., Carrere, C. & Lavieri, J. C. Dysplasia of the fibrous sheath: an ultrastructural defect of human spermatozoa associated with sperm immotility and primary sterility. *Fertil. Steril.* **48,** 664–669 (1987).
53. Hotaling, J. M. & Walsh, T. J. Male infertility: a risk factor for testicular cancer. *Nat. Rev. Urol.* **6,** 550–556 (2009).
54. Nathanson, K. L. *et al.* The Y deletion gr/gr and susceptibility to testicular germ cell tumor. *Am. J. Hum. Genet.* **77,** 1034–1043 (2005).
55. Kuroda-Kawaguchi, T. *et al.* The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat. Genet.* **29,** 279–286 (2001).
56. Hemminki, K. & Li, X. Familial risk in testicular cancer as a clue to a heritable and environmental aetiology. *Br. J. Cancer* **90,** 1765–1770 (2004).
57. Hasle, H., Mellemgaard, A., Nielsen, J. & Hansen, J. Cancer incidence in men with Klinefelter syndrome. *Br. J. Cancer* **71,** 416–420 (1995).
58. Turnbull, C. *et al.* Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nat. Genet.* **42,** 604–607 (2010).
59. Gori, S. *et al.* Germ cell tumours of the testis. *Crit. Rev. Oncol. Hematol.* **53,** 141–164 (2005).
60. Mai, P. L. *et al.* The international testicular cancer linkage consortium: a clinicopathologic descriptive analysis of 461 familial malignant testicular germ cell tumor kindred. *Urol. Oncol.* **28,** 492–499 (2010).
61. Forman, D. *et al.* Familial testicular cancer: a report of the UK family register, estimation of risk and an HLA class 1 sib-pair analysis. *Br. J. Cancer* **65,** 255–262 (1992).

© 2015 Macmillan Publishers Limited. All rights reserved.

62. Usanova, S. *et al.* Cisplatin sensitivity of testis tumour cells is due to deficiency in interstrand-crosslink repair and low ERCC1-XPF expression. *Mol. Cancer* **9,** 248 (2010).
63. Cavallo, F. *et al.* Reduced proficiency in homologous recombination underlies the high sensitivity of embryonal carcinoma testicular germ cell tumors to cisplatin and poly (ADP-ribose) polymerase inhibition. *PLoS ONE* **7,** e51563 (2012).
64. Huddart, R. A., Wooster, R., Horwich, A. & Cooper, C. S. Microsatellite instability in human testicular germ cell tumours. *Br. J. Cancer* **72,** 642–645 (1995).
65. Orkin, S. Molecular-cloning—a laboratory manual, 2nd edn (Sambrook, J., Fritsch, E.F., Maniatis, T.) *Nature* **343,** 604–605 (1990).
66. Lahiri, D. K. & Nurnberger, J. I. A Rapid Nonenzymatic Method for the Preparation of Hmw DNA from Blood for Rflp Studies. *Nucleic Acids Res.* **19,** 5444 (1991).
67. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10,** 1081–1082 (2013).
68. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).

Acknowledgements

We thank the patients and their clinicians for participation in this study. We acknowledge the National Health Service funding to the National Institute for Health Research Biomedical Research Centre. We acknowledge the facilities and expertise of the Cancer Genetics Core Laboratory Facility and the Cancer Genetics Sequencing Facility made available at the Institute of Cancer Research by Professor Nazneen Rahman. This study was supported by the Movember foundation and the Institute of Cancer Research. K. Litchfield is supported by a PhD fellowship from Cancer Research UK. R.S.H. and P.B. are supported by Cancer Research UK (C1298/A8362 Bobby Moore Fund for Cancer Research UK).

Author contributions

C.T. designed the study. J.S. and R.A.H. provided the samples. D.D., B.S. and R.A.-S. coordinated sample administration and tracking. K. Litchfield and S.S coordinated sample management. J.S and B.S. provided CGH validation data. C.T., R.S.H., P.B. A.R. and K. Litchfield designed laboratory experiments. K. Litchfield and A.R conducted laboratory experiments. K. Litchfield, R.S.H., K.L., N.C.T., S.Y. and R.S. designed bioinformatic analyses. K. Litchfield, S.Y. and R.S. carried out bioinformatics analyses. K. Litchfield performed statistical analyses. K. Litchfield drafted the manuscript with assistance from C.T., R.S.H. and J.S. All authors reviewed and contributed to the manuscript.

Additional information

Accession codes: Whole-exome sequencing have been deposited in the European Genome–phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI), under the accession code EGAS00001001084.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Litchfield, K. *et al.* Whole-exome sequencing reveals the mutational spectrum of testicular germ cell tumours. *Nat. Commun.* 6:5973 doi: 10.1038/ncomms6973 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© 2015 Macmillan Publishers Limited. All rights reserved.

Mutational analysis of primary central nervous system lymphoma

Aurélie Bruno[1,2,3,4], Blandine Boisselier[1,2,3,4,5], Karim Labreche[1,2,3,4], Yannick Marie[5,6], Marc Polivka[7], Anne Jouvet[8], Clovis Adam[9], Dominique Figarella-Branger[10], Catherine Miquel[11], Sandrine Eimer[12], Caroline Houillier[13], Carole Soussain[14], Karima Mokhtari[1,2,3,4,6] Romain Daveau[15] and Khê Hoang-Xuan[1,2,3,4,13]

[1] Sorbonne Universités, UPMC Univ Paris 06, UM 75, ICM, F-75013 Paris, France

[2] Institut National de la Santé et de la Recherche Médicale, U1127, ICM, Paris, F-75013 Paris, France

[3] Centre National de la Recherche Scientifique, UMR 7225, ICM, Paris, F-75013 Paris, France

[4] ICM, Paris, 75013 France

[5] Plateforme de Génotypage Séquençage, ICM, F-75013, Paris, France

[6] Onconeurothèque, Groupe Hospitalier Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, Paris, France

[7] Centre Hospitalier Universitaire Lariboisière, Assistance Publique-Hôpitaux de Paris, Service d'Anatomopathologie, Paris, France

[8] Hospices Civils de Lyon, Hôpital Neurologique, Bron, France and Université Lyon 1, Institut National de la Santé et de la Recherche Médicale Unité 842, Lyon, France

[9] Centre Hospitalier Universitaire Bicêtre, Assistance Publique-Hôpitaux de Paris, Service d'anatomopathologie, Bicêtre, France

[10] Centre Hospitalier Universitaire La Timone, Assistance Publique-Hôpitaux de Marseille, Institut National de la Santé et de la Recherche Médicale Unité 911, Centre de Recherches en Oncologie biologique et Onco-pharmacologie, Université de la Méditerranée and Tumorothèque de l'Assistance Publique-Hôpitaux de Marseille (AC 2013-1786), Marseille, France

[11] Centre hospitalier Sainte Anne, Université Paris Descartes, Sorbonne Paris Cité, Paris, France

[12] Service de Pathologie, CRB Tumorothèque, Centre Hospitalier Universitaire Bordeaux, Bordeaux, France

[13] Assistance Publique-Hôpitaux de Paris, Hôpital de la Pitié-Salpêtrière, Service de Neurologie 2-Mazarin, Paris, France; and the LOC network (INCa)

[14] Hôpital René Huguenin, Institut Curie, Service d'Hématologie, Saint Cloud, France; and the LOC network (INCa)

[15] Institut National de la Santé et de la Recherche Médicale Unité 830, Génétique et Biologie des Cancers, Institut Curie, Paris, France

Correspondence to: *Khê Hoang-Xuan,* ***email***: *khe.hoang-xuan@psl.aphp.fr*

Keywords: Primary CNS lymphoma, exome sequencing, somatic mutations, NFKB, B cell differentiation

Received: *April 21, 2014* ***Accepted***: *June 7, 2014* ***Published***: *June 8, 2014*

This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ABSTRACT

Little is known about the genomic basis of primary central nervous system lymphoma (PCNSL) tumorigenesis. To investigate the mutational profile of PCNSL, we analyzed nine paired tumor and germline DNA samples from PCNSL patients by high throughput exome sequencing. Eight genes of interest have been further investigated by focused resequencing in 28 additional PCNSL tumors to better estimate their incidence. Our study identified recurrent somatic mutations in 37 genes, some involved in key signaling pathways such as NFKB, B cell differentiation and cell cycle control. Focused resequencing in the larger cohort revealed high mutation rates for genes already described as mutated in PCNSL such as *MYD88* (38%), *CD79B* (30%), *PIM1* (22%) and *TBL1XR1* (19%) and for genes not previously reported to be involved in PCNSL tumorigenesis such as *ETV6* (16%), *IRF4* (14%), *IRF2BP2* (11%) and *EBF1* (11%). Of note, only 3 somatically acquired SNVs were annotated in the COSMIC database. Our results demonstrate a high genetic heterogeneity of PCNSL

and mutational pattern similarities with extracerebral diffuse large B cell lymphomas, particularly of the activated B-cell (ABC) subtype, suggesting shared underlying biological mechanisms. The present study provides new insights into the mutational profile of PCNSL and potential targets for therapeutic strategies.

INTRODUCTION

Primary central nervous system lymphoma (PCNSL) represents a rare subgroup of diffuse large B-cell lymphoma (DLBCL) that arises in the brain, eyes, meninges or spinal cord, accounting for up to 5% of primary malignant brain tumors and 1% of non-Hodgkin's lymphomas (NHL) in adults. Despite the application of intensive treatment including high-dose methotrexate based poly-chemotherapy with or without whole brain radiotherapy, the median overall survival ranges from 2 to 4 years with a poorer prognosis than extracerebral DLBCL [1]. The pathogenesis of PCNSL remains largely unclear, which is partly due to the rarity of the tumor tissue available for research studies. Transcriptomic studies have identified deregulated genes involved in the IL4/JAK/STAT6, cell adhesion-related, unfolded protein response (UPR) and apoptosis signaling pathways [2–5]. Copy number variation studies [4, 6–8] have revealed frequent chromosome losses affecting the 6q, 6p21.32 and 9p21 regions. However, the mutational landscape of PCNSL is still poorly known. A whole exome sequencing strategy has successfully identified pivotal gene mutations in several hematologic and brain malignancies [9, 10]. In a previous study, we have reported preliminary results based on four PCNSL cases investigated by this technique and identified recurrent mutations in *MYD88* and *TBL1XR1* [8]. Here, we have expanded our series and we present the results of nine paired germline and tumour samples, allowing for the identification of recurrent gene mutations that have not yet been reported in PCNSL. We confirmed the most relevant mutations and genes in a validation set of 28 PCNSL cases.

RESULTS

Mutational pattern of PCNSL revealed by whole exome sequencing

To investigate the mutational profile of PCNSL, we performed high throughput exome sequencing on 9 cases. DNA from case-matched blood was also sequenced to screen out germline polymorphisms. On average, 9.8e7 (8.1e7-1.4e8) 75-bp paired reads were sequenced per sample, 5.8e7 (3.8e7-8.3e7) of these were specifically positioned onto the human reference exome (as defined by the Agilent SureSelect 50 Mb probes) after the removal of both low-quality mapped reads and potential PCR-derived duplicates (Supplementary Figure 1). This provided 76% (64-86) coverage over the targeted regions at a minimum depth of 20X (Supplementary Figure 2), wherein 82% of the bases were suitable for variant detection. Across the coding regions of the 9 matched tumor and germline pairs we investigated, we detected 17e3 (15e3-19e3) SNVs and 226 (176-263) indels. A total of 25e3 (20e3-32e3) SNVs and 21e2 (18e2-27e2) indels were also called outside of the targeted exons, but those primarily fell into neighboring introns and, in most cases, were already described as known polymorphisms. To assess the quality of our calls, we reviewed population-scale variant distributions from the 1000-genomes project and found no difference with either paired germline or PCNSL samples when considering all high-quality called SNVs and comparing the (i) transition to transversion rates, (ii) mutational spectrum and (iii) variant annotation (Supplementary Figure 3). Then, we focused on somatic SNVs identified in tumor DNA and not present in germline DNA (Figure 1). On average, we identified 220 (126-358) somatically acquired point mutations per sample and no hypermutated tumors were found. Among them, 62 (26-101) and 143 (89-231) were synonymous and non-synonymous, respectively. The non-synonymous to synonymous ratio was thus 2.4 (1.8-3.4), and there was a non-silent mutation rate of 2.9 (1.8-4.6) per Mb, the latter being lower than previously published estimations in DLBCL [11]. Half of those non-synonymous SNVs, i.e., 74 (49-111) were predicted as functionally deleterious in the dbNSFP database [12, 13]. Transitions accounted for 68% of somatic events (Figure 1B) similar to pattern observed in DLBCL [14, 15]. To confirm the depth at the somatic mutation sites, reads aligned at these genomic positions were visualized using IGV software (Broad Institute).

Identification of 37 genes recurrently affected by somatic non-synonymous mutations

Only SNVs located within coding regions were considered. After having removed germline variations, synonymous SNVs, indels and known polymorphisms, we identified 37 genes, harboring 142 somatically point mutations (Supplementary Table 1), that were mutated in at least 2 patients. Among these 142 mutations, 133 led to an amino acid exchange while the remaining nine led to the gain or loss of a stop codon. These 37 recurrently mutated genes were prioritized based on (i) the number of mutated tumors, (ii) the prediction of the functional impact

Table 1: Prioritization of the 37 genes of interest identified by whole exome sequencing. The present study identified 37 genes affected by non-synonymous somatic SNVs in at least 2 of the 9 patients of the discovery set. In this table are listed all these genes prioritized according to (i) number of mutated patients, (ii) functional impact prediction (FISM), (iii) number of mutations per gene. Functional impact prediction columns indicate the number of patients harboring at least one mutation for each FISM category (1 corresponds to the highest impact). Of importance, number of mutated patients and number of mutations per gene take into account all somatic mutations identified by exome sequencing before any attempt of validation.

				Functional prediction impact (FISM)						
Genes	Chromosome	Mutations	Patients	NA	≥0.5	≥0.6	≥0.7	≥0.8	≥0.9	=1
PIM1	6	32	8	0	8	8	7	6	6	5
IGLL5	22	12	6	6	0	0	0	0	0	0
MYD88	3	2	5	5	0	0	0	0	0	0
TBL1XR1	3	4	4	0	4	4	4	4	4	3
CSMD3	8	4	4	0	4	4	4	3	3	1
CD79B	17	3	3	0	2	2	2	2	2	1
HIST1H2AC	6	8	3	0	3	3	3	3	1	1
ETV6	12	5	3	0	3	3	2	2	1	1
KLHL14	18	7	2	0	2	2	2	2	2	2
IRF4	6	3	2	0	2	2	2	2	2	2
PRKCD	3	2	2	0	2	2	2	2	2	2
ABCC8	11	2	2	0	2	2	2	2	2	1
ZFHX4	8	2	2	0	2	2	2	2	2	1
SALL3	18	2	2	0	2	2	2	1	1	1
IRF2BP2	1	3	2	0	2	2	1	1	1	1
CD37	19	2	2	0	2	2	1	1	1	1
OSBPL10	3	7	2	0	2	2	2	2	2	0
EBF1	5	3	2	0	2	2	2	2	2	0
DST	6	2	2	0	2	2	2	2	1	0
MIF4GD	17	2	2	0	2	2	2	2	1	0
HIST1H1D	6	3	2	0	2	2	2	1	1	0
BTG1	12	2	2	0	2	2	2	1	1	0
MEP1B	18	2	2	0	2	2	2	1	1	0
THBS4	5	2	2	0	2	2	2	1	1	0
ADAMTS5	21	2	2	0	2	2	1	1	1	0
HIST1H1E	6	2	2	0	2	1	1	1	1	0
MPEG1	11	3	2	1	1	1	1	1	1	0
OBSCN	1	2	2	0	2	2	2	2	0	0
C10orf71	10	2	2	0	2	2	2	1	0	0
HMCN1	1	2	2	0	2	2	2	1	0	0
MYH4	17	2	2	0	2	2	1	1	0	0
TBC1D4	13	2	2	0	2	1	1	1	0	0
SLC2A12	6	2	2	0	2	2	1	0	0	0
ETS1	11	2	2	0	2	2	0	0	0	0
MUC16	19	2	2	2	0	0	0	0	0	0
UNC80	2	2	2	2	0	0	0	0	0	0
ACTG1	17	1	2	2	0	0	0	0	0	0

and (iii) the number of SNVs per gene (Table 1). Then, somatic mutations were verified by Sanger sequencing on tumor and germline DNA. For *PIM1* and *MYD88* genes, only "hot spot" mutations E226K and L265P, respectively, were validated. To better understand the biological processes that are potentially altered by somatic mutations, we used gene ontology [16] annotations for these 37 genes. This functional categorization highlighted the variability of the biological processes that are altered in PCNSL (Figure 2), including transcription (e.g., *ETV6*, *IRF2BP2*, *EBF1*, *IRF4*, *TBL1XR1*), cell cycle (e.g., *PIM1*, *BTG1*), nucleosome assembly (e.g., *HIST1H1D*, *HIST1H2AC*) and cell adhesion (e.g., *MUC16*, *ACTG1*). In terms of signaling pathways, we identified mutations in the genes involved in the NFKB, WNT and B-cell or T-cell receptor signaling pathways.

Analysis of 8 relevant genes in an independent series of 28 PCNSL

In order to specify their mutation frequency in PCNSL, we selected 8 genes for further investigation in an independent validation panel of PCNSL tumors (n=28). This selection was based both on high mutation rate in our discovery set and biological relevance. *PIM1*, *TBL1XR1*, *ETV6*, *IRF4*, *IRF2BP2* and *EBF1* were resequenced for their coding exons by pyrosequencing. We identified 133 variations, including 122 SNVs and 11 deletions. Among them, 39 variations were missense mutations (Supplementary Table 2), including 35 variations that were not previously described in the dbSNP database as known polymorphisms. For each missense SNV, functional impact was predicted using SIFT or Polyphen2 tools and identified 11 SNVs with putative damaging consequences predicted by both softwares. Twenty-five out of the 35 missense mutations were validated by Sanger sequencing and corresponded to 22 SNVs and 3 frameshift deletions. The somatic state of the validated mutations was verified with direct sequencing. Considering the whole cohort, including the discovery and the validation sets, somatic variations were found in 22% (8/37) of the PCNSL cases for *TBL1XR1*, 19% (7/37) for *PIM1*, 16% (6/37) for *ETV6*, 14% (5/37) for *IRF2BP2* and 11% (4/37) for *IRF4* and *EBF1* each (Fig 3A). Of note, 3 non-sense mutations affecting *ETV6* and *IRF2BP2* genes and 3 deletions leading to a frameshift in *TBL1XR1*, *ETV6* and *EBF1* were observed (Figure 3B). Somatic mutations on the hot spots L265P of *MYD88* and Y196 of *CD79B* were already referenced in the COSMIC database. One somatic mutation within the *PIM1* gene was also identified in this database (e.g., COSM220740) as reported in DLBCL cases [9, 14]. Four other somatic mutations identified within the *PIM1*, *ETV6* and *IRF4* genes in this study occur in the same codon as the alterations that are mainly reported in hematopoietic or lymphoid malignancies.

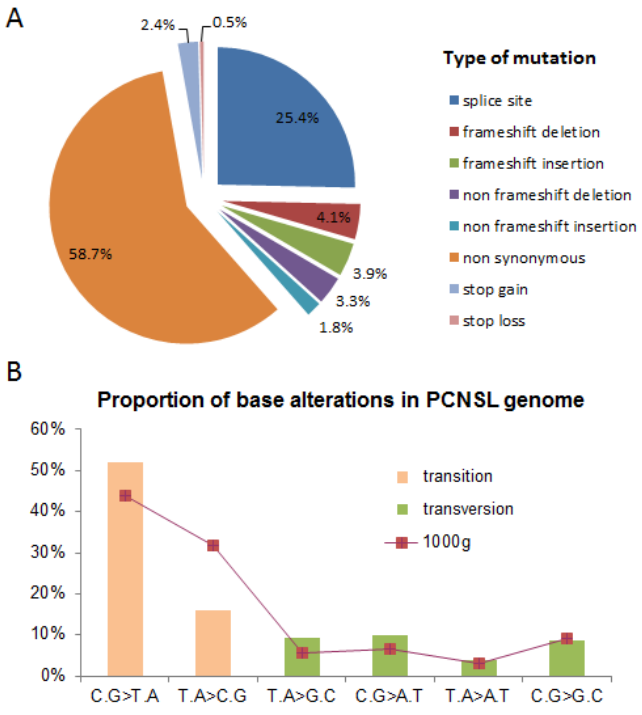


Figure 1: Mutation pattern of PCNSL samples investigated by whole exome sequencing. To address PCNSL mutational profile, whole exome sequencing was conducted in nine paired blood and tumor samples. (A) Pie chart represents relative distribution of somatically acquired mutations classified according to their type. (B) Histogram depicts the proportion of PCNSL somatic mutations in each mutational class of transitions and transversions compared with 1000-genomes project data (red line).

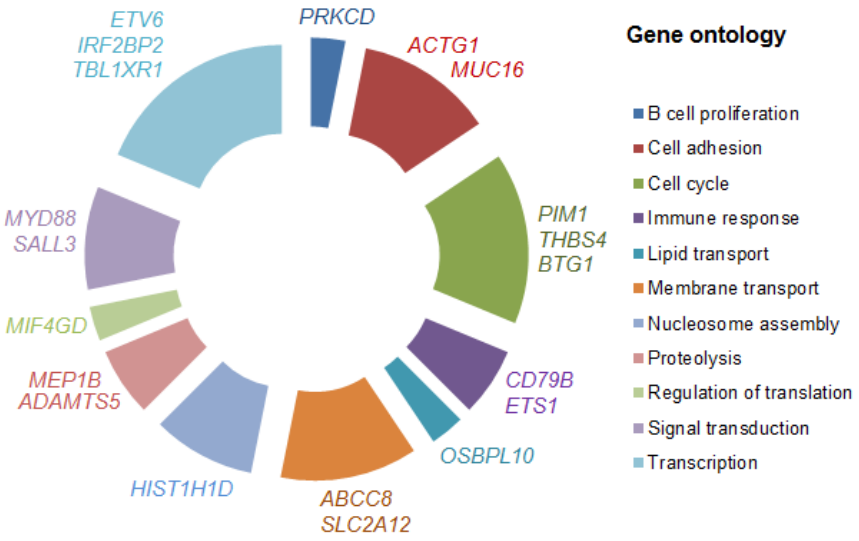


Figure 2: Gene ontology of PCNSL genes. Relative distribution of the 37 genes somatically mutated in PCNSL by gene ontology categories. The spans of the arcs indicate the relative numbers of genes annotated with respect to gene ontology terms. Representative genes in each category are shown next to each arc.

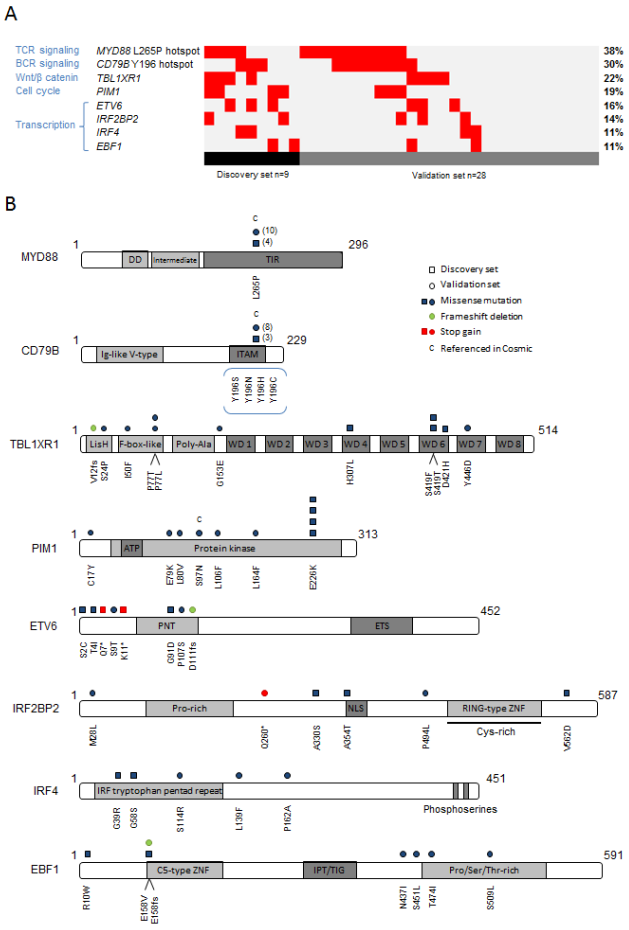


Figure 3: Investigation of 8 relevant genes recurrently affected by point mutations in PCNSL. Based on genes identified by whole exome sequencing, we selected 8 relevant genes to be sequenced in a larger cohort: *CD79B*, *EBF1*, *ETV6*, *IRF4*, *IRF2BP2*, *MYD88*, *PIM1* and *TBL1XR1*. (A) Repartition of validated mutations by gene within the whole population of 37 PCNSL cases. (B) Schematic representation of all validated mutations identified in the discovery (□) and the validation sets (○) with their position according to protein domains. Symbol color indicates mutation type. Number of □ or ○ indicates the number of mutated patients except for L265P *MYD88* and Y196 *CD79B* mutations.

Direct sequencing of *MYD88* and *CD79B* focused on the hot spot mutations identified in the discovery panel; the L265P mutation was found in 4/9 cases, and Y196 mutations were found in 3/9 cases. In the validation panel, 10 additional patients harbored the *MYD88* L265P mutation and 8 additional cases harbored *CD79B* Y196 mutations. Considering the whole population, *MYD88* L265P and *CD79B* Y196 mutations were identified in 38% (14/37) and 30% (11/37) of PCNSL tumors, respectively (Figure 3A), representing the most recurrently mutated genes in our series.

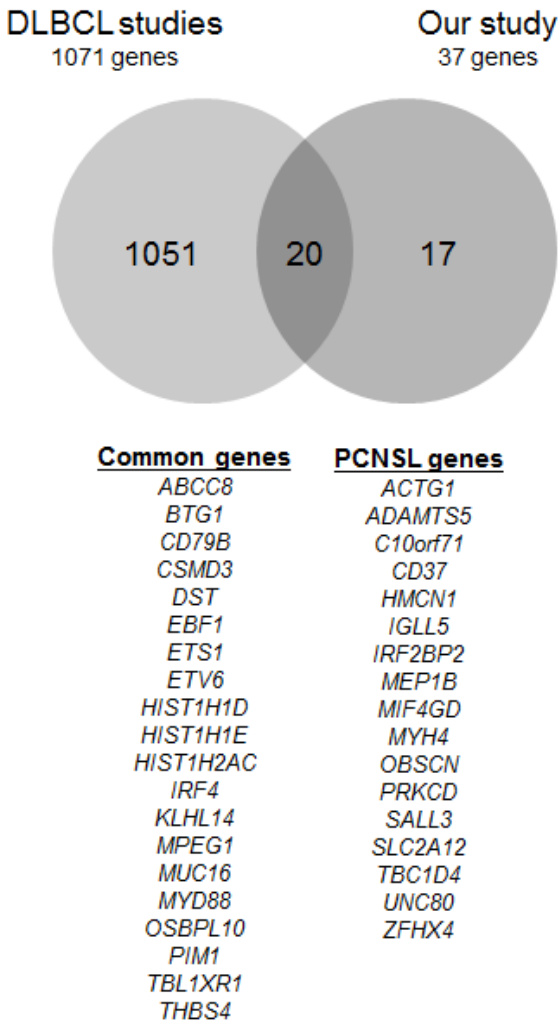


Figure 4: Overlaps in genes discovered in DLBCL studies and our 37 genes of interest. The Venn diagram depicts the comparison between gene mutations from the five DLBCL exomes studies and the present PCNSL study. The gene lists used were as follows: Lohr et al. (Table 1 in Ref. 11, n=72 genes), Pasqualucci et al. (Table S3 and Fig. S4 in Ref. 15, n=108 validated somatic genes), Zhang et al. (Table S3 in Ref. 14, n=322 genes), Morin et al. (in Ref. 9, n=315 known and confirmed somatic genes; Table S3 in Ref. 37, n=588 known and confirmed somatic genes).

DISCUSSION

The present study investigated the coding genomes of PCNSL in order to provide information on the mutational landscape of these tumors. We described an overview of the genes that are recurrently mutated in PCNSL, including (i) genes previously known to be mutated in PCNSL, such as *MYD88*, *CD79B*, *PIM1* and *TBL1XR1*; (ii) genes altered by somatic mutations in other B cell malignancies that have not yet been reported in PCNSL, such as *ETV6*, *IRF4* or *EBF1*; and (iii) genes that are altered in solid tumors, such as *IRF2BP2*. These results reveal the genetic heterogeneity of this disease and highlight the major signaling pathways that are deregulated in PCNSL.

In our series, genes coding for nuclear factor-κB (NFκB) pathway regulators (i.e., *MYD88*, *CD79B* and *TBL1XR1*) represented the most frequently altered genes. *MYD88* encodes a signaling adaptor protein that induces NFκB and JAK/STAT3 pathway activation after the stimulation of the Toll-like and IL1/IL18 receptors as well as interferon β production [17, 18]. *CD79B* encodes a B-cell receptor (BCR) subunit that is essential for BCR signaling, leading to NFκB activation [19]. We identified *MYD88* L265P and *CD79B* Y196 hot spot mutations in 38% and 30% of the PCNSL patients, respectively. We confirm and expand the results of Montesinos-Rongen et al [20, 21] who have recently investigated PCNSL for mutations in several genes involved in the BCR signaling cascade and reported a 36% (7/14) and 20% (5/25) mutation rate in *MYD88* and *CD79B*, respectively. These two hot-spot mutations have been described as oncogenic activating alterations leading to constitutive NFκB activation in DLBCL [22, 23]. Additionally, we found a significant association (p=0.0044, Chi-square test) between the *MYD88* L265P and *CD79B* Y196 mutations, suggesting collaborative effects of the NFκB activating pathways in PCNSL. The *TBL1XR1* gene, which encodes for a transcriptional regulator involved both in the Wnt/B catenin [24, 25] and NFκB pathways [26], was mutated in 22% of our PCNSL cohort. The *TBL1XR1* mutation rate in our series and the recurrent deletions of 3q26.32 (*TBL1XR1* locus) reported in PCNSL [7], extracerebral DLBCL [15], and acute lymphoblastic leukemia [27, 28] suggest its potential role as a tumor suppressor. Taken together, mutations in *MYD88*, *CD79B* and *TBL1XR1* affected 54% (20/37) of our cohort, suggesting that NFκB pathway deregulation is a driving mechanism in PCNSL tumorigenesis. Other genes, such as *CARD11* and *TNFAIP3*, which belong to this pathway are also reported to be mutated at lower rates in 16% and 3% of PCNSL, respectively [29].

A second set of alterations was detected in genes involved in B-cell proliferation and differentiation, such as *ETV6*, *EBF1*, *IRF4* and *ETS1*. To our knowledge these gene mutations have never been reported in

PCNSL. The *ETV6* tumor suppressor gene encodes an Ets family transcriptional repressor factor required for hematopoeisis [30] and largely described as a partner of gene translocation in lymphoid and myeloid hematopoietic tumors [31]. In our series, we found *ETV6* mutations in 16% of cases, including 2 cases with non-sense mutations. In line with this, several studies have reported heterozygous and homozygous deletions of 12p13.2 corresponding to the *ETV6* locus (15% in the present series) in PCNSL [6, 7]. *IRF4*, also known as *MUM1* encoding a lymphocyte-specific transcription factor [32], and *EBF1*, encoding an activator of transcription involved in lymphoid development [33], were found to be mutated in 11% of our cohort. We also reported, in our discovery set, 2 somatic mutations affecting *ETS1*, encoding another Ets family transcription factor involved in the negative regulation of plasmocytic differentiation [34]. A variety of *ETS1* alterations, including deletions [35] or gains [36] and somatic mutations [9, 37], have been reported in B cell malignancies. Finally, 11 tumors from our 37 samples (30%) harbored one or more mutation of genes involved in B cell proliferation and differentiation, supporting the role of B lymphoid development deregulation in PCNSL tumorigenesis.

A hallmark of oncogenesis is the alteration of genes controlling the cell cycle. We and others have previously identified *CDKN2A* homozygous deletions as a frequent alteration in PCNSL [4, 6–8] with an unfavorable impact on the prognosis [8]. In the present study, we found recurrent mutations in cell cycle regulator genes such as *PIM1* [38] (7/37; 19%), *IRF2BP2* [39] (5/37; 14%) and *BTG1* [40] (2/9). *PIM1* is a proto-oncogene that encodes a serine/threonine kinase and is known to be frequently targeted by somatic hypermutation in PCNSL [41]. Of note, 6 of the 7 seven mutations identified on *PIM1* in the present study were located on the protein kinase domain. A variety of inhibitors are currently under development for PIM family proteins [42](Tab2), rendering these proteins attractive targets for therapy [5]. *IRF2BP2* encodes a zinc finger protein that interacts with partners such as *TP53* and the oncogene *IRF2*. *IRF2BP2* acts as a repressor of *IRF2*, leading to the inhibition of interferon responsive gene expression and *NFAT1*, which is involved in the cell cycle. Recently, a novel fusion between *IRF2BP2* and the *CDX1* homeobox gene was described in a patient suffering from a mesenchymal chondrosarcoma [43]. Intriguingly, the patient also had PCNSL; unfortunately the brain tumor tissue was not investigated.

Our results revealed many similarities between genomic abnormalities of extracerebral DLBCL and PCNSL. Indeed, among the 37 genes of interest identified in this study, 20 have described mutations in DLBCL exome studies [9, 11, 14, 15, 37] (Figure 4). More specifically, mutations in the genes involved in the NFκB signaling pathway and in *PIM1*, as observed in PCNSL, are likely associated with the activated B-cell like (ABC) subtype of DLBCL. In contrast, histone-modifying genes, such as *CREBBP*, *EZH2* and *MLL2*, which are recurrently altered in the germinal center B-cell like (GCB) subtype of DLBCL [9, 14, 15], were not found in our series. These observations are in agreement with previous studies showing that the PCNSL gene expression profile is more closely related to post-GCB and ABC cells than to GCB cells [2, 44].

The present study has several limitations. Even if the small number of cases analyzed is generally acceptable given the rarity of the disease and small amount of available tissue, it provides a limited power of analysis and we likely underestimate the PCNSL gene mutations. In addition, the sequencing methods used do not investigate noncoding portions of the genome. Altogether, this could explain the relatively low overlap with a recent study of the Mayo Clinic including 10 PCNSL investigated by whole exome sequencing (O'Neill BP *et al.*, 2013, ASH Annual Meeting Abstract). Alternatively, these results could also illustrate a high molecular heterogeneity within PCNSL as observed in extracerebral DLBCL exome studies [14]. However, our results contribute to the description of the PCNSL mutational landscape and provide insights into the prominent signaling pathways that are disrupted in PCNSL tumorigenesis. Genomic similarities with the ABC subtype of extracerebral DLBCL may open the possibility for parallels in therapeutic strategies of both lymphomas. For example, lenalidomide which induces IRF4 levels decrease [45], and ibrutinib which targets B-cell receptor signaling (Wilson WH *et al.*, 2012, ASH Annual Meeting Abstract) have shown promising results in extracerebral ABC-DLBCL. In this setting, they might also be attractive therapeutic strategies for PCNSL.

METHODS

PCNSL sample selection and patient characteristics

Thirty-seven PCNSL patients were selected for the present study. All tumors were classified as CD20+ DLBCL according to the WHO classification and demonstrated to contain at least 90% tumor cells based on morphology and immunohistochemistry. All the patients were newly diagnosed and immunocompetent. The participants provided written consent for sample collection and genetic analysis. This study was approved by the local ethical committee (CPPRB Pitié-Salpêtrière). Based on the high quality and sufficient levels of DNA, nine paired frozen tumor and blood tissues were selected to constitute the discovery set investigated by whole exome sequencing, and 28 tumor samples constituted the validation set investigated by direct sequencing. The sex ratio was 1.18 (male/female) and the median age at

diagnosis was 61 years, ranging from 17 to 83.

Isolation and quality assessment of DNA

Tumor DNA from 34 cryopreserved and 3 FFPE samples was extracted using the QIAamp DNA Mini Kit (Qiagen) and iPrep™ ChargeSwitch® Forensic Kit (Life Technologies), respectively, according to the manufacturer's instructions. A conventional saline method was used for the extraction of germline DNA from the blood samples. DNA was quantified using a NanoDrop spectrophotometer, and the quality was assessed on a 1% agarose gel.

Whole exome sequencing

Whole exome sequencing was possible for PCNSL patients with available paired frozen tumor and blood samples and with a minimal amount of 5 µg of tumor and germline DNA. Genomic DNA capture was performed using biotinylated oligonucleotides probes library (Human All Exon v2 – 46 Mb, Agilent) according to Agilent in-solution enrichment methodology (SureSelect Human All Exon Kits Version 2, Agilent). Sequence capture, enrichment and elution were performed according to manufacturer's instructions and protocols (SureSelect, Agilent). Massively parallel sequencing was realized on an Illumina GAIIX as paired-end 75 b reads.

Mapping and variant calling

Mapping of high-quality paired-end sequenced reads onto the GRCh37 build of the human reference genome was performed by Integragen using the Illumina ELAND 2 software tool. Raw alignments were first filtered for both low-quality mapped reads and assumed PCR duplicates with the SAMtools view (-q 20) and the Picard MarkDuplicates utilities, respectively [46]. The resulting filtered BAM files were subsequently confined to the genomic coordinates delineating the Agilent SureSelect 50-Mb probes using the intersectBed command of the BEDtools suite [47]. A commonly used combination of SAMtools mpileup and BCFtools view was then applied to the latter bounded alignments in order to call single nucleotide variations (SNVs) as well as short insertions and deletions (indels) within the targeted genomic regions. Mapping and coverage summary statistics were additionally obtained by an in-house post-processing of SAMtools idxstats and mpileup outputs.

Annotating called variants

Variant annotation was performed with the unpublished Genomic and Functional Annotation Pipeline (GFAP) software, developed and routinely used at Institut Curie (http://gfap.curie.fr/). Briefly, GFAP consists of a set of tools that automatically: (i) retrieve and store suitable information from public variant databases such as 1000-genomes [48], dbSNP [49] or COSMIC [50, 51], (ii) match submitted variants against built-in databases and annotate them with respect to their genomic localization, (iii) assign an integrated functional impact prediction to non-synonymous variants (including stop-gains and losses) using dbNSFP database [12, 13] which compiles several tools such as SIFT [52] or Polyphen2 [53].

Validation set

Samples were selected based on the availability of tumor DNA. The validation set was investigated for known hotspot mutations by Sanger sequencing and for all exons of highly mutated genes by pyrosequencing. The tumor DNA was amplified using the primers listed in Supplementary Table 3. The amplification conditions were 94°C for 3 min followed by 45 cycles of 94°Cx15 sec, 60°Cx45 sec and 72°Cx1 min, with a final step at 72°C for 8 min. Exon 13 of *TBL1XR1* was amplified using Touch Down PCR with a gradient from 62 to 55°C during 6 cycles followed by 30 cycles at 55°C for primer annealing. The PCR products were purified according to the Agencourt® AMPure® XP PCR purification protocol (Beckman Coulter) with the Biomek® 3000 Automation Workstation.

Sanger sequencing

Sequencing reactions were performed in both orientations using the Big-Dye® Terminator Cycle Sequencing Ready Reaction (Perkin Elmer). The extension products were purified with the Agencourt® CleanSEQ® protocol according to the manufacturer's instructions (Beckman Coulter). The purified sequences were analyzed on an ABI Prism 3730 DNA Analyzer (Applied Biosystems). The forward and reverse sequences were visualized using Chromas Lite software.

Pyrosequencing

The universal tailed amplicon resequencing approach (454 Sequencing Technology, Roche) was used for coding exons sequencing. This system employs a second PCR, aiming MID (multiplex identifier) and 454 adaptors incorporation, an emulsion PCR according to the emPCR Amplification Method Manual Lib-A protocol (GS Junior Titanium Series, Roche), enrichment and pyrosequencing according to the Sequencing Method Manual (Roche). Sequences analysis was performed using CLC Genomics Workbench software.

ACKNOWLEDGMENTS

This work is part of the national program Cartes d'Identité des Tumeurs ® (CIT) http://cit.ligue-cancer.net/ funded and developed by the Ligue nationale contre le cancer. This study benefited from the LOC study group network (réseau national de centres experts des lymphomes primitifs du SNC, INCa). The research leading to these results has received funding from the program "Investissements d'avenir" ANR-10-IAIHU-06 and from the Association pour la recherche sur les tumeurs cérébrales (ARTC). The biological resource centers of CHU Bordeaux and of Groupe Hospitalier Pitié Salpêtrière took part to samples collection.

Conflict of interest

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

A.B., B.B., K.L. R.D., Y.M., K.H-X. designed and performed research, analyzed data and wrote the paper. K.M., M.P., A.J., D.F-B, C.A., C.M., S.E., C.H., C.S. contributed to collect biological tissue and analyzed data.

REFERENCES

1. Ricard D, Idbaih A, Ducray F, Lahutte M, Hoang-Xuan K, Delattre J-Y. Primary brain tumours in adults. Lancet. 2012; 379: 1984-1996.
2. Montesinos-Rongen M, Brunn A, Bentink S, Basso K, Lim WK, Klapper W, Schaller C, Reifenberger G, Rubenstein J, Wiestler OD, Spang R, Dalla-Favera R, Siebert R, et al. Gene expression profiling suggests primary central nervous system lymphomas to be derived from a late germinal center B cell. Leukemia. 2008; 22: 400-405.
3. Tun HW, Personett D, Baskerville KA, Menke DM, Jaeckle KA, Kreinest P, Edenfield B, Zubair AC, O'Neill BP, Lai WR, Park PJ, McKinney M. Pathway analysis of primary central nervous system lymphoma. Blood. 2008; 111: 3200-3210.
4. Sung CO, Kim SC, Karnan S, Karube K, Shin HJ, Nam D-H, Suh Y-L, Kim S-H, Kim JY, Kim SJ, Kim WS, Seto M, Ko Y-H. Genomic profiling combined with gene expression profiling in primary central nervous system lymphoma. Blood. 2011; 117: 1291-1300.
5. Rubenstein JL, Fridlyand J, Shen A, Aldape K, Ginzinger D, Batchelor T, Treseler P, Berger M, McDermott M, Prados M, Karch J, Okada C, Hyun W, et al. Gene expression and angiotropism in primary CNS lymphoma. Blood. 2006; 107: 3716-3723.
6. Schwindt H, Vater I, Kreuz M, Montesinos-Rongen M, Brunn A, Richter J, Gesk S, Ammerpohl O, Wiestler OD, Hasenclever D, Deckert M, Siebert R. Chromosomal imbalances and partial uniparental disomies in primary central nervous system lymphoma. Leukemia. 2009; 23: 1875-1884.
7. Braggio E, McPhail ER, Macon W, Lopes MB, Schiff D, Law M, Fink S, Sprau D, Giannini C, Dogan A, Fonseca R, O'Neill BP. Primary central nervous system lymphomas: a validation study of array-based comparative genomic hybridization in formalin-fixed paraffin-embedded tumor specimens. Clin Cancer Res. 2011; 17: 4245-4253.
8. Gonzalez-Aguilar A, Idbaih A, Boisselier B, Habbita N, Rossetto M, Laurenge A, Bruno A, Jouvet A, Polivka M, Adam C, Figarella-Branger D, Miquel C, Vital A, et al. Recurrent mutations of MYD88 and TBL1XR1 in primary central nervous system lymphomas. Clin Cancer Res. 2012; 18: 5203-5211.
9. Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett RD, Johnson NA, Severson TM, Chiu R, Field M, Jackman S, Krzywinski M, Scott DW, et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. 2011; 476: 298-303.
10. Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu I-M, Gallia GL, Olivi A, McLendon R, Rasheed BA, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008; 321: 1807-1812.
11. Lohr JG, Stojanov P, Lawrence MS, Auclair D, Chapuy B, Sougnez C, Cruz-Gordillo P, Knoechel B, Asmann YW, Slager SL, Novak AJ, Dogan A, Ansell SM, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc Natl Acad Sci USA. 2012; 109: 3879-3884.
12. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011; 32: 894-899.
13. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. Hum Mutat. 2013; 34: 2393-2402.
14. Zhang J, Grubor V, Love CL, Banerjee A, Richards KL, Mieczkowski PA, Dunphy C, Choi W, Au WY, Srivastava G, Lugar PL, Rizzieri DA, Lagoo AS, et al. Genetic heterogeneity of diffuse large B-cell lymphoma. Proc Natl Acad Sci USA. 2013; 110: 1398-1403.
15. Pasqualucci L, Trifonov V, Fabbri G, Ma J, Rossi D, Chiarenza A, Wells VA, Grunn A, Messina M, Elliot O, Chan J, Bhagat G, Chadburn A, et al. Analysis of the coding genome of diffuse large B-cell lymphoma. Nat Genet. 2011; 43: 830-837.
16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25: 25-29.

17. Iwasaki A, Medzhitov R. Regulation of adaptive immunity by the innate immune system. Science. 2010; 327: 291-295.
18. Ishii KJ, Akira S. Innate immune recognition of, and regulation by, DNA. Trends Immunol. 2006; 27: 525-532.
19. Dal Porto JM, Gauld SB, Merrell KT, Mills D, Pugh-Bernard AE, Cambier J. B cell antigen receptor signaling 101. Mol Immunol. 2004; 41: 599-613.
20. Montesinos-Rongen M, Godlewska E, Brunn A, Wiestler OD, Siebert R, Deckert M. Activating L265P mutations of the MYD88 gene are common in primary central nervous system lymphoma. Acta Neuropathol. 2011; 122: 791-792.
21. Montesinos-Rongen M, Schäfer E, Siebert R, Deckert M. Genes regulating the B cell receptor pathway are recurrently mutated in primary central nervous system lymphoma. Acta Neuropathol. 2012; 124: 905-906.
22. Ngo VN, Young RM, Schmitz R, Jhavar S, Xiao W, Lim K-H, Kohlhammer H, Xu W, Yang Y, Zhao H, Shaffer AL, Romesser P, Wright G, et al. Oncogenically active MYD88 mutations in human lymphoma. Nature. 2011; 470: 115-119.
23. Davis RE, Ngo VN, Lenz G, Tolar P, Young RM, Romesser PB, Kohlhammer H, Lamy L, Zhao H, Yang Y, Xu W, Shaffer AL, Wright G, et al. Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. Nature. 2010; 463: 88-92.
24. Li J, Wang C-Y. TBL1-TBLR1 and beta-catenin recruit each other to Wnt target-gene promoter for transcription activation and oncogenesis. Nat Cell Biol. 2008; 10: 160-169.
25. Perissi V, Scafoglio C, Zhang J, Ohgi KA, Rose DW, Glass CK, Rosenfeld MG. TBL1 and TBLR1 phosphorylation on regulated gene promoters overcomes dual CtBP and NCoR/SMRT transcriptional repression checkpoints. Mol Cell. 2008; 29: 755-766.
26. Perissi V, Aggarwal A, Glass CK, Rose DW, Rosenfeld MG. A corepressor/coactivator exchange complex required for transcriptional activation by nuclear receptors and other regulated transcription factors. Cell. 2004; 116: 511-526.
27. Parker H, An Q, Barber K, Case M, Davies T, Konn Z, Stewart A, Wright S, Griffiths M, Ross FM, Moorman AV, Hall AG, Irving JA, et al. The complex genomic profile of ETV6-RUNX1 positive acute lymphoblastic leukemia highlights a recurrent deletion of TBL1XR1. Genes Chromosomes Cancer. 2008; 47: 1118-1125.
28. Zhang J, Mullighan CG, Harvey RC, Wu G, Chen X, Edmonson M, Buetow KH, Carroll WL, Chen I-M, Devidas M, Gerhard DS, Loh ML, Reaman GH, et al. Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group. Blood. 2011; 118: 3080-3087.
29. Montesinos-Rongen M, Schmitz R, Brunn A, Gesk S, Richter J, Hong K, Wiestler OD, Siebert R, Küppers R, Deckert M. Mutations of CARD11 but not TNFAIP3 may activate the NF-kappaB pathway in primary CNS lymphoma. Acta Neuropathol. 2010; 120: 529-535.
30. Hock H, Meade E, Medeiros S, Schindler JW, Valk PJM, Fujiwara Y, Orkin SH. Tel/Etv6 is an essential and selective regulator of adult hematopoietic stem cell survival. Genes Dev. 2004; 18: 2336-2341.
31. De Braekeleer E, Douet-Guilbert N, Morel F, Le Bris M-J, Basinko A, De Braekeleer M. ETV6 fusion genes in hematological malignancies: a review. Leuk Res. 2012; 36: 945-961.
32. De Silva NS, Simonetti G, Heise N, Klein U. The diverse roles of IRF4 in late germinal center B-cell differentiation. Immunol Rev. 2012; 247: 73-92.
33. Pongubala JMR, Northrup DL, Lancki DW, Medina KL, Treiber T, Bertolino E, Thomas M, Grosschedl R, Allman D, Singh H. Transcription factor EBF restricts alternative lineage options and promotes B cell fate commitment independently of Pax5. Nat Immunol. 2008; 9: 203-215.
34. John SA, Clements JL, Russell LM, Garrett-Sinha LA. Ets-1 regulates plasma cell differentiation by interfering with the activity of the transcription factor Blimp-1. J Biol Chem. 2008; 283: 951-962.
35. Overbeck BM, Martin-Subero JI, Ammerpohl O, Klapper W, Siebert R, Giefing M. ETS1 encoding a transcription factor involved in B-cell differentiation is recurrently deleted and down-regulated in classical Hodgkin's lymphoma. Haematologica. 2012; 97: 1612-1614.
36. Flossbach L, Holzmann K, Mattfeldt T, Buck M, Lanz K, Held M, Möller P, Barth TFE. High-resolution genomic profiling reveals clonal evolution and competition in gastrointestinal marginal zone B-cell lymphoma and its large cell variant. Int J Cancer. 2013; 132: 116-127.
37. Morin RD, Mungall K, Pleasance E, Mungall AJ, Goya R, Huff R, Scott DW, Ding J, Roth A, Chiu R, Corbett RD, Chan FC, Mendez-Lago M, et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole genome sequencing. Blood. 2013; 122: 1256-1265.
38. Shirogane T, Fukada T, Muller JM, Shima DT, Hibi M, Hirano T. Synergistic roles for Pim-1 and c-Myc in STAT3-mediated cell cycle progression and antiapoptosis. Immunity. 1999; 11: 709-719.
39. Koeppel M, van Heeringen SJ, Smeenk L, Navis AC, Janssen-Megens EM, Lohrum M. The novel p53 target gene IRF2BP2 participates in cell survival during the p53 stress response. Nucleic Acids Res. 2009; 37: 322-335.
40. Rouault JP, Rimokh R, Tessa C, Paranhos G, Ffrench M, Duret L, Garoccio M, Germain D, Samarut J, Magaud JP. BTG1, a member of a new family of antiproliferative genes. EMBO J. 1992; 11: 1663-1670.
41. Montesinos-Rongen M, Van Roost D, Schaller C, Wiestler OD, Deckert M. Primary diffuse large B-cell lymphomas of the central nervous system are targeted by aberrant somatic hypermutation. Blood. 2004; 103: 1869-1875.
42. Narlik-Grassow M, Blanco-Aparicio C, Carnero A. The PIM family of serine/threonine kinases in cancer. Med Res

Rev. 2014; 34: 136-159.

43. Nyquist KB, Panagopoulos I, Thorsen J, Haugom L, Gorunova L, Bjerkehagen B, Fosså A, Guriby M, Nome T, Lothe RA, Skotheim RI, Heim S, Micci F. Whole-transcriptome sequencing identifies novel IRF2BP2-CDX1 fusion gene brought about by translocation t(1;5)(q42;q32) in mesenchymal chondrosarcoma. PLoS ONE. 2012; 7: 49705.

44. Camilleri-Broët S, Crinière E, Broët P, Delwail V, Mokhtari K, Moreau A, Kujas M, Raphaël M, Iraqi W, Sautès-Fridman C, Colombat P, Hoang-Xuan K, Martin A. A uniform activated B-cell-like immunophenotype might explain the poor prognosis of primary central nervous system lymphomas: analysis of 83 cases. Blood. 2006; 107: 190-196.

45. Yang Y, Shaffer AL 3rd, Emre NCT, Ceribelli M, Zhang M, Wright G, Xiao W, Powell J, Platig J, Kohlhammer H, Young RM, Zhao H, Yang Y, et al. Exploiting synthetic lethality for the therapy of ABC diffuse large B cell lymphoma. Cancer Cell. 2012; 21: 723-737.

46. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078-2079.

47. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26: 841-842.

48. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491: 56-65.

49. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29: 308-311.

50. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008; 10: 10-11.

51. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011; 39: 945-950.

52. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4: 1073-1081.

53. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7: 248-249.

Ce manuscrit de thèse traitant de la prédisposition génétique aux gliomes comporte 5 chapitres. 1/ une introduction, 2/ une présentation des méthodes utilisées, 3/ les résultats d'une étude largement collaborative identifiant des allèles à risque pour les glioblastomes et les gliomes non - glioblastomes ; 4/ un chapitre consacré à l'identification précise des allèles à risque pour les sous-types moléculaires de gliomes classés selon l'OMS 2016; 5/ la présentation d'une analyse d'exomes portant sur des oligodendrogliomes anaplasiques.

Plus en détail, le premier chapitre se consacre à une présentation générale des prédispositions et susceptibilités génétiques aux gliomes. Il place le sujet dans le contexte des connaissances actuelles dans le champ de la susceptibilité génétique au cancer, distinguant les allèles rares avec forte pénétrance et les allèles fréquents avec faible pénétrance. Après avoir présenté les principes des analyses par GWAS, ce chapitre évoque les différents travaux réalisés antérieurement, ayant permis d'identifier 12 allèles à risque par approches de GWAS.

Le deuxième chapitre est entièrement consacré à une présentation des méthodes utilisées pour réaliser les travaux bio-informatiques présentés par la suite.

Dans un premier temps, une description des populations étudiées pour l'approche GWAS est apportée, à savoir une compilation des études antérieures réalisés par des grands collaborations internationales; il s'agit des données du GICC (5189 cas de gliomes), le UK GWAS (636 cas), le German GWAS (880 cas), le MDA GWAS (1281 cas), le UCSF GWAS (677 cas), le GliomaScan (1653 cas) et les données françaises. De plus une description des techniques de génotype à partir de puces SNP-arrays d'illumina ainsi que les techniques de préparation des librairies pour séquençage « whole exome » et « RNA-seq » est également présentées.

Dans un deuxième temps, j'expose en détail les outils utilisés pour ses analyses bio-informatiques, comprenant: 1/les outils de séquençage, d'alignement, d'appels de variants, de filtration et de prédiction, 2/ les techniques d'analyses de GWAS, 3/les modélisations in silico des altérations fonctionnelles, 4/ les outils de visualisation et de construction de figure.

Le troisième chapitre est consacré à la présentation des résultats portant sur une méta-analyse des GWAS réalisées jusqu'en mai 2017. Le travail a donc porté sur 12496 cas, comparés à 18190 contrôles. Ce travail collaboratif international a donné lieu à une publication dans Nat Genet en 2017. Ce travail a d'abord permis de confirmer 10/12 allèles à risque précédemment identifiés. Cinq allèles à risque pour les glioblastomes et 8 pour les non glioblastomes sont nouvellement identifiés, amenant à un total de 25 allèles à risque, dont deux non confirmés par cette méta-analyse. La suite du travail a consisté à inférer l'impact des SNP identifiés sur l'expression des gènes localisés dans un rayon de 1 Mb. Cette approche permet de proposer un lien fonctionnel entre les SNP candidats et plusieurs fonctions cellulaires et voies de signalisation, dont *TP53-MDM4, P13K-AKT3, LRIG1-EGFR*, et le maintien des télomères. Ce travail représente la nouvelle référence internationale pour la connaissance des allèles de susceptibilité aux gliomes de l'adulte.

Le quatrième chapitre porte sur l'identification des allèles à risque pour chaque type moléculaire de gliomes, tels que définis par l'OMS 2016. A cette fin, j'ai cherché à établir des liens entre les 25 loci à risques identifiés à la suite de la méta-analyse et les sous-types de gliomes définis selon le statut IDH, 1p/19q et *TERT*. Les populations d'intérêt étaient celle du TCGA et celles du groupe de la Pitié-Salpêtrière (French GWAS et French sequencing). Les associations testées ont compris 1/le lien statistique entre les allèles à risque et le sous-type

moléculaire de gliome, 2/le nombre d'allèles à risque et l'âge de survenue du gliome, 3/les allèles à risque et la survie. La signification biologique des allèles à risque a été explorée in silico par data mining concernant à la fois l'expression des gènes candidats dans diverses régions du cerveau et la conformation des régions chromatinennes d'intérêt dans des cellules souches neuronales et progéniteurs (« Hi-C »). Au total, les associations entre 5 types moléculaires de gliomes et 25 allèles à risque ont été testées sur 2648 cas et 9365 contrôles. Les allèles à risque identifiés précédemment comme associés aux « nonglioblastomes » se révèlent essentiellement être associés aux gliomes IDH1 mutés. Les analyses poursuivies pour chaque sous-type moléculaire permettent d'aboutir à l'identification d'allèles à risques associés aux IDH mutés, et « triple positifs » en particulier, ainsi qu'à les « TERT only ». Par ailleurs des SNP à risque ont aussi été identifiés en 7p11.2 pour les gliomes avec amplification de EGFR, et en 9p21.3 pour les gliomes avec délétion de CDKN2A. Les logiciels permettant d'inférer les impacts fonctionnels amènent à identifier 5 grandes fonctions biologiques impliquant les SNP issus des analyses: 1/ une voie du métabolisme, en lien avec les mutations de IDH, 2/ une voie de maintenance des télomères, associée aux mutations de TERT, 3/ une voie de signalisation par EGFR et AKT, 4/ le rôle de TP53, et enfin 5/des gènes impliqués dans le développement neuronal.

Le cinquième chapitre présente les résultats d'une analyse par « whole exome sequencing » (WES) de 51 oligodendrogliomes anaplasiques, enrichis pour l'analyse de données publiques (TCGA) portant sur 43 échantillons supplémentaires. Le message principal de ce travail est l'identification de mutations hétérozygotes du gène TCF12 dans environ 8% des tumeurs. Plus en détail, les mutations identifiées sont hétérozygotes; l'expression protéique des formes mutées est en général augmentée et à la fois cytoplasmique et nucléaire, au contraire de la protéine sauvage qui n'est d'expression que nucléaire; certaines mutations tronquantes s'accompagnent d'une perte d'expression. Les protéines mutées perdent leur activité transcriptionnelle, testée in vitro avec un rapporteur luciférase sous une E-box. Enfin, une corrélation non significative est proposée avec un pronostic plus défavorable.

Au total, le thème principal du projet de thèse est l'identification de gènes de susceptibilité aux gliomes de l'adulte, à partir d'analyses de GWAS. Cette thématique s'inscrit dans un travail plus large portant sur l'identification de loci de susceptibilité aux tumeurs cérébrales.

La deuxième thématique du travail porte sur l'identification d'altérations génétiques somatiques, acquises dans les tumeurs, à partir d'analyses d'exomes.

université PARIS-SACLAY | ÉCOLE DOCTORALE Cancérologie, biologie, médecine, santé (CBMS)

Titre : Susceptibilité génétique et caractérisation moléculaire des gliomes

Mots clés : Génomique du Cancer, Susceptibilité génétique au cancer, Gliomes, Étude d'association génomique, Séquençage à haut débit.

Résumé : Les gliomes constituent les plus fréquentes des tumeurs malignes primaires du système nerveux central. Les liens qui existent entre ces tumeurs et un certain nombre de cancers rares héréditaires, comme les Neurofibromatoses I et II ou les syndromes de Turcot et de Li-Fraumeni, attestent d'une prédisposition génétique aux gliomes. L'observation d'un risque deux fois plus élevé de développer un gliome chez les parents de premier degré de patients atteints suggère aussi une possible prédisposition génétique dans les gliomes sporadiques. Par ailleurs, l'analyse à haut débit permet de préciser le profil somatique des gliomes et d'identifier des biomarqueurs pronostiques voire prédictifs et s'inscrire dans une démarche de traitement personnalisé du patient.

Durant ma thèse, je me suis focalisé sur deux axes de recherches complémentaires; l'identification de gènes de susceptibilité et la découverte de nouveaux gènes fréquemment mutés dans les gliomes, afin de déterminer les voies de signalisation contribuant à la gliomagenèse.

Dans leur ensemble, les résultats obtenus dans cette thèse apportent non seulement des informations importantes sur la nature de la prédisposition génétique aux gliomes mais également de son association spécifique pour les différents sous-types de tumeurs. La découverte d'un nouveau gène muté, offre la perspective à plus long terme d'un traitement personnalisé pour chaque patient sur la base du profil génétique de sa tumeur.

Title : Genetic susceptibility and molecular characterization of glioma

Keywords: Cancer genomics, Genetic Susceptibility to Cancer, Glioma, Genome Wide Association Studies, High Throughtput Sequencing.

Abstract : Gliomas are the most common adult malignant primary tumour of the central nervous system. Thus far, no environmental exposures has been linked to risk except for ionizing radiation, which only accounts for a very small number of cases. Direct evidence for inherited predisposition to glioma is provided by a number of rare inherited cancer syndromes, such as Turcot's and Li–Fraumeni syndromes, and neurofibromatosis. Even collectively, these diseases however account for little of the twofold increased risk of glioma seen in first-degree relatives of glioma patients.

My research was centred on two complementary research activities: Identifying susceptibility genes for glioma to delineate key biological pathways contributing to disease pathogenesis and to identify new recurrent mutated genes for glioma to provide for further insights into glial oncogenesis and suggesting targets for novel therapeutic strategies.

Collectively the findings in this thesis provide increased insight into the nature of genetic predisposition to glioma and substantiate the often distinct associations between susceptibility variants and glioma molecular groups. In addition the discovery of a new mutated gene in glioma offers the potential to support drug development and advance precision medicine for this tumours.