

Investigating intratumour heterogeneity analysis  
methods and their application in GBM

Georgette Nicola Tanner

Submitted in accordance with the requirements for the degree of  
Doctor of Philosophy

The University of Leeds

School of Medicine

October 2020

The PGR confirms that the work submitted is his/her/their own, except where work which has formed part of jointly authored publications has been included. The contribution of the PGR and the other authors to this work has been explicitly indicated below. The PGR confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The following chapters contain work that has been included in joint author publications:

## **Chapter 2**

The majority of this chapter is included in:

**Tanner, G.**, Westhead, D.R., Droop, A. and Stead, L.F. (2019)

Simulation of Heterogeneous Tumour Genomes with HeteroGenesis and In Silico Whole Exome Sequencing.

Bioinformatics. 35(16), pp. 2850-2852, <https://doi.org/10.1093/bioinformatics/bty1063>

This work was carried out entirely by myself, with the exception of creating parts of Figure 15, which were created by Stead L.F.. The material is reproduced under the Creative Commons CC BY license (<https://creativecommons.org/licenses/>).

## **Chapter 4**

The analysis using the SubClonalSelection model with the GLASS dataset is included in:

Barthel FP, Johnson KC, Varn FS, Moskalik AD, **Tanner G**, ... Verhaak RGW. (2019)

Longitudinal Molecular Trajectories of Diffuse Glioma in Adults.

Nature. <https://www.nature.com/articles/s41586-019-1775-1>

The model was run entirely by myself. The statistical tests used to explore associations with the results were carried out in collaboration with other authors on the publication. Figure 31 was created by Barthel F.P., although it illustrates results that I generated. The material is reproduced with the lead author's permission.

## Chapter 5

The majority of the work presented in this chapter is included in:

Golebiewska A, Hau AC, Oudin A, Stieber D, Yabo YA, Baus V, Barthelemy V, Klein E, Bougnaud S, Keunen O, Wantz M, Michelucci A, Neirinckx V, Muller A, Kaoma T, Nazarov PV, Azuaje F, De Falco A, Flies B, Richart L, Poovathingal S, Arns T, Grzyb K, Mock A, Herold-Mende C, Steino A, Brown D, May P, Miletic H, Malta TM, Noushmehr H, Kwon YJ, Jahn W, Klink B, **Tanner G**, Stead LF, Mittelbronn M, Skupin A, Hertel F, Bjerkvig R, Niclou SP. Patient-derived organoids and orthotopic xenografts of primary and recurrent gliomas represent relevant patient avatars for precision oncology. *Acta Neuropathol.* 2020 Oct 3. doi: 10.1007/s00401-020-02226-7. Epub ahead of print. PMID: 33009951.

All the analyses presented in the chapter were carried out entirely by myself. The material is reproduced under the Creative Commons Attribution 4.0 International License

(<https://creativecommons.org/licenses/by/4.0/>).

## Acknowledgements

I first want to thank the University for funding my studies, through a Leeds Anniversary Research Scholarship. I also want to thank my supervisors, Associate Professor Lucy Stead, Dr Alastair Droop, and Professor David Westhead, for their valuable support and guidance. My project additionally benefitted from the opportunity for me to visit Professor Roel Verhaak and his group at the Jackson Laboratory, to whom I am very grateful for the experience. This was made possible by grants from both the British Association for Cancer research, and The European Association for Cancer research, to whom I am also grateful. Lastly, I wish to acknowledge the support and resources provided by the university's Research Computing team.

## Abstract

Glioblastoma (GBM) is an incurable cancer with a median survival of 15 months. Despite debulking surgery, cancer cells are inevitably left behind in the surrounding brain, with a minority able to resist subsequent chemoradiotherapy and eventually form a recurrent tumour. This resistance is likely influenced by the cells' genotypes, which show high variability (intratumour heterogeneity), as a result of tumour evolution. Characterising changes in the genetic architecture of tumours through therapy, may allow us to understand the effect that different mutations and pathways have on cell survival, and potentially identify novel targets for counteracting resistance in GBM. Such analyses involve detection of mutations from bulk tumour samples, and then delineating them into individual genetically distinct 'subclones', through subclonal deconvolution. This is a complex process, with no reliable guidelines for the best pipelines to use. I therefore developed methods to allow simulation and *in silico* sequencing of genomes from realistically complex, artificial tumour samples, so that I could benchmark such pipelines. This revealed that no tested pipelines, using single bulk samples, showed a high level of accuracy, though mutation calling with Mutect2 and FACETS, followed by subclonal deconvolution with Ccube, showed the best results. I then used alternative approaches with the largest longitudinal GBM dataset investigated to date. I found that evidence of strong subclonal selection is absent in many samples, and not associated with therapy. Nonetheless, this does not negate the possibility of smaller, or less frequent, pockets of altered fitness. Using pathway analysis combined with variants that are informative of tumour progression, I identified processes that may confer increased resistance, or sensitisation to therapy, and which warrant further investigation. Lastly, I apply subclonal deconvolution to investigate mouse-specific evolution in GBM patient-derived orthotopic xenografts and found no clear evidence to suggest these models are unsuitable for investigations relevant to humans.

# Contents

<b>ABBREVIATIONS .....</b>	<b>8</b>
<b>CHAPTER 1 – INTRODUCTION .....</b>	<b>10</b>
1.1 OVERVIEW .....	10
1.2 TUMOUR EVOLUTION .....	12
1.3 INTRATUMOUR HETEROGENEITY .....	16
1.4 TREATMENT RESISTANCE IN CANCER .....	16
1.5 GLIOBLASTOMA .....	19
1.6 TREATMENT RESISTANCE IN GLIOBLASTOMA .....	23
1.7 MODELS OF GBM .....	25
1.8 COMPUTATIONAL INVESTIGATION OF ITH .....	26
1.9 HYPOTHESIS .....	30
1.10 AIMS AND OBJECTIVES .....	30
1.11 REFERENCES .....	24
<b>CHAPTER 2 – SIMULATION OF HETEROGENEOUS TUMOUR GENOMES AND <i>IN SILICO</i> WES DATA SETS .....</b>	<b>33</b>
2.1 INTRODUCTION .....	40
2.1.1 Overview .....	40
2.1.2 Limitations of existing benchmarking studies .....	41
2.1.3 Limitations of existing somatic genome simulation programs .....	44
2.2 METHODS AND RESULTS .....	47
2.2.1 Creation of a program for simulating realistically complex tumour genomes .....	47
2.2.1.1 HeteroGenesis Overview .....	47
2.2.1.2 HeteroGenesis Workflow .....	48
2.2.1.2.1 <i>heterogenesis_vargen</i> .....	48
2.2.1.2.2 <i>heterogenesis_varincorp</i> .....	50
2.2.1.2.3 <i>freqcalc</i> .....	57
2.2.1.3 How HeteroGenesis models scenarios to recreate tumour complexity .....	57
2.2.1.4 Modifications since publication .....	58
2.2.2 Optimisation of <i>In silico</i> whole-exome sequencing .....	58
2.2.2.1 Wessim .....	58
2.2.2.2 GemSIM error model .....	59
2.2.2.2.1 Error model creation .....	59
2.2.2.2.2 Error model validation .....	59
2.2.2.3 Wessim in probe hybridisation mode .....	62
2.2.2.4 w-Wessim .....	62
2.2.3 Demonstration of the combined use of HeteroGenesis with w-Wessim .....	66
2.3 DISCUSSION .....	70
2.4 REFERENCES .....	71
<b>CHAPTER 3 – BENCHMARKING OF MUTATION CALLING AND SUBCLONAL DECONVOLUTION METHODS .....</b>	<b>75</b>
3.1 INTRODUCTION .....	75
3.1.1 Overview .....	75
3.1.2 Variant calling .....	76
3.1.3 CNA calling .....	78
3.1.4 Subclonal deconvolution .....	81
3.2 RESULTS .....	83
3.2.1 Creation of datasets for use in benchmarking .....	83
3.2.1.1 Genome simulation .....	83
3.2.1.2 <i>In silico</i> sequencing and creation of BAM files .....	86
3.2.2 Benchmarking of mutation calling and CCF estimation methods .....	86
3.2.2.1 Variant calling .....	86
3.2.2.2 Copy number calling .....	92
3.2.2.3 Subclonal deconvolution .....	94
3.3 DISCUSSION .....	100
3.3.1 Overview .....	100
3.3.2 Variant calling .....	101
3.3.3 CNA calling and subclonal deconvolution .....	102
3.3.4 Future directions and conclusions .....	103

3.4 METHODS.....	104
3.4.1 Creating datasets for use in benchmarking.....	104
3.4.2 Variant calling.....	105
3.4.3 CNA calling.....	105
3.4.4 Subclonal deconvolution.....	106
3.5 APPENDIX.....	107
3.5.1 Variant calling commands.....	107
3.5.2 CNA heatmaps.....	110
3.6 REFERENCES.....	123
<b>CHAPTER 4 – IDENTIFICATION OF PATHWAYS RELEVANT TO GBM PROGRESSION THROUGH THERAPY.....</b>	<b>127</b>
4.1 INTRODUCTION.....	127
4.1.1 Overview.....	127
4.1.2 The GLASS dataset.....	127
4.1.3 Investigating the mode of tumour evolution.....	128
4.1.4 Identification of biological pathways relevant in GBM progression through therapy.....	129
4.2 RESULTS.....	133
4.2.1 Investigating the mode of tumour evolution.....	133
4.2.3 Pathway analysis.....	134
4.2.4 Running PathScore.....	135
4.2.5 Pathways differentially enriched between primary and recurrent tumours.....	137
4.2.6 Pathways enriched in clonally expanding cells.....	141
4.2.7 Pathways under-enriched in cells surviving through therapy.....	146
4.3 DISCUSSION.....	149
4.3.1 Summary.....	149
4.3.2 Investigating the mode of tumour evolution.....	149
4.3.3 Pathway analysis.....	150
4.3.4 Pathways differentially enriched between primary and recurrent tumours.....	151
4.3.5 Pathways enriched in clonally expanding cells.....	153
4.3.6 Pathways under-enriched in cells surviving through therapy.....	155
4.3.7 Conclusions and future directions.....	157
4.4 METHODS.....	158
4.4.1 SubClonalSelection methods.....	158
4.4.2 Pathway analysis methods.....	159
4.5 REFERENCES.....	160
<b>CHAPTER 5 - COMPARISON OF INTRATUMOUR HETEROGENEITY BETWEEN GBM PATIENT BIOPSIES AND PATIENT-DERIVED ORTHOTOPIC XENOGRAFTS.....</b>	<b>165</b>
5.1 INTRODUCTION.....	165
5.2 RESULTS.....	166
5.2.1 Sample details.....	166
5.2.2 Variant calling.....	167
5.2.3 Copy number calling.....	167
5.2.4 Correlation of CCFs.....	173
5.2.5 Further investigation of <i>EGFR</i> .....	181
5.3 DISCUSSION.....	182
5.4 METHODS.....	185
5.4.1 Variant calling.....	185
5.4.2 Copy number calling.....	186
5.4.3 Correlation of CCFs.....	186
5.5 REFERENCES.....	186
<b>CHAPTER 6 - DISCUSSION.....</b>	<b>188</b>
REFERENCES.....	191

# Abbreviations

24OH-CHOL	24S-hydroxycholesterol
27OH-CHOL	27-hydroxycholesterol
7 $\alpha$ OH-CHOL	7 $\alpha$ -hydroxycholesterol
ABCB11	ATP binding cassette subfamily B member 11
aCGH	array Comparative Genomic Hybridisation
ACOT8	Acyl-coenzyme A thioesterase 8
ACOX2	Acyl-CoA oxidase 2
AKR1[C4/D1]	Aldo-Keto reductase family 1 member [C4/D1]
Akt	Protein kinase B
AMACR	Alpha-methylacyl-CoA racemase
AMHR2	Anti-Mullerian hormone receptor Type 2
BAAT	Bile acid-CoA:amino acid N-acyltransferase
BCAT1	Branched chain amino acid transaminase 1
BMP	Bone morphogenetic protein
BMPR2	Bone morphogenetic protein receptor type 2
CCF	Cancer cell fraction
CDKN2A/B	Cyclin-dependent kinase inhibitor 2A/B
CH25H	Cholesterol 25-hydroxylase
CHOL	Cholesterol
CLDN8	Claudin-8
CN	Copy number
CNV	Copy number variant (germline)
CNA	Copy number aberration (somatic)
CSC	Cancer stem cell
CYP[7A/7B/8B/27A/39A/46A]1	Cytochrome P450 family [7/8/27/39/46] subfamily [A/B] member 1
ecDNA	Extrachromosomal DNA
EGFR	Epidermal growth factor receptor
EMT	Epithelial-to-mesenchymal transition
FXR	Farnesoid X receptor
GAD2	Glutamate decarboxylase 2
GATK	Genome analysis toolkit
GBM	Glioblastoma
GLASS	Glioma Longitudinal AnalySiS consortium
GSEA	Gene set enrichment analysis
HMGCS2	3-hydroxy-3-methylglutaryl-CoA synthase 2
HSD3B7	Hydroxy-Delta-5-Steroid Dehydrogenase, 3 Beta- And Steroid Delta-Isomerase 7
HSD17B4	Hydroxysteroid 17-Beta Dehydrogenase 4
ICT	Isocitrate
IDH[1/2]	Isocitrate dehydrogenase [1/2]
InDel	Insertion or deletion
ITH	Intratumour heterogeneity
LDL	Low density lipoprotein
LXR	Liver X receptor
MADif	Mean absolute difference
MAPK	Mitogen-activated protein kinase
MGMT	O6-methylguanine-DNA methyl-transferase
MPP5	Membrane palmitoylated protein 5
MTOR	Mechanistic target of rapamycin (serine/threonine kinase)
NF1	Neurofibromatosis type 1
OXCT[1/2]	3-oxoacid CoA-transferase [1/2]
PARD3	Par-3 family cell polarity regulator



PDGFR	Platelet-derived growth factor receptor
PIK3C[A/G]	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit [alpha/gamma]
PIK3R1	Phosphoinositide-3-kinase regulatory subunit 1
PI3K	Phosphoinositide 3-kinases
PTEN	Phosphatase and tensin homolog
RB1	Retinoblastoma protein
ROC	Receiver operator curve
RTK	Receptor tyrosine kinase
SCP2	Sterol carrier protein 2
SLC27A[2/5]	Solute carrier Family 27 member [2/5]
SMAD5	Mothers against decapentaplegic homolog 5
SNP	Single nucleotide polymorphism (germline)
SNV	Single nucleotide variant (somatic)
TERT	Telomerase reverse transcriptase
TGF $\beta$	Transforming growth factor beta
TGF $\beta$ R[2/3]	Transforming growth factor beta receptor [2/3]
TKI	Tyrosine kinase inhibitor
TMZ	Temozolomide
TP53	Tumour protein P53
VAF	Variant allele fraction
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
wt	Wild-type

# Chapter 1 – Introduction

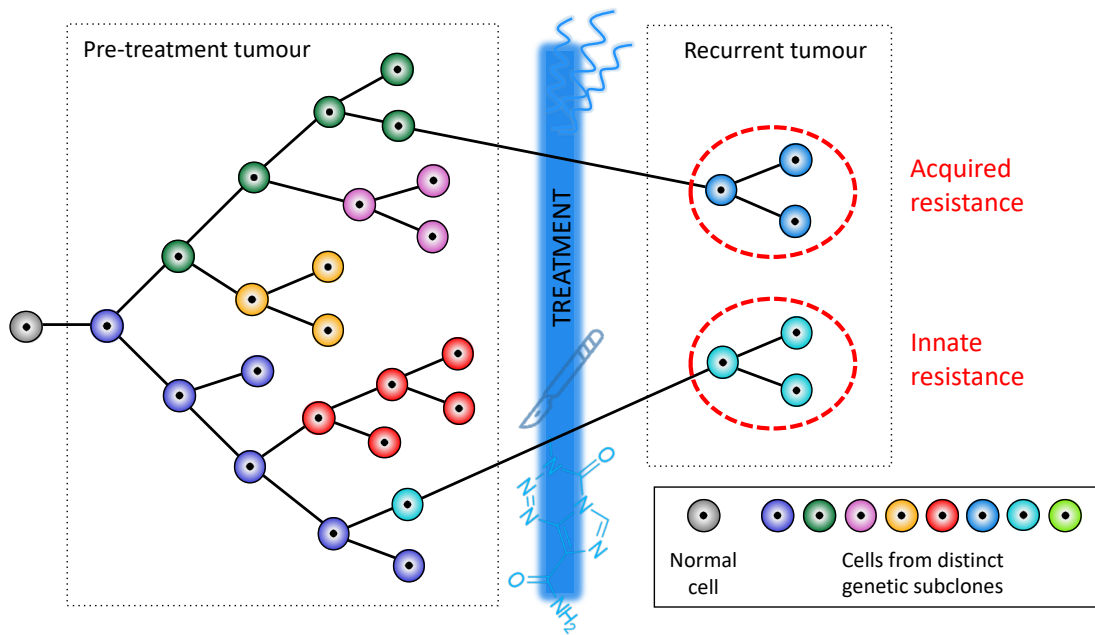
## 1.1 Overview

Cancer is a progressive and adaptable disease, and as a result, one of the biggest challenges in treating it is due to the ability of tumours to become resistant to therapies. An important goal in developing more effective therapies is, therefore, to understand the mechanisms that drive therapy resistance so that they can be avoided, or specifically targeted. Our group is focussed on understanding how glioblastomas (GBMs) survive treatment. GBM is a particularly aggressive cancer that, despite chemoradiotherapy, almost invariably recurs, resulting in a median patient survival of just ~15 months (Johnson and O’Neill, 2012; Stupp *et al.*, 2017).

Tumours may become intrinsically resistant to therapy in one of three ways, illustrated in Figure 1. All are a consequence of cell-to-cell variation within a tumour, known as intratumour heterogeneity (ITH), which may stem from either genetic or epigenetic effects. Adaptive resistance results from epigenetic change of cell state into an inherently more resistant cell type, whereas, innate and acquired resistance come from genetic mutations that first occurred prior to or after the onset of therapy, initially in a small minority of cells (Sharma *et al.*, 2017).

One way to identify these resistance mechanisms is through comparing matched primary and recurrent samples, using omics datasets that inform on their genomes or transcriptomes. Our group has access to such samples for GBM, from which I aim to investigate resistance mechanisms using genome sequencing datasets. This involves characterising genetic mutations and their influence on the genomic architectures of the samples through therapy, thereby informing us of cellular processes that affect GBM’s response to therapy. Part of this work also includes assessing the accuracy of certain steps involved in analysing genomic datasets, providing us and other researchers with a guide for the most suitable methods to use, and illustrating their limitations.

A Genetic ITH



B Epigenetic ITH

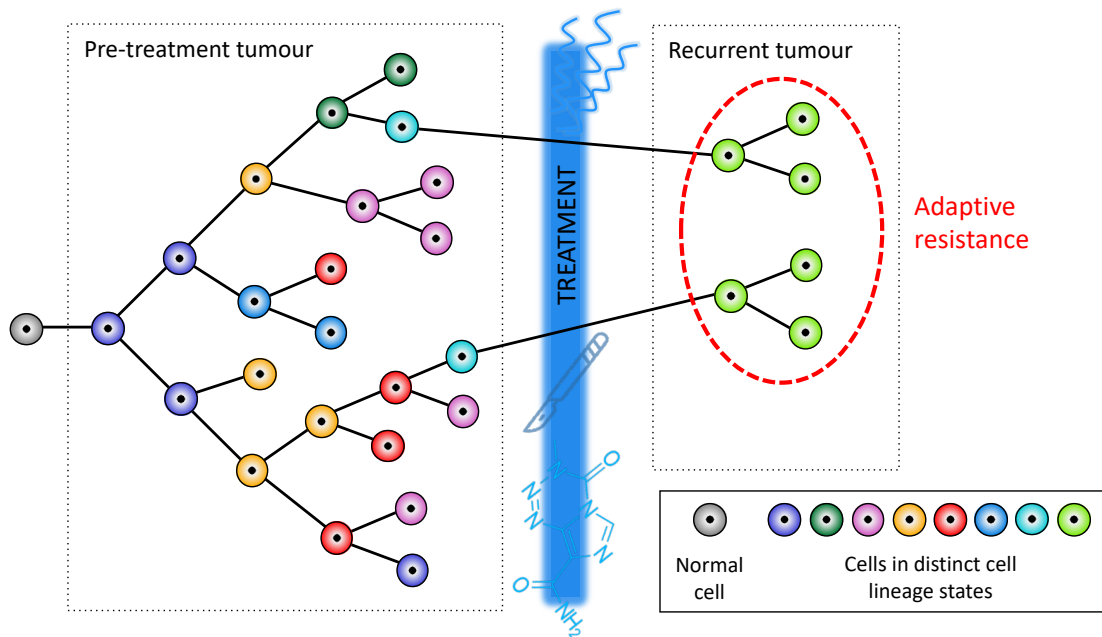


Figure 1. A representation of ITH within a primary and recurrent tumour, and how these enable treatment resistance. A) Differently coloured cells represent those in different genetic subclones, as a result of tumour evolution. Resistance to treatment may occur from a genetic mutation that was present before (innate) or after (acquired) the onset of treatment. B) Differently coloured cells represent those in different cell states, as a result of epigenetic plasticity. Resistance to therapy may occur from an adaptive epigenetic change in cell states (adaptive).








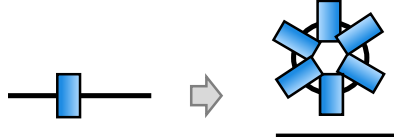
## 1.2 Tumour Evolution

Tumours are dynamic systems that evolve over time, enabling them to become more aggressive or adapt to changes in their environment, through both genetic and epigenetic mechanisms. At the genetic level, evolution occurs as a result of inherited genome variation from parent to daughter cells. Somatic mutations early on in a tumour's development lead to increased cell division and reduced activity of DNA repair pathways or apoptosis checkpoints. This fuels continuous development and accumulation of subsequent mutations throughout the tumour, causing a phylogenetic tree structure of related cell populations. Each contains distinct sets of mutations that results in further variation in phenotypes. The terms 'subclone' or 'clone', are used to refer to these genetically distinct cell populations. They can be used flexibly, and technically refer to any group of cells that differ to others at any position in their genome, which may describe each individual cell in a tumour. More commonly, they're used to refer to a group of genetically similar cells that share a common ancestor with a mutation conferring a distinct phenotype (Sottoriva *et al.*, 2017) (Figure 1A).

A variety of categories of genetic mutations exist in a cancer genome, illustrated in Table 1 (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes, 2020; Li *et al.*, 2020; Zack *et al.*, 2013; Rode *et al.*, 2016). These are either germline mutations present in every cell in the body, or somatic mutations specific to cancer cells. Additionally, somatic mutations may be present in either every cell in a tumour, and described as 'clonal', or alternatively, present in only a subset of cells, and described as 'subclonal'.

Cancer cells are under constant pressure, from both competition against each other for space and resources, as well as attempting to avoid destruction by the body's immune system. The variation in genotypes in a tumour results in some subclones having an advantage that allows them to outcompete others and grow at a faster rate. Eventually, a subclone may expand to such an extent that it dominates the whole tumour, thereby allowing the tumours to become more aggressive over time, or adapt in response to changes in pressures, such as the initiation of drugs. This selection of specific subclones is analogous to the Darwinian evolution of populations of organisms, where germlines acquire new mutations and the fittest members eventually outcompete, or diverge from, others in their species. As such, many of the terms and approaches to studying population evolution dynamics have become common in studies of tumour evolution (Cross *et al.*, 2016; Davis *et al.*, 2017; McGranahan and Swanton, 2017; Sottoriva *et al.*, 2017).

Table 1. Genetic mutations common in cancer genomes.

Mutation type		Description		Potential effects on the cell
Point variants	Single nucleotide polymorphisms (SNPs)	Germline single base substitutions.		Altered or shortened protein sequences. Dysregulation of genes.
	Single nucleotide variants (SNVs)	Somatic single base substitutions.		
	Insertions and deletions (InDels)	Germline or somatic short losses or gains, generally ≤50bp.		
Structural mutations	Translocations and mobile-element transpositions	Sections of genome that have moved to different positions		Disruption of genes over break points. Creation of gene fusions.
	Inversions	Regions that been inverted in place.		
	Copy number variations (CNVs)	Regions that have been replicated or deleted in the germline genome, generally >50bp.		Disruption of genes over break points. Increased expression of amplified genes. Decreased expression of deleted genes.
	Copy number aberrations (CNAs)	Regions that have been replicated or deleted in a somatic genome, generally >50bp.		
	Chromothrypsis	A chromosomal shattering event that occurs in a minority of samples across cancer types, resulting in widespread genome rearrangements and CNAs.		Combined effects of above structural mutations.
	Extrachromosomal DNA (ecDNA)	A specific type of amplification when a region of DNA forms a circular structure outside of the chromosomes.		Very high expression of genes. Uneven segregation and distribution of ecDNA allows for rapid changes in tumour genotypes.

The dynamics of tumour evolution can be categorised into four modes; neutral, linear, branched, and punctuated evolution (Figure 2) (Davis *et al.*, 2017). The timing and level of selective advantage that new mutations confer to cells, over others, as well as spatial constraints of cell populations, are the major factors in determining which mode is present in each tumour, and the degree of genetic ITH observed:

**Neutral evolution:** New mutations confer no selective advantage to cells, which continue to grow and develop further mutations at a constant rate. Such scenarios are possible for prolonged periods if an adaptive peak has been reached (Eldredge *et al.*, 2005). Importantly, while neutral evolution lacks any active selection of individual subclones, 'neutral drift' may still cause subclones to passively expand over others due to stochastic processes (Sottoriva *et al.*, 2017). Nonetheless, this mode results in the highest level of ITH seen in tumours.

**Linear evolution:** New driver mutations provide strong selective advantages to cells, causing them to dominate the tumour as a result of selective sweeps. This results in a low level of ITH.

**Branched evolution:** Similar to linear evolution, but where full clonal sweeps are prevented due to i) advantageous mutations developing in multiple competing cells simultaneously and continuously, ii) subclone cooperation, or iii) spatial separation of subclones (de Bruin *et al.*, 2014; West *et al.*, 2019; Noble *et al.*, 2019). ITH is moderately high in this mode.

**Punctuated evolution:** A 'big-bang' of many mutations and subsequent selection of one or a few subclones, followed by prolonged stasis with few new mutations or further selection (Cross *et al.*, 2016). This results in a low level of ITH.

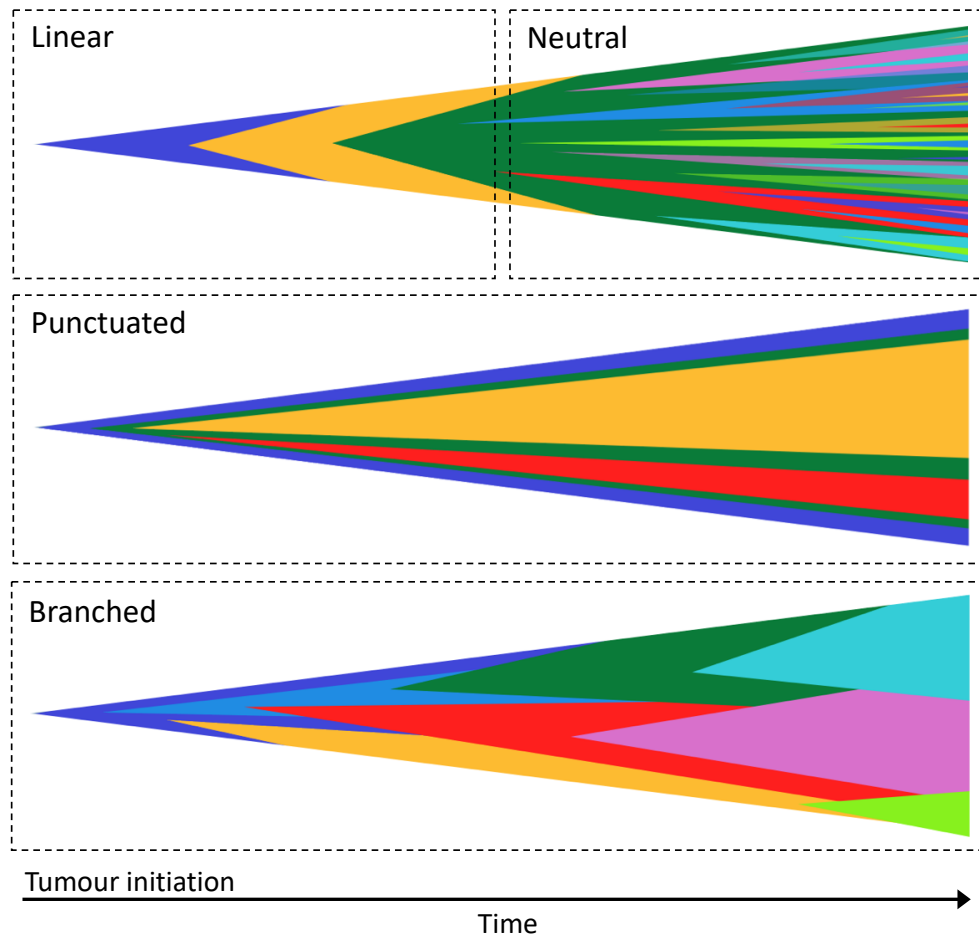


Figure 2. Representations of tumour evolution modes, with each colour corresponding to a different subclone. The top panel shows a tumour that started evolving linearly, but then switched to neutral evolution in the absence of further clonal sweeps. Though not represented in the diagram, each subclone in the branched and linear evolution panels is subject to continuous within-clone neutral evolution over time, whereas punctuated evolution tends to describe only copy number evolution, which may remain stable for prolonged periods without additional mutations.

These are not distinct scenarios, and tumours go through multiple evolutionary modes in their lifetime (West *et al.*, 2019). Furthermore, different mutation types have a propensity for different evolutionary modes; evidence supporting linear and branched evolution is mainly derived from point variants, whereas, evidence for punctuated evolution has generally come from copy number analyses (Newburger *et al.*, 2013; Wang *et al.*, 2014; Gerstung *et al.*, 2020). An additional factor that influences tumour evolution is the fact that subclones interact with each other through secretion of molecules and changes to the microenvironment. This allows for non-autonomous selection, where non-dominant genetic subclones drive overall tumour growth and maintain clonal diversity (Marusyk *et al.*, 2014), and, in some cases, enables metastases to develop (Sanborn *et al.*, 2015).

## 1.3 Intratumour heterogeneity

Intratumour heterogeneity (ITH) is a fundamental feature of all cancers, and provides tumours with the ability to adapt and develop resistance to therapies.

Genetic ITH results in a large pool of distinct genomes for resistance conferring mutations to exist by chance. The mutations may have a neutral, or even deleterious effect on cells prior to treatment, but result in a beneficial effect after its onset, leading to selection and clonal expansion of those cells. As previously mentioned, the mode of evolution influences the level of genetic ITH seen in a tumour, with strong selective forces reducing ITH, and weak selection or neutral evolution allowing a build-up of ITH, and therefore an increased chance of resisting therapy through genetic mechanisms (Davis *et al.*, 2017).

A further factor that influences the level of genetic ITH is the presence of ecDNA. These molecules undergo uneven segregation during cell division, due to their lack of centromeres, leading to an uneven distribution in otherwise genetically similar populations. This causes a major increase in ITH, and additionally, allows for rapid changes in genotypes in response to changes in environment, such as the initiation of drugs (Nathanson *et al.*, 2014; Decarvalho *et al.*, 2018; Turner *et al.*, 2017; Xu *et al.*, 2019; Verhaak *et al.*, 2019; Kim *et al.*, 2020).

At the transcriptomic level, ITH of cancer cell profiles may occur due to regional differences in the microenvironment, such as the extracellular matrix, availability of blood supply, secreted molecules from nearby cells, or interaction with immune cells, all of which affect the regulation of cancer cells (Yu *et al.*, 2020; Nelson and Bissell, 2006; Darmanis *et al.*, 2017). More importantly, cell plasticity enables cancer cells to undergo epigenetic reprogramming and revert back to developmental-like stem cell states, with others differentiated into more proliferative cell types that correspond, to some extent, with varying stages along lineage pathways, dependant on the cell of origin for that tumour (Lathia *et al.*, 2015; Liao *et al.*, 2017; Neftel *et al.*, 2019; Couturier *et al.*, 2020). This plasticity in cell states allows tumours to transition to more resistant states in response to therapy, as discussed below.

## 1.4 Treatment Resistance in Cancer

Cancer therapies come in many forms: Cytotoxic therapies, including radiotherapy and many chemotherapies, aim to damage cells by interfering with essential cellular processes or inducing DNA damage, effects which are less tolerated or unable to be repaired specifically in cancer cells; targeted therapies aim to disrupt specific proteins or cellular processes that are dysregulated in cancer cells; monoclonal antibodies and vaccines aim to direct a patient's natural immune system to specific antigens; and adoptive T cell therapy, immune checkpoint inhibitors, immune system modulators, and oncolytic



viruses aim to cause a general increase of the immune response to tumours (Kruger *et al.*, 2019; Farkona *et al.*, 2016).

These therapies are all susceptible to tumours developing resistance to them, via innate or acquired genetic changes, or an adaptive transcriptomic change in cell states. Such mechanisms allow cancer cells to develop resistance to a particular therapy, or to develop multidrug resistance, as outlined below (Cree and Charlton, 2017).

### **Alteration of drug target**

Cancer cells can become resistant to the effects of a drug if they acquire mutations that prevent it binding with its target. This can be achieved through altering the localization of the target, changing its conformation so that it no longer binds to the drug, or reducing levels of the target in the cell. For example, nutlins are a class of pre-clinical small molecule activators of the tumour suppressor protein p53. They enhance p53 mediated apoptosis by preventing its binding to the ubiquitin ligase Mdm2, which otherwise blocks p53 transcriptional activity and induces its degradation (Pflaum *et al.*, 2014). Resistance to nutlins has been shown in multiple cancer cell lines as a result of acquired missense mutations in the DNA binding domain of the p53 gene, thereby preventing its activation (Aziz *et al.*, 2011; Michaelis *et al.*, 2011).

### **Increased resistance to apoptosis**

Most cancer therapies aim to trigger apoptosis in cancer cells. Therefore, an important mechanism of therapy resistance is for cells to increase their tolerance to such signalling pathways (Mohammad *et al.*, 2015). In order for a cell to avoid apoptosis, antiapoptotic proteins must sequester proapoptotic activator proteins. The small-molecule drug ABT-737, is an antagonist of Bcl-2 antiapoptotic protein that underwent evaluation in clinical trials for several cancers. In non-Hodgkin lymphoma cell lines, long-term exposure of this drug was found to initiate resistance through transcriptional upregulation of alternative antiapoptotic proteins BFL-1 or Mcl-1, thereby allowing the cells to sufficiently sequester the activator proteins and avoid apoptosis (Yecies *et al.*, 2010).

### **Increased DNA repair**

Many cancer drugs aim to take advantage of cancer cells' reduced ability to repair DNA. However, these cells often subsequently increase DNA damage in response. Such effects have been demonstrated in ovarian cancer patients receiving platinum-based therapy, where cells taken from patients prior to and after the onset of therapy were subjected to the drug *ex vivo* for 1h. The proportion of the resulting DNA crosslinking was measured over time, with similar peak levels in

the two groups. After 24 hours the previously drug naïve cells had repaired 2.85% of the damage, whereas cells from patients who had previously received the drug, had repaired 71.23% of the damage, demonstrating a significantly increased DNA damage repair response following drug exposure (Wynne *et al.*, 2007).

### **Drug inactivation**

Glutathione S-transferases (GSTs) are a family of enzymes involved in detoxication reactions, and found to be associated with therapy resistance across multiple cancers. GSTs enable cancer cells to better resist treatments through both conjugating with drugs, and reducing reactive oxygen radicals caused by chemo and radiotherapy (Singh, 2015; Allocati *et al.*, 2018). For example, *GSTP1* was found to be upregulated in osteosarcoma cell lines following platinum-based chemotherapy (Huang *et al.*, 2007), and increased expression of *GSTP1* in osteosarcomas from patients receiving the same therapy was associated with a significantly higher relapse rate and a worse prognosis (Pasello *et al.*, 2008).

### **Increased drug efflux**

One method to reduce the negative effects of a drug on a cell is to reduce its concentration by increasing the drug's efflux from the cell (Yu *et al.*, 2013). ATP binding cassette (ABC) transporters are a family of transporters involved in this process (Domenichini *et al.*, 2019). In multiple-myeloma patients, 40 genes have been identified as being prognostic of progression-free survival. 7 of these are from the ABC transporter family, suggesting their involvement in primary drug resistance (Hassen *et al.*, 2015).

### **Epithelial-to-mesenchymal transition**

Cancer stem cells (CSCs) are a subpopulation of neoplastic cells with self-renewing properties, thought to give rise to more differentiated cell types and drive tumour progression. These cells are also thought to be largely responsible for driving therapy resistance and relapse, owing to a lack of proliferation and survival checkpoints (Yamada and Nakano, 2012; Prager *et al.*, 2020; Lathia *et al.*, 2015). This 'stemness' phenotype has been found to be largely attributable to epigenetic changes resulting from an epithelial-to-mesenchymal transition (EMT) programme, initiated via pathways including transforming growth factor beta (TGF $\beta$ ) – SMAD and Wnt signalling. Cells undergoing EMT lose their epithelial cell junctions and apical–basal polarity, and instead acquire mesenchymal like features such as an elongated fibroblast-like morphology, with increased capacity for migration and invasion (Shibue and Weinberg, 2017). EMT's involvement in prognosis is illustrated in a study where resistance of both non-small-cell lung cancer (NSCLC) cell lines and patient tumours to

phosphoinositide 3-kinases (PI3K)/protein kinase B (Akt) or epidermal growth factor receptor (EGFR) inhibitors was higher for those with a mesenchymal EMT gene expression signature. Additionally, inhibition of the EMT marker Axl receptor tyrosine kinase, sensitised mesenchymal NSCLC cell lines and xenograft tumours to EGFR inhibitors. The same effects were not seen for common cytotoxic chemotherapy drugs, suggesting EMT confers resistance only to certain types of therapies (Byers *et al.*, 2013).

### **Reduced cell proliferation**

Studies are finding accumulating evidence of subpopulations of cells in tumours that evade the effects of therapy by entering into a reversible, quiescent persister cell state, capable of only transient partial differentiation. Similar to the phenomenon seen in bacterial populations (Dawson *et al.*, 2011), cell quiescence prevents drugs from having the same level of corruption in the cells. This has been demonstrated in lung cancer cells treated with the tyrosine kinase inhibitor (TKI) erlotinib, from which, resistant populations arose with negligible growth. Many of these populations, across different replicates, were found to have a variety of classic erlotinib specific resistance mechanisms, such as mutations in the mitogen-activated protein kinase (MAPK) pathway. This led to speculation that persister cells act as a reservoir for subsequent acquired resistance mechanisms (Ramirez *et al.*, 2016; Dawson *et al.*, 2011).

Another mode of therapy resistance has been recently identified in melanoma cells. These were found to have extensive transcriptional variability, with semi-coordinated transcription of high levels of multiple resistance markers, such as receptor tyrosine kinases, in a very small number of cells prior to therapy. Addition of a B-Raf inhibitor resulted in fixation of resistance markers and epigenetic reprogramming of those cells, following loss of SOX10-mediated differentiation. Interestingly, the cells that went on to become stably resistant through this process, had high rates of proliferation both prior to, and during, drug onset, suggesting this is a different mechanism to those resulting in persister cell states (Shaffer *et al.*, 2017).

## **1.5 Glioblastoma**

Glioblastoma (GBM) is the most common and aggressive malignant primary brain cancer in adults, responsible for around half of all cases. Despite surgical intervention and chemoradiotherapy, the tumours almost invariably become treatment resistant and recur, causing mortality with a median survival of just 15 months (Johnson and O'Neill, 2012; Stupp *et al.*, 2017). GBMs are grade IV diffuse gliomas, which are characterised by their ability for cancer cells to infiltrate into surrounding brain parenchyma, with lower grade diffuse gliomas typically being grade II/III diffuse astrocytoma and oligodendroglioma. The 5-year survival rate for GBM is around 5% (Stupp *et al.*, 2009; Marenco-Hillebrand *et al.*, 2020; Ostrom *et al.*,

2016), far lower than any other malignant brain or CNS cancer, which together have an average 5-year survival rate of 34.9% (Ostrom *et al.*, 2016). Grade II/III diffuse astrocytoma and oligodendroglioma have a much better prognosis, with 5-year survival rates of 49.7% and 80.9% (Ostrom *et al.*, 2016). GBM affects 2.05 per 100,000 annually in the UK, with estimates in other countries varying from 0.59 (Korea) to 3.40 (Australia) per 100,000 (Tamimi and Juweid, 2017). Cases increase with age, peaking at 75–84 years (Tamimi and Juweid, 2017). Other risk factors include being male (around 3:2 ratio to females), white ethnicity, previous radiation exposure, and carrying germline mutations in apoptotic or DNA repair pathways, such as *TP53* mutation (Li Fraumeni syndrome) or *NF1* (Neurofibromatosis type 1) (Rice *et al.*, 2016; Ostrom *et al.*, 2016; Thakkar *et al.*, 2014).

GBM, diffuse astrocytoma, and oligodendroglioma are further categorised into 3 World Health Organisation (WHO) classifications based on the presence of co-occurring deletion of the chromosomes arms 1p and 19q (1p/19q-codeletion) and the mutation status of isocitrate dehydrogenase 1 and 2 (*IDH1/IDH2*) (Figure 3). In healthy cells, IDH1 and IDH2 enzymes function primarily in the reversible conversion between isocitrate (ICT) and  $\alpha$ -ketoglutarate ( $\alpha$ -KG), within the citric acid cycle. In many glioma tumours, however, mutations within the ICT binding site allow IDH to perform an additional reaction whereby  $\alpha$ -KG is converted into 2-hydroxyglutarate (2HG) (Wise *et al.*, 2011), which causes genome-wide histone alterations and DNA hypermethylation (Xu *et al.*, 2011; Ye *et al.*, 2018). *IDH*-mutant gliomas are most commonly slower-growing grade II or III oligodendrogliomas (1p/19q-codeletion) or astrocytomas (1p/19q-wildtype (wt)), but may later progress to a fast-growing grade IV GBM. *IDHwt* gliomas are most commonly grade IV (GBMs), with a smaller fraction being a lower grade astrocytoma, and are associated with worse survival and therapy response compared to IDH mutant tumours (Louis *et al.*, 2016; Sepúlveda-Sánchez *et al.*, 2018; Wesseling and Capper, 2018; Perry and Wesseling, 2016; Christians *et al.*, 2019).

Gliomas can be further categorised into three subgroups (proneural, mesenchymal, and classical), identified through clustering of tumour transcriptomes. Within each tumour, all three subgroups are supported by varying proportions of individual cells, the most prevalent of which reflects the overall bulk tumour subtype, and is influenced by the genetic alterations present (Wang *et al.*, 2017; Verhaak *et al.*, 2010).

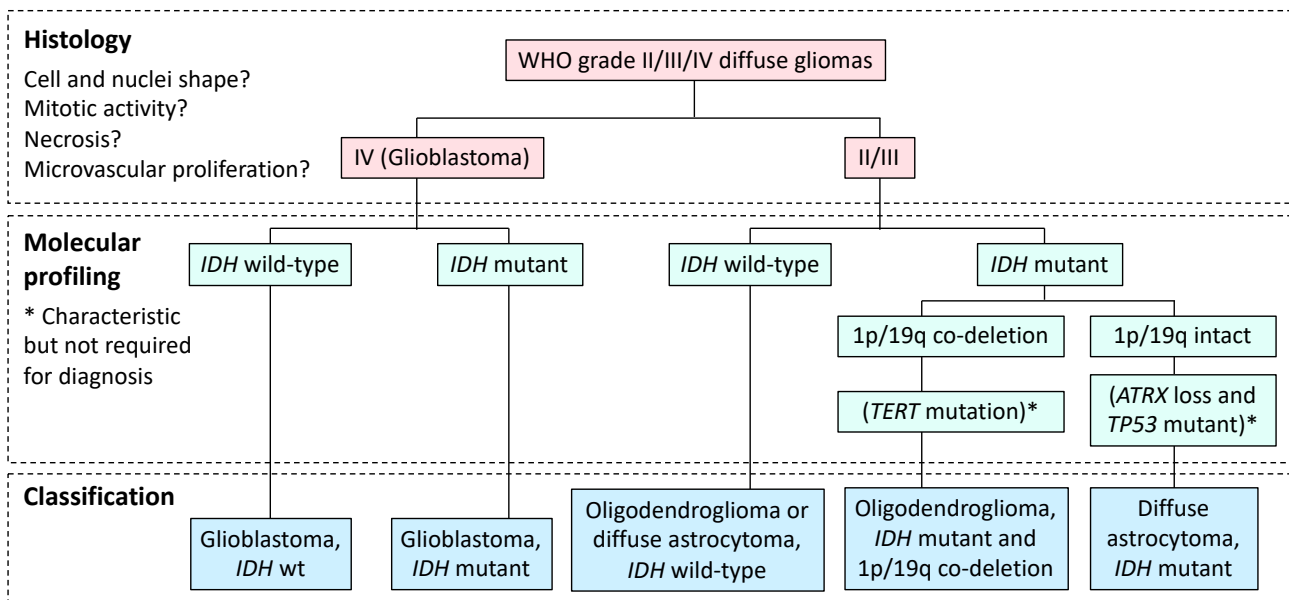


Figure 3. Flow diagram of the processes involved in classifying diffuse gliomas. (*ATRX*: alpha thalassemia/mental retardation syndrome X-linked.)

GBM, and other diffuse gliomas, have several characteristic features that make them particularly challenging to treat (Aldape *et al.*, 2019). Firstly, surgeons are unable to resect a wide margin around the tumour, as they do with other cancers, owing to the damaging consequences of removing healthy brain cells. The infiltrative nature of GBM and other diffuse glioma cells to migrate and grow deep into surrounding brain parenchyma, further exacerbates the issue and, as a result, many cancer cells are left remaining (Lee *et al.*, 2018). Secondly, the blood brain barrier that surrounds the brain and spinal cord, severely reduces the delivery of drugs to the brain, limiting the choice of effective treatments (Riganti *et al.*, 2014). Lastly, GBMs in particular have extensive ITH. This is seen at both the genetic level, through spatially distinct multisampling genome sequencing studies (Sottoriva *et al.*, 2013), as well as at the transcriptomic level, through histology (Perry and Wesseling, 2016) and single-cell RNA sequencing. The latter has undergone extensive investigation lately, with GBM cells found to cluster into a spectrum of distinct cell states that recapitulate different neural developmental pathways (Nefitel *et al.*, 2019; Couturier *et al.*, 2020).

As with all cancers, GBMs are individually highly variable in their genetic makeup, but show general trends in their progression (Körber *et al.*, 2019; Gerstung *et al.*, 2020; Barthel *et al.*, 2018). Loss of chromosomes 10 (covering phosphatase and tensin homolog (*PTEN*)) and 9p (covering cyclin-dependent kinase inhibitors 2A and 2B (*CDKN2A* and *CDKN2B*)), homozygous loss specifically of *CDKN2A* and *CDKN2B*, and gain of chromosome 7 (covering *EGFR*) are common early events, with at least two of these found as clonal in 81% of IDHwt GBMs. Other common early events include coding variants in *EGFR* and *TP53*, and promotor variants in telomerase reverse transcriptase (*TERT*). Point variants in *PTEN*, and aneuploid changes such as

gain of chromosomes 19q and 20q, and loss of chromosome 22q are common intermediate or later events. (Körber *et al.*, 2019; Barthel *et al.*, 2018; Brastianos *et al.*, 2017).

These, and other common mutations in GBM, lead to the dysregulation of pathways including receptor tyrosine kinase (RTK), PI3K, MAPK, Wnt, and Akt signalling, as well as the p53 DNA damage repair pathway. Additionally, activating variants in the promoter of *TERT*, involved in the rate limiting step in elongation of telomeres, immortalises GBM cells early on in development (Ceccarelli *et al.*, 2016). Loss of *CDKN2A/B*, which are involved in regulating the cell cycle, further enables GBM cells to proliferate uncontrollably (Ellis *et al.*, 2019; Barthel *et al.*, 2018; Sottoriva *et al.*, 2013; The Cancer Genome Atlas (TCGA) Research Network, 2008).

A common feature of GBM is ecDNA, where genes are found in high copy numbers on circular DNA outside of centromere containing chromosomes. As a result of the uneven segregation and distribution of these molecules, ecDNA has shown to be an important contributing factor to GBM's high levels of genetic ITH and allows for rapid changes in genotypes (Nathanson *et al.*, 2014; Decarvalho *et al.*, 2018; Turner *et al.*, 2017; Xu *et al.*, 2019; Verhaak *et al.*, 2019; Kim *et al.*, 2020). *EGFR* is one example of a gene that is commonly present on ecDNA in GBM, and codes for a transmembrane receptor tyrosine kinase in the ERBB family, with numerous signalling ligands that activate the PI3K/Akt, MAPK/extracellular signal-related kinase (ERK), and Janus kinases (JAK)/signal transducer and activator of transcription (STAT) pathways (An *et al.*, 2018). In addition to amplifications, structural variants of *EGFR* are common in GBM, particularly specific loss of exons 2-7, resulting in the protein variant EGFRvIII, present in around a third of GBMs (Yang *et al.*, 2017). EGFRvIII is a constitutively activated form of the receptor, resulting in signalling independent of ligand, causing increased DNA mismatch repair, proliferation, invasiveness and angiogenesis, and reduced apoptosis. Additionally, it induces chromosomal instability and is, alone, sufficient to initiate gliomagenesis after a long latency (Noorani *et al.*, 2020; Struve *et al.*, 2020).

Standard treatment for GBM includes surgical resection, followed by radiotherapy and temozolomide (TMZ) chemotherapy. These are both cytotoxic therapies that aim to kill cancer cells through damaging their DNA. Unlike healthy cells, which can recover from cytotoxicity, cancer cells often lack the DNA damage repair mechanisms. Radiotherapy, delivered via an external beam to gliomas, directly damages cells' DNA, primarily through double-strand breaks (Huang and Zhou, 2020). TMZ is one of the few chemotherapy drugs able to cross the blood brain barrier, owing to its ability to downregulate mechanisms responsible for its efflux back into the blood stream, and has been part of standard treatment since 2005 (Stupp *et al.*, 2005; Riganti *et al.*, 2014). It induces damage through methylation of DNA, including the formation of O<sup>6</sup>-methylguanine lesions. These are highly cytotoxic in cells unable to repair them, resulting in mismatched guanine to thymine bases after replication which, if also not repaired, ultimately result in a double-strand breaks (Drabløs *et al.*, 2004). Importantly, expression of the DNA repair protein gene O<sup>6</sup>-

methylguanine-DNA methyl-transferase (*MGMT*) greatly reduces the efficacy of TMZ by removing the methyl groups it deposits. In about half of GBMs the *MGMT* promoter is methylated, silencing its expression and typically conferring treatment sensitivity (Hegi *et al.*, 2005).

## 1.6 Treatment resistance in glioblastoma

While treatment resistance is a potential issue in almost all cancers, it is of particular significance in GBM, where it contributes to the almost 100% mortality in patients. In the past year, longitudinal genomic studies of matched primary and recurrent GBMs from patients having undergone standard therapy, have aimed to identify genetic mechanisms responsible for resistance (Kraboth and Kalman, 2020; Barthel *et al.*, 2019; Körber *et al.*, 2019; Wang *et al.*, 2016). However, they've shown a general lack of evidence supporting such mechanisms in GBM. Few genes are found commonly altered specifically in recurrent tumours, and the subclonal genetic ITH of the primary tumours are at least partially maintained through to recurrent. Together, these studies suggest that, instead, transcriptomic adaptive mechanisms are likely the primary driving force behind standard therapy resistance in GBM.

However, evidence for some mutations conferring genetic resistance has been identified in a minority of patients. Mutations in genes encoding DNA mismatch-repair proteins are commonly seen in recurrent gliomas in response to TMZ (Wang *et al.*, 2016). In GBM specifically, the mismatch-repair protein gene *MSH6* was found altered in 3 out of 14 post standard therapy recurrent tumours, but in none of 40 primary tumours examined. Furthermore, its expression was lost in 7 of 17 recurrent tumours, but in none of 17 primary tumours examined (Cahill *et al.*, 2007). Gain of function mutations in latent TGF-beta binding protein 4 (*LTBP4*) were identified in 10/93 patients, and in none of the matched primary recurrent tumours (Wang *et al.*, 2016). This gene activates TGF- $\beta$  signalling, thereby inducing EMT and increasing proliferation (Comaills *et al.*, 2016).

Studies investigating resistance mechanisms at the transcriptomic level have identified numerous genes and pathways that are differentially regulated in recurrent tumours, and implicated in increased therapy resistance (Ali *et al.*, 2020; Chien *et al.*, 2019; Lee, 2016; Yi *et al.*, 2019). However, the most important observations are likely those that show widespread epigenetic remodelling of cell states (Oliver *et al.*, 2020). The traditional view of GBM, and other cancers, is that a population of pre-existing CSCs within a unidirectional hierarchical system of cell states are largely responsible for driving adaptive therapy resistance (Yamada and Nakano, 2012; Prager *et al.*, 2020; Lathia *et al.*, 2015; Lan *et al.*, 2017). Some recent studies, however, contradict this, with GBM cells instead found to cluster into four or five distinct states, each of which is able transition to another (Neftel *et al.*, 2019; Couturier *et al.*, 2020). Part of this discrepancy in observations may result from a lack of reliable CSC markers, or confounding influences of unrealistic media conditions in which GBM cell lines are traditionally cultured, causing extensive phenotypic

differences between patient derived and cultured cell line derived stem cells (Lee *et al.*, 2006). Further studies instead suggest GBM cells resist therapy by transitioning to slow cycling persister states. A study using single-cell lineage tracing in stem-like GBM cells, with an amplification of platelet-derived growth factor receptor (*PDGFR*), found that after application of a TKI targeting PDGFR, the cells reversibly transitioned to resistant slow-cycling persister cell states. These were dependent on Notch signalling and characterised by widespread chromatin remodelling with upregulation of developmental programs. Furthermore, this occurred in combination with expansions of genetically resistant subclones containing focal amplifications of either insulin receptor substrate 1 or 2 (*IRS-1* or *IRS-2*), both of which conferred resistance to the PDGFR inhibitor when overexpressed. This therefore demonstrates sequential adaptive and innate or acquired resistance mechanisms in the same GBM cells (Eyler *et al.*, 2020; Liau *et al.*, 2017). Similarly, rapid and reversible transition to slow-cycling persister cell states has been observed in response to TMZ. These persister cells had increased expression of histone lysine demethylase genes relative to the primary tumours, and transiently acquired characteristics of more differentiated neuronal glial cell types (Banelli *et al.*, 2015). Additionally, when a cell line was exposed to TMZ, a small number of drug tolerant cells from across different genetic subpopulations survived, with dysregulation of most basal metabolism processes and reduced proliferation. These then expanded following epigenetic remodelling to acquire full resistance. In contrast to previous studies, this resistance was not reversible (Rabé *et al.*, 2020).

In an attempt to overcome the poor efficacy of standard therapy, studies have investigated alternative and more targeted treatment approaches, but with limited success. (An *et al.*, 2018; Touat *et al.*, 2017; Platten, 2017). EGFRvIII, in particular, has received a lot of attention (Yang *et al.*, 2017). The EGFRvIII vaccine rindopepimut failed recurrent GBM phase III clinical trials (Platten, 2017; Zussman and Engh, 2015), although it has shown promising survival benefits more recently in a phase II trial when used in combination with the vascular endothelial growth factor (VEGF) inhibitor bevacizumab in recurrent EGFRvIII-positive GBM (Reardon *et al.*, 2020). The EGFR antibody therapy cetuximab was used in combination with bevacizumab and irinotecan, a cytotoxic alkylid, in a phase II trial of recurrent GBM, but was found to have no benefit compared to the use of bevacizumab and irinotecan alone (Hasselbalch *et al.*, 2010). Chimeric antigen receptor (CAR) T cells targeting EGFRvIII showed anti-tumour activity in recurrent GBM patients in a phase I trial. However, they also induced adaptive resistance to the drug through significantly reduced expression of EGFRvIII in the tumours, along with increased expression of compensatory immunosuppressive molecule genes, including *IDO1*, *PD-L1*, and *IL-10* (O'Rourke *et al.*, 2017). Other drugs have included TKIs targeting EGFR. Only a minority of patients gain a significant benefit from these, with response found to be associated with co-expression of PTEN and EGFRvIII (Mellinghoff *et al.*, 2005). In a phase II clinical trial, the TKI dacomitinib generally showed poor results against recurrent glioblastoma, although with a significant benefit in 4 of the 49 patients, who remained progression free for at least 6 months. Various genetic resistance mechanisms have been identified for TKIs in GBM. One study



found that after receiving dacomitinib, a patient's recurrent GBM acquired a six base in-frame deletion in *EGFR*, resulting in their death just two months after therapy onset (Brastianos *et al.*, 2017). A different patient who received a TKI targeting MET due to their recurrent GBM containing over 70 copies of the *MET* gene, died 6 weeks later with no evidence of the amplification in the autopsy tumour sample (Brastianos *et al.*, 2017). Similarly, in mice, GBMs became resistant to the TKI erlotinib through elimination of *EGFRvIII* ecDNA, which then re-emerged clonally once treatment was stopped (Nathanson *et al.*, 2014). These examples illustrate the rapid changes that can occur in the tumour subclonal architecture in response to targeted therapies in GBM, and highlight the need for alternative approaches that are less susceptible to therapy resistance.

Owing to the limited successes of existing therapies, it is important to further characterise mechanisms associated with resistance in GBM, particularly against the standard cytotoxic chemoradiotherapy, where studies have shown a lack of a genetic bottleneck. Accumulating evidence suggests that adaptive epigenetic reprogramming of cell states is an important factor in driving such resistance. Nonetheless, genetic factors are still likely to play a role. Even if a strong genetic bottleneck is not seen, some subclones may show smaller selective advantages over others, or confer resistance in a minority of patients, as has been found previously (Wang *et al.*, 2016; Cahill *et al.*, 2007). It is also shown that specific gene alterations predispose cells to transition into particular cellular states seen in GBMs (Neftel *et al.*, 2019), and therefore they may influence the ability of cells to undergo adaptive resistance. There might also be pathways that are essential for the reprogramming process and detrimental when altered, thereby preventing cells with such mutations from surviving through to the recurrent. It's therefore important to characterise the effect of standard therapy on the presence and cellular frequencies of mutations in GBM. While studies have already identified individual genes that are more commonly altered in recurrent tumours in GBM (Kraboth and Kalman, 2020; Wang *et al.*, 2016), they have not applied extensive methods to investigate alterations across cellular pathways, and which take into account changes in cellular frequencies. Such analyses may inform on the processes that influence GBM cells' ability to survive through therapy, and potentially identify new therapeutic targets.

## 1.7 Models of GBM

In order for researchers to carry out functional analyses or assessment of drug efficacies in cancers, *in vitro* and *in vivo* models that sufficiently recapitulate patient tumour biology are required. However, obtaining such models is a challenge.

Traditional *in vitro* models of GBM, where cells are cultured in 2D, fail to account for the influences of microenvironmental features, such as the brain's extracellular matrix and circulatory structures, or interactions with non-cancer cells. More state-of-the-art innovative approaches are being developed to

overcome these limitations (Caragher *et al.*, 2019). Organotypic spheroids, where tissue taken directly from patient biopsies is grown in 3D culture, are able to preserve intact blood vessels and the tumour-associated extracellular matrix, as well as the presence of patient macrophages. Alternatively, scaffolds that artificially replicate the extracellular matrix of the brain are being developed for culturing cells in 3D. More advanced approaches include the creation of cerebral organoids from pluripotent stem cells, grown and differentiated under specific sequential media conditions to create what are being referred to as ‘mini-brains’. These newer methods of culturing GBM cells better recreate the original tumour biology, and have resulted in improved accuracies of *in vitro* drug screening (Caragher *et al.*, 2019).

*In vivo* cancer models are commonly used to avoid some of the limitations of traditional *in vitro* models. Use of such animal models is a controversial topic, not least because of the significant ethical issues with inflicting pain and suffering on animals (Joffe *et al.*, 2016; Ferdowsian and Beck, 2011). Additional concerns relate to the fact that the models often do not recapitulate human biology. Mouse models for cancers are commonly developed by injecting cancer cells from a patient biopsy into the same location in an animal, creating a patient-derived orthotopic xenograft (PDOX). Treatment of cells prior to injection, such as cell dissociation and passaging in culture, results in them drifting from the biology of the original biopsy. As well as undergoing transcriptomic and epigenomic changes, the cells may also undergo clonal evolution with selection of subclones, in a way that differs from that of the patient’s tumour (Ben-David *et al.*, 2017; Yoshida, 2020). As a result, these models show poor accuracy in predicting therapeutic responses in humans (Aldape *et al.*, 2019). Researchers are therefore aiming to improve how GBM PDOX’s are created, through minimising cell processing steps from biopsy to PDOX, with the goal of preserving more of the biology of the patient tumour (Golebiewska *et al.*, 2020; Yoshida, 2020).

In order for any *in vitro* or *in vivo* models to provide accurate predictions of GBM biology, they must be able to maintain the subclonal architecture of the originating patient biopsy. It is therefore necessary to characterise the ITH of matched models and biopsies, and investigate any changes between them. Such analysis can be achieved using the same computational methods as those for investigating GBM progression through therapy.

## 1.8 Computational investigation of ITH

The importance of ITH in tumour progression and response to therapies, makes characterising it a priority for researchers and clinicians. Reconstruction of the clonal architecture and evolutionary dynamics of tumours allows for predictions of their clinical course, as well as hypotheses on the consequences of individual mutations within them, and the role of cellular processes they affect. Many researchers aim to achieve this using genome sequencing of samples taken from patient tumours. Ideally this would include many samples per tumour so that all spatially distinct areas can be sufficiently represented. Longitudinal

plasma biopsies are also useful in tracking DNA from circulating cancer cells, and inferring changes in subclonal populations in response to therapies over time, including those from metastatic cells not detectable at the original tumour site (Dagogo-Jack and Shaw, 2018). Often, however, only a single bulk sample is available from a tumour, owing to financial constraints or tissue availability. As a result, a major limitation of any analysis into genetic ITH from these samples is subject to the confounding effects of sampling bias. This raises a number of issues, such as i) parts of a tumour's phylogenetic tree not being captured, hindering phylogenetic reconstruction, ii) inaccurate estimation of a mutation's prevalence in a tumour, particularly when subclonal mutations falsely appear clonal or completely absent, or iii) underestimation of the level of ITH. Such effects from single samples have been extensively assessed through spatial simulations of tumour evolution (Sun *et al.*, 2017; Chkhaidze *et al.*, 2019), or through multi-region sequencing (Bhandari *et al.*, 2018; Siegmund and Shibata, 2016; Watkins and Schwarz, 2018; Mahlokozera *et al.*, 2018).

Nonetheless, a large field of research has been undertaken to study ITH from both single and multi-sample datasets. One approach has been to quantify the extent of ITH in tumours in a simple metric, or from the numbers of detectable subclones (Mroz *et al.*, 2015). This allows observations to be made, such as the survival prognostic value of ITH or number of detectable subclones. Reports from these studies can be confusing, as many refer to a high or low number of detectable subclones as being the same as high or low ITH, which is not always the case; ITH is highest in neutral evolution, but due to the lack of selection and clonal expansions, subclones are often not detected (Figure 2). In general, increased numbers of detectable subclones have been associated with poor survival across a range of cancers, possibly reflecting an increased aggressiveness of the cells under selection (Morris *et al.*, 2016; Turajlic *et al.*, 2018; Davis *et al.*, 2017; Espiritu *et al.*, 2018). Conversely, very high numbers, beyond four subclones, is associated with increased survival, possibly as a result of the negative effect of genomic instability on tumour cell survival (Andor *et al.*, 2016). Other studies aim to investigate the specific events underlying ITH, through characterising individual subclones in a tumour by identifying the sets of mutations that define them, and their cellular frequencies in each sample. The ideal approach to achieve this would be through sequencing a representative proportion of individual cells in a tumour, so that unbiased quantification of distinct genotypes can be calculated. However, current high-throughput single-cell sequencing methods are expensive and suffer from extensive technical noise, including gene drop-outs (Kuipers *et al.*, 2017). Such datasets are therefore not readily available for most tumour samples. An alternative, and commonly performed approach, is to identify mutations from bulk tumour sequencing (either whole-exome (WES) or whole-genome (WGS), often with a matched normal sample), and then attempt to infer the underlying subclones from these through a process known as subclonal deconvolution. Phylogenetic reconstruction is then also often performed, in order to infer the evolutionary history of these subclones within the tumour.

Several large-scale projects have been undertaken to investigate tumour ITH and evolution. The TRACERx (TRACKing Cancer Evolution through therapy (Rx)) study has collected tumour specimens over multiple time points from hundreds of patients across lung, melanoma, prostate, and renal cancers, allowing unprecedented details into tumour evolution to be studied (Turajlic *et al.*, 2018; Mitchell, 2018; McGranahan *et al.*, 2016; Jamal-Hanjani, 2017). More recently, the ICGC-TCGA Pan-Cancer Analysis of Whole Genomes Project (PCAWG) has investigated ITH in 2,778 cancer samples from 2,658 patients, allowing comparisons between 38 cancer types, and identified cancer specific trends in tumour evolutionary trajectories (Gerstung *et al.*, 2020; D'Antonio *et al.*, 2020). This study largely confirmed previous reports of the ordering of common genetic alterations in primary GBMs, described in section 1.5.

Subclonal deconvolution has been a key feature in these projects to allow predictions of the clonal make up of tumours. Such an analysis can be achieved for a number of different types of mutations. The most common methods, and the type focussed on in this study, perform deconvolution of point variants. These utilise variant allele frequencies (VAFs) calculated from the ratios of reference to variant supporting reads, in order to estimate the proportions of cells containing each variant, known as their cancer cell fraction (CCF) (Figure 4). This inference of CCFs from VAFs is a challenging process due to a number of factors that distort a direct correlation between them, including tumour purity, technical noise, and overlapping CNAs, all for which require normalisation. Then, variants are grouped into distinct subclones, either by clustering CCFs or through mathematical phylogenetic modelling approaches. In themselves, CCFs are valuable to researchers, as they allow for inferences on the relative fitness conferred by variants through changes in CCFs between longitudinal samples (Barthel *et al.*, 2019). However, once deconvoluted into distinct subclones, these can be more reliably tracked between samples, or fitted into a phylogenetic tree, and observations can be made about co-occurring mutations within the same subclone.

Many methods are available for performing subclonal deconvolution of point variants (Eg. PyClone (Roth *et al.*, 2014), Ccube (Yuan *et al.*, 2018), Sclust (Cun *et al.*, 2018), MOBSTER (Caravagna *et al.*, 2020)), and for generating the mutation call set inputs (Eg. Mutect2 (Benjamin *et al.*, 2019), VarScan2 (Koboldt *et al.*, 2012), Strelka2 (Kim *et al.*, 2018), Lancet (Narzisi *et al.*, 2018), TITAN (Ha *et al.*, 2014), FACETS (Shen and Seshan, 2016), Sequenza (Favero *et al.*, 2015)), but all of which give highly varied results (Abécassis *et al.*, 2019; Andor *et al.*, 2016; Salcedo *et al.*, 2020; Bhandari *et al.*, 2018). Such methods therefore require benchmarking where the results obtained from each are compared to known ground-truths. This allows researchers to assess how reliable these analyses are, as well as which methods provide the most accurate predictions. While most mutation calling and subclonal deconvolution methods have undergone some level of previous benchmarking, many such studies have limitations, largely due to the difficulty in obtaining suitably realistic test data sets with known ground truths of the underlying mutation profiles. This requires computationally simulated datasets. However, available methods for creating these are limited in their ability to model the complexity of real tumour sequencing datasets, which include overlapping and flexibly

ordered point and structural mutations from multiple related subclones, with sequencing reads reflecting noise from both sequencing errors and alignment errors, and in the case of WES, variations in exon capture along the genome. These are all likely to impact the performance of subclonal deconvolution pipelines and therefore need to be modelled. Additionally, investigations into the suitability and effects of different mutation calling methods used to generate the inputted call sets, particularly allele-specific CNA callers, are often overlooked. Such limitations of previous subclonal deconvolution benchmarking studies are discussed in detail in Chapter 2.

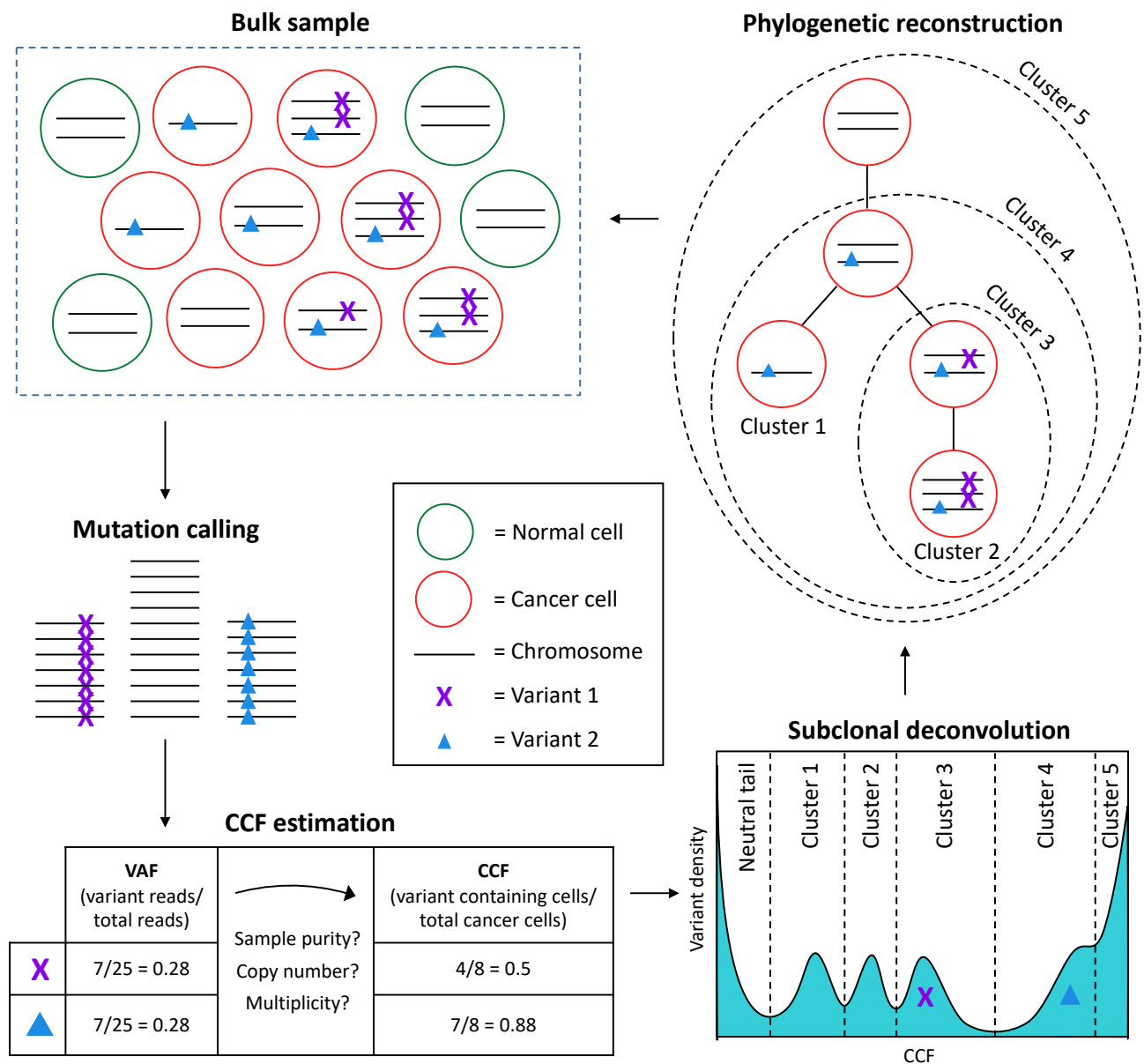


Figure 4. Illustration of the processes involved in subclonal deconvolution. **Bulk sample:** Circles represent normal (green) and cancerous (red) cells within the sample. The lines represent copies of a chromosome, which may differ from two copies in the cancer cells. Variants (cross and triangle) may be present on one, multiple, or no copies of a chromosome within a cell depending on the order of alterations and phylogenetic history of the cell. **Mutation calling:** The bulk sample is

sequenced and variant allele frequencies (VAFs) are identified from the proportions of sequencing reads that support each variant allele, compared to the total reads across those positions. **CCF estimation:** Cancer cell fractions (CCFs) are estimated from VAFs by normalising for sample purity, copy number and multiplicity. **Subclonal deconvolution:** CCFs are clustered into groups, thereby inferring variants that are present in the same subclones. The neutral tail cluster results from ongoing neutral evolution across all subclones. **Phylogenetic reconstruction:** CCF clusters are arranged onto a phylogenetic tree in order to infer the evolutionary history of a tumour. This history is reflected in the cells present in the bulk tumour.

## 1.9 Hypothesis

GBMs inevitably become resistant to standard chemoradiotherapy, and accumulating evidence suggests this is mediated largely through adaptive epigenetic reprogramming. Innate or acquired genetic factors are also likely to contribute to therapy resistance, either through influencing the ability of cells to undergo adaptive epigenetic reprogramming, or through distinct mechanisms.

Characterising ITH in paired primary and recurrent GBMs allows identification of changes in the presence, or cellular frequencies, of variants through therapy, and may indicate those that either increase or decrease the ability of cells to survive. Investigating such variants using pathway analysis could uncover cellular processes underlying resistance or sensitisation mechanisms, and which may provide novel targets for treating GBM.

Similar methods also allow for an assessment of the extent that cancer models maintain the subclonal architecture, and therefore biology, of the originating patient biopsies. Our collaborators at the NorLux Neuro-oncology laboratory based at the Luxembourg Institute of health, have developed glioma PDOX models using 3D organoids created without cell dissociation or passage. Characterising the ITH of these PDOXs, and assessing the level of correlation with the originating biopsies, will aid in determining their suitability as models.

## 1.10 Aims and objectives

I aim to investigate the effect of therapy on intratumour heterogeneity in GBM in order to uncover new targets for overcoming therapy resistance, using genome sequencing datasets for a set of matched primary and recurrent GBM samples. This approach involves mutation calling and subclonal deconvolution to characterise subclones within the tumours. However, it's not clear what the most suitable methods are for performing these processes, so I therefore first aim to carry out benchmarking to assess their accuracy. This requires developing novel methods to simulate artificial tumour sequencing reads with known ground truths, and which improve on existing methods in recapitulating the complexity of real tumours. I will then

use the simulated datasets with different analysis pipelines, and compare their outputs to the known ground truths. The best performing pipeline will then be applied to the real tumour datasets. I next aim to use the results from this to investigate how therapy affects subclone frequencies. This will include the use of a model for predicting whether subclones are undergoing selection in the tumours and whether this is associated with undergoing therapy, as well as performing pathway analysis to uncover specific cellular mechanisms that may be driving an increased or decreased resistance to therapy. Lastly, I aim to use subclonal deconvolution to determine whether PDOXs, created without cell dissociation or passage, maintain the intratumour heterogeneity of the originating patient biopsies.

The specific aims and objectives of each chapter are outlined below and illustrated in Figure 5.

## **Chapter 2 - Simulation of Heterogeneous Tumour genomes and *in silico* WES Data Sets**

Aim: Develop methods for simulating realistically complex artificial WES datasets for heterogeneous tumours.

Objectives:

- Develop a novel method for simulating heterogeneous tumour genomes, with mutations that recreate the complexity seen in real tumours.
- Improve an existing method for *in silico* WES, to allow creation of artificial sequencing datasets from simulated genomes, with realistic read distributions.

## **Chapter 3 – Benchmarking of mutation calling and subclonal deconvolution methods**

Aim: Determine the accuracies of different pipelines for mutation calling and subclonal deconvolution.

Objectives:

- Create test datasets with known ground truths using the methods developed in Chapter 2.
- Run somatic variant calling and CNA calling methods on the test datasets and assess the accuracy of these methods by comparing their outputs to the known ground truths.
- Provide the called variants and CNAs as input into subclonal deconvolution methods, and assess the accuracy of these methods by comparing their outputs to known ground truths.

## **Chapter 4 – Identification of pathways relevant to GBM progression through therapy**

Aim: Use subclonal deconvolution, or alternative methods, to characterise GBM progression through therapy, and identify variants and pathways that may confer increased resistance or sensitivity.

Objectives:

- Use the SubClonalSelection model with VAFs from paired primary and recurrent GBMs, to predict whether the tumours are evolving neutrally or under selection through therapy.
- Identify variants that show evidence of being in subclones that increase or decrease in frequency through therapy.
- Perform pathway gene set enrichment analysis (GSEA) on the identified variants, to highlight pathways that are candidates for conferring increased resistance or sensitivity to therapy.

### **Chapter 5 – Comparison of intratumour heterogeneity between GBM patient biopsies and patient-derived orthotopic xenografts**

Aim: Use subclonal deconvolution to examine the extent to which GBM PDOXs, created without cell dissociation or passage, are able to maintain the clonal architecture of the originating patient biopsies.

Objectives:

- Develop and apply pipelines for calling variants and CNAs from targeted sequencing and array Comparative Genomic Hybridisation (aCGH) datasets, from paired biopsy and PDOX GBM samples.
- Perform subclonal deconvolution using the called variants and CNAs, to generate variant CCFs.
- Determine the correlations of variant CCFs between biopsies and PDOXs, and investigate those that differ.



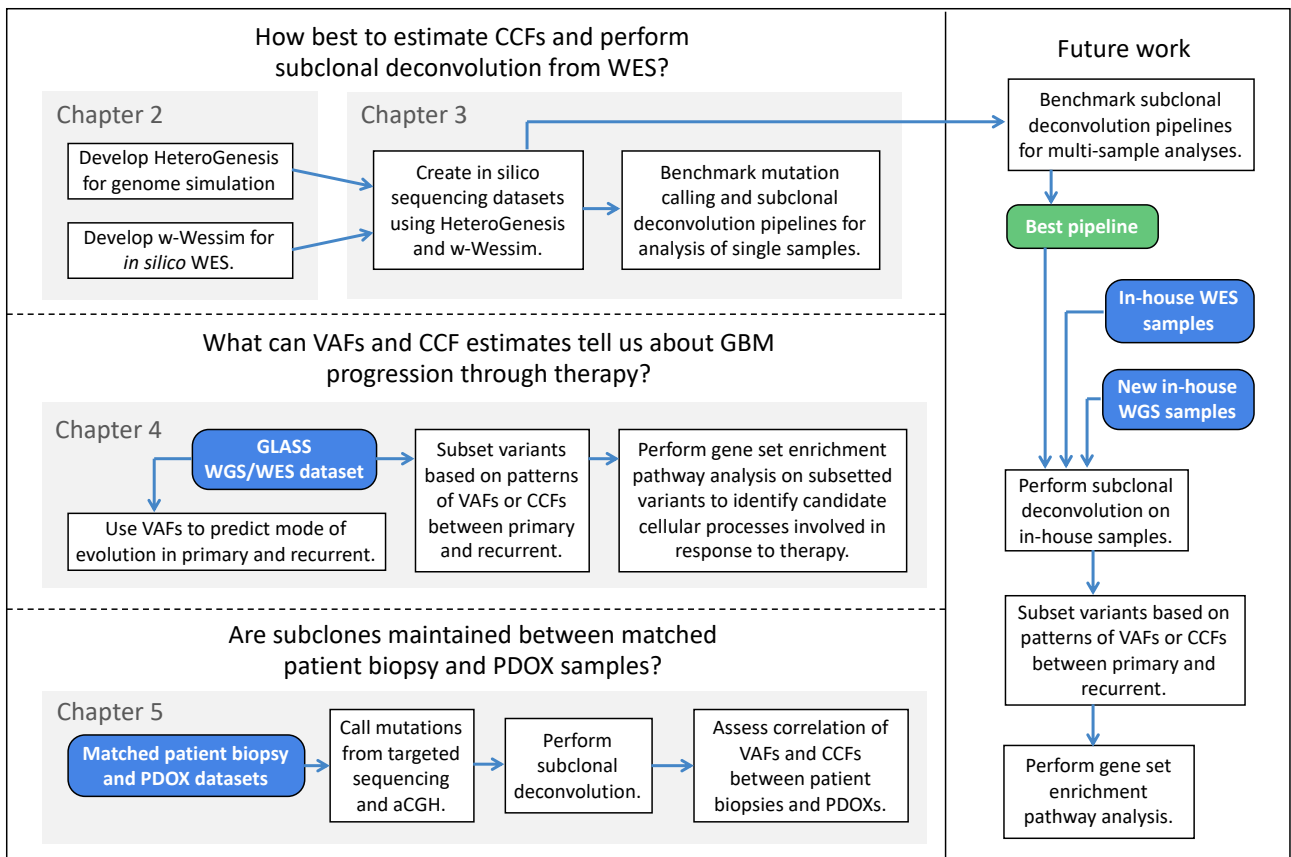


Figure 5. A diagram of the objectives involved in each of the chapters in this study, as well as future planned analyses that will follow on from, and are made possible, by this work.

## 1.11 References

- Abécassis, J. *et al.* (2019) Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data. *PLoS One*, **14**.
- Aldape, K. *et al.* (2019) Challenges to curing primary brain tumours. *Nat. Rev. Clin. Oncol.*, **16**, 509–520.
- Ali, M.Y. *et al.* (2020) Radioresistance in Glioblastoma and the Development of Radiosensitizers. *Cancers (Basel)*, **12**, 2511.
- Allocati, N. *et al.* (2018) Glutathione transferases: Substrates, inhibitors and pro-drugs in cancer and neurodegenerative diseases. *Oncogenesis*, **7**, 1–15.
- An, Z. *et al.* (2018) Epidermal growth factor receptor and EGFRvIII in glioblastoma: Signaling pathways and targeted therapies. *Oncogene*, **37**, 1561–1575.
- Andor, N. *et al.* (2016) Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.*, **22**, 105–13.
- Aziz, M.H. *et al.* (2011) Acquisition of p53 mutations in response to the non-genotoxic p53 activator Nutlin-3. *Oncogene*, **30**, 4678–4686.
- Banelli, B. *et al.* (2015) The histone demethylase KDM5A is a key factor for the resistance to temozolomide in glioblastoma. *Cell Cycle*, **14**, 3418–3429.
- Barthel, F.P. *et al.* (2019) Longitudinal molecular trajectories of diffuse glioma in adults. *Nature*, **576**, 112–120.
- Barthel, F.P. *et al.* (2018) Reconstructing the molecular life history of gliomas. *Acta Neuropathol.*, **135**, 649–670.

- Ben-David,U. *et al.* (2017) Patient-derived xenografts undergo mouse-specific tumor evolution. *Nat. Genet.*, **49**, 1567–1575.
- Benjamin,D. *et al.* (2019) Calling Somatic SNVs and Indels with Mutect2.
- Bhandari,V. *et al.* (2018) Quantifying the Influence of Mutation Detection on Tumour Subclonal Reconstruction. *bioRxiv*, 418780.
- Brastianos,P.K. *et al.* (2017) Resolving the phylogenetic origin of glioblastoma via multifocal genomic analysis of pre-treatment and treatment-resistant autopsy specimens. *npj Precis. Oncol.*, **1**, 33.
- de Bruin,E.C. *et al.* (2014) Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science (80-. )*, **346**, 251 LP – 256.
- Byers,L.A. *et al.* (2013) An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin. Cancer Res.*, **19**, 279–290.
- Cahill,D.P. *et al.* (2007) Loss of the mismatch repair protein MSH6 in human glioblastomas is associated with tumor progression during temozolomide treatment. *Clin. Cancer Res.*, **13**, 2038–2045.
- Caragher,S. *et al.* (2019) Glioblastoma’s next top model: Novel culture systems for brain cancer radiotherapy research. *Cancers (Basel)*, **11**.
- Caravagna,G. *et al.* (2020) Subclonal reconstruction of tumors by using machine learning and population genetics. *Nat. Genet.*, **52**, 898–907.
- Ceccarelli,M. *et al.* (2016) Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*, **164**, 550–563.
- Chien,C.-H. *et al.* (2019) Role of autophagy in therapeutic resistance of glioblastoma. *J. Cancer Metastasis Treat.*, **2019**.
- Chkhaidze,K. *et al.* (2019) Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *bioRxiv*, 544536.
- Christians,A. *et al.* (2019) The prognostic role of IDH mutations in homogeneously treated patients with anaplastic astrocytomas and glioblastomas. *Acta Neuropathol. Commun.*, **7**, 156.
- Comaills,V. *et al.* (2016) Genomic Instability Is Induced by Persistent Proliferation of Cells Undergoing Epithelial-to-Mesenchymal Transition. *Cell Rep.*, **17**, 2632–2647.
- Couturier,C.P. *et al.* (2020) Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nat. Commun.*, **11**.
- Cree,I.A. and Charlton,P. (2017) Molecular chess? Hallmarks of anti-cancer drug resistance. *BMC Cancer*, **17**, 10.
- Cross,W.C.H. *et al.* (2016) New paradigms in clonal evolution: punctuated equilibrium in cancer. *J. Pathol.*, **240**, 126–136.
- Cun,Y. *et al.* (2018) Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust. *Nat. Protoc.*, **13**, 1488–1501.
- Dagogo-Jack,I. and Shaw,A.T. (2018) Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.*, **15**, 81–94.
- Darmanis,S. *et al.* (2017) Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell Rep.*, **21**, 1399–1410.
- Davis,A. *et al.* (2017) Tumor evolution: Linear, branching, neutral or punctuated? *Biochim. Biophys. Acta - Rev. Cancer*, **1867**, 151–161.
- Dawson,C.C. *et al.* (2011) ‘Persisters’: Survival at the cellular level. *PLoS Pathog.*, **7**.
- Decarvalho,A.C. *et al.* (2018) Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. *Nat. Genet.*, **50**, 708–717.
- Dentro,S.C. *et al.* (2020) Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *bioRxiv*, 312041.

- Domenichini,A. *et al.* (2019) ABC transporters as cancer drivers: Potential functions in cancer development. *Biochim. Biophys. Acta - Gen. Subj.*, **1863**, 52–60.
- Drabløs,F. *et al.* (2004) Alkylation damage in DNA and RNA - Repair mechanisms and medical significance. *DNA Repair (Amst)*., **3**, 1389–1407.
- Eldredge,N. *et al.* (2005) The dynamics of evolutionary stasis. *Paleobiology*, **31**, 133–145.
- Ellis,H.P. *et al.* (2019) Clinically Actionable Insights into Initial and Matched Recurrent Glioblastomas to Inform Novel Treatment Approaches. *J. Oncol.*, **2019**.
- Espiritu,S.M.G. *et al.* (2018) The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. *Cell*, **173**, 1003-1013.e15.
- Eyler,C.E. *et al.* (2020) Single-cell lineage analysis reveals genetic and epigenetic interplay in glioblastoma drug resistance. *Genome Biol.*, **21**.
- Farkona,S. *et al.* (2016) Cancer immunotherapy: The beginning of the end of cancer? *BMC Med.*, **14**.
- Favero,F. *et al.* (2015) Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, **26**, 64–70.
- Ferdowsian,H.R. and Beck,N. (2011) Ethical and Scientific Considerations Regarding Animal Testing and Research. *PLoS One*, **6**, e24059.
- Gerstung,M. *et al.* (2020) The evolutionary history of 2,658 cancers. *Nature*, **578**, 122–128.
- Golebiewska,A. *et al.* (2020) Patient-derived organoids and orthotopic xenografts of primary and recurrent gliomas represent relevant patient avatars for precision oncology. *Acta Neuropathol.*, **10**, 16.
- Ha,G. *et al.* (2014) TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.*, **24**, 1881–93.
- Hasselbalch,B. *et al.* (2010) Cetuximab, bevacizumab, and irinotecan for patients with primary glioblastoma and progression after radiation therapy and temozolomide: A phase II trial. *Neuro. Oncol.*, **12**, 508–516.
- Hassen,W. *et al.* (2015) Drug metabolism and clearance system in tumor cells of patients with multiple myeloma. *Oncotarget*, **6**, 6431–6447.
- Hegi,M.E. *et al.* (2005) MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.*, **352**, 997–1003.
- Huang,G. *et al.* (2007) Expression of human glutathione S-transferase P1 mediates the chemosensitivity of osteosarcoma cells. *Mol. Cancer Ther.*, **6**, 1610–1619.
- Huang,R.X. and Zhou,P.K. (2020) DNA damage response signaling pathways and targets for radiotherapy sensitization in cancer. *Signal Transduct. Target. Ther.*, **5**, 1–27.
- ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**, 82–93.
- Jamal-Hanjani,M. (2017) Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.*, **376**, 2109–2121.
- Joffe,A.R. *et al.* (2016) The ethics of animal research: A survey of the public and scientists in North America. *BMC Med. Ethics*, **17**, 17.
- Johnson,D.R. and O’Neill,B.P. (2012) Glioblastoma survival in the United States before and during the temozolomide era. *J. Neurooncol.*, **107**, 359–364.
- Kim,H. *et al.* (2020) Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat. Genet.*, 1–7.
- Kim,S. *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.
- Koboldt,D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–76.
- Körber,V. *et al.* (2019) Evolutionary Trajectories of IDH WT Glioblastomas Reveal a Common Path

- of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis. *Cancer Cell*, **35**, 692–704.e12.
- Krboth,Z. and Kalman,B. (2020) Longitudinal Characteristics of Glioblastoma in Genome-Wide Studies. *Pathol. Oncol. Res.*, **26**, 2035–2047.
- Kruger,S. *et al.* (2019) Advances in cancer immunotherapy 2019 - Latest trends. *J. Exp. Clin. Cancer Res.*, **38**.
- Kuipers,J. *et al.* (2017) Advances in understanding tumour evolution through single-cell sequencing. *Biochim. Biophys. Acta - Rev. Cancer*, **1867**, 127–138.
- Lan,X. *et al.* (2017) Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature*, **549**, 227–232.
- Lathia,J.D. *et al.* (2015) Cancer stem cells in glioblastoma. *Genes Dev.*, **29**, 1203–17.
- Lee,J. *et al.* (2006) Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell*, **9**, 391–403.
- Lee,Joo Ho *et al.* (2018) Human glioblastoma arises from subventricular zone cells with low-level driver mutations. *Nature*, **560**, 243–247.
- Lee,S.Y. (2016) Temozolomide resistance in glioblastoma multiforme. *Genes Dis.*, **3**, 198–210.
- Li,Y. *et al.* (2020) Patterns of somatic structural variation in human cancer genomes. *Nature*, **578**, 112–121.
- Liau,B.B. *et al.* (2017) Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance. *Cell Stem Cell*, **20**, 233-246.e7.
- Louis,D.N. *et al.* (2016) The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol.*, **131**, 803–820.
- Mahlokozera,T. *et al.* (2018) Biological and therapeutic implications of multisector sequencing in newly diagnosed glioblastoma. *Neuro. Oncol.*, **20**, 472–483.
- Marenco-Hillebrand,L. *et al.* (2020) Trends in glioblastoma: outcomes over time and type of intervention: a systematic evidence based analysis. *J. Neurooncol.*, **147**, 297–307.
- Marusyk,A. *et al.* (2014) Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature*, **514**, 54–8.
- McGranahan,N. *et al.* (2016) Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science (80-. )*, **351**, 1463–1469.
- McGranahan,N. and Swanton,C. (2017) Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*, **168**, 613–628.
- Mellinghoff,I.K. *et al.* (2005) Molecular determinants of the response of glioblastomas to EGFR kinase inhibitors. *N. Engl. J. Med.*, **353**, 2012–2024.
- Michaelis,M. *et al.* (2011) Adaptation of cancer cells from different entities to the MDM2 inhibitor nutlin-3 results in the emergence of p53-mutated multi-drug-resistant cancer cells. *Cell Death Dis.*, **2**, e243.
- Mitchell,T. (2018) Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx Renal. *Cell*, **173**, 611–623.
- Mohammad,R.M. *et al.* (2015) Broad targeting of resistance to apoptosis in cancer. *Semin. Cancer Biol.*, **35**, S78–S103.
- Morris,L.G.T. *et al.* (2016) Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*, **7**, 10051–63.
- Mroz,E.A. *et al.* (2015) Intra-tumor Genetic Heterogeneity and Mortality in Head and Neck Cancer: Analysis of Data from The Cancer Genome Atlas. *PLoS Med.*, **12**.
- Narzisi,G. *et al.* (2018) Genome-wide somatic variant calling using localized colored de Bruijn graphs. *Commun. Biol.*, **1**, 1–9.
- Nathanson,D.A. *et al.* (2014) Targeted therapy resistance mediated by dynamic regulation of

- extrachromosomal mutant EGFR DNA. *Science (80-. )*, **343**, 72–76.
- Neftel,C. *et al.* (2019) An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell*, **178**, 835-849.e21.
- Nelson,C.M. and Bissell,M.J. (2006) Of extracellular matrix, scaffolds, and signaling: tissue architecture regulates development, homeostasis, and cancer. *Annu. Rev. Cell Dev. Biol.*, **22**, 287–309.
- Newburger,D.E. *et al.* (2013) Genome evolution during progression to breast cancer. *Genome Res.*, **23**, 1097–1108.
- Noble,R. *et al.* (2019) Spatial structure governs the mode of tumour evolution. *bioRxiv*, 586735.
- Noorani,I. *et al.* (2020) PiggyBac mutagenesis and exome sequencing identify genetic driver landscapes and potential therapeutic targets of EGFR-mutant gliomas. *Genome Biol.*, **21**, 181.
- O'Rourke,D.M. *et al.* (2017) A single dose of peripherally infused EGFRvIII-directed CAR T cells mediates antigen loss and induces adaptive resistance in patients with recurrent glioblastoma. *Sci. Transl. Med.*, **9**.
- Oliver,L. *et al.* (2020) Drug resistance in glioblastoma: are persisters the key to therapy? *Cancer Drug Resist.*, **3**, 287–301.
- Ostrom,Q.T. *et al.* (2016) CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2009–2013. *Neuro. Oncol.*, **18**, v1–v75.
- Pasello,M. *et al.* (2008) Overcoming glutathione S-transferase P1-related cisplatin resistance in osteosarcoma. *Cancer Res.*, **68**, 6661–6668.
- Perry,A. and Wesseling,P. (2016) Histologic classification of gliomas. In, *Handbook of Clinical Neurology*. Elsevier, pp. 71–95.
- Pflaum,J. *et al.* (2014) P53 family and cellular stress responses in cancer. *Front. Oncol.*, **4**.
- Platten,M. (2017) EGFRvIII vaccine in glioblastoma-InACT-IVE or not ReACTive enough? *Neuro. Oncol.*, **19**, 1425–1426.
- Prager,B.C. *et al.* (2020) Glioblastoma Stem Cells: Driving Resilience through Chaos. *Trends in Cancer*, **6**, 223–235.
- Rabé,M. *et al.* (2020) Identification of a transient state during the acquisition of temozolomide resistance in glioblastoma. *Cell Death Dis.*, **11**, 1–14.
- Ramirez,M. *et al.* (2016) Diverse drug-resistance mechanisms can emerge from drug-tolerant cancer persister cells. *Nat. Commun.*, **7**.
- Reardon,D.A. *et al.* (2020) Rindopepimut with bevacizumab for patients with relapsed EGFRvIII-expressing glioblastoma (REACT): Results of a double-blind randomized phase II trial. *Clin. Cancer Res.*, **26**, 1586–1594.
- Rice,T. *et al.* (2016) Understanding inherited genetic risk of adult glioma – a review. *Neuro-Oncology Pract.*, **3**, 10–16.
- Riganti,C. *et al.* (2014) Temozolomide down-regulates P-glycoprotein in human blood-brain barrier cells by disrupting Wnt3 signaling. *Cell. Mol. Life Sci.*, **71**, 499–516.
- Rode,A. *et al.* (2016) Chromothripsis in cancer cells: An update. *Int. J. Cancer*, **138**, 2322–2333.
- Roth,A. *et al.* (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396–8.
- Salcedo,A. *et al.* (2020) A community effort to create standards for evaluating tumor subclonal reconstruction. *Nat. Biotechnol.*, **38**, 97–107.
- Sanborn,J.Z. *et al.* (2015) Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 10995–1000.
- Sepúlveda-Sánchez,J.M. *et al.* (2018) SEOM clinical guideline of diagnosis and management of low-grade glioma (2017). *Clin. Transl. Oncol.*, **20**, 3–15.
- Shaffer,S.M. *et al.* (2017) Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, **546**, 431–435.

- Sharma,P. *et al.* (2017) Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy. *Cell*, **168**, 707–723.
- Shen,R. and Seshan,V.E. (2016) FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.*, **44**, e131–e131.
- Shibue,T. and Weinberg,R.A. (2017) EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. *Nat. Publ. Gr.*
- Siegmund,K. and Shibata,D. (2016) At least two well-spaced samples are needed to genotype a solid tumor. *BMC Cancer*, **16**, 250.
- Singh,S. (2015) Cytoprotective and regulatory functions of glutathione S-transferases in cancer cell proliferation and cell death. *Cancer Chemother. Pharmacol.*, **75**, 1–15.
- Sottoriva,A. *et al.* (2017) Catch my drift? Making sense of genomic intra-tumour heterogeneity. *Biochim. Biophys. Acta - Rev. Cancer*, **1867**, 95–100.
- Sottoriva,A. *et al.* (2013) Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 4009–4014.
- Struve,N. *et al.* (2020) EGFRvIII upregulates DNA mismatch repair resulting in increased temozolomide sensitivity of MGMT promoter methylated glioblastoma. *Oncogene*, **39**, 3041–3055.
- Stupp,R. *et al.* (2017) Effect of tumor-treating fields plus maintenance temozolomide vs maintenance temozolomide alone on survival in patients with glioblastoma a randomized clinical trial. *JAMA - J. Am. Med. Assoc.*, **318**, 2306–2316.
- Stupp,R. *et al.* (2009) Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol.*, **10**, 459–466.
- Stupp,R. *et al.* (2005) Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.*, **352**, 987–996.
- Sun,R. *et al.* (2017) Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.*, **49**, 1015–1024.
- Tamimi,A.F. and Juweid,M. (2017) Epidemiology and Outcome of Glioblastoma. In, *Glioblastoma*. Codon Publications, pp. 143–153.
- Thakkar,J.P. *et al.* (2014) Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiol. Biomarkers Prev.*, **23**, 1985–1996.
- The Cancer Genome Atlas (TCGA) Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
- Touat,M. *et al.* (2017) Glioblastoma targeted therapy: updated approaches from recent biological insights. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, **28**, 1457–1472.
- Turajlic,S. *et al.* (2018) Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell*, **173**, 595-610.e11.
- Turner,K.M. *et al.* (2017) Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature*, **543**, 122–125.
- Verhaak,R.G.W. *et al.* (2019) Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nat. Rev. Cancer*, **19**, 283–288.
- Verhaak,R.G.W. *et al.* (2010) Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98–110.
- Wang,J. *et al.* (2016) Clonal evolution of glioblastoma under therapy. *Nat. Genet.*, **48**, 768–776.
- Wang,Q. *et al.* (2017) Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell*, **32**, 42-56.e6.
- Wang,Y. *et al.* (2014) Clonal evolution in breast cancer revealed by single nucleus genome

- sequencing. *Nature*, **512**, 155–160.
- Watkins, T.B.K. and Schwarz, R.F. (2018) Phylogenetic Quantification of Intratumor Heterogeneity. *Cold Spring Harb. Perspect. Med.*, **8**, a028316.
- Wesseling, P. and Capper, D. (2018) WHO 2016 Classification of gliomas. *Neuropathol. Appl. Neurobiol.*, **44**, 139–150.
- West, J. *et al.* (2019) Tissue structure accelerates evolution: premalignant sweeps precede neutral expansion. *bioRxiv*, 542019.
- Wise, D.R. *et al.* (2011) Hypoxia promotes isocitrate dehydrogenase-dependent carboxylation of  $\alpha$ -ketoglutarate to citrate to support cell growth and viability. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 19611–19616.
- Wynne, P. *et al.* (2007) Enhanced repair of DNA interstrand crosslinking in ovarian cancer cells from patients following treatment with platinum-based chemotherapy. *Br. J. Cancer*, **97**, 927–933.
- Xu, K. *et al.* (2019) Structure and evolution of double minutes in diagnosis and relapse brain tumors. *Acta Neuropathol.*, **137**, 123–137.
- Xu, W. *et al.* (2011) Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of  $\alpha$ -ketoglutarate-dependent dioxygenases. *Cancer Cell*, **19**, 17–30.
- Yamada, R. and Nakano, I. (2012) Glioma stem cells: Their role in chemoresistance. *World Neurosurg.*, **77**, 237–240.
- Yang, J. *et al.* (2017) Targeting EGFRvIII for glioblastoma multiforme. *Cancer Lett.*, **403**, 224–230.
- Ye, D. *et al.* (2018) Metabolism, Activity, and Targeting of D- and L-2-Hydroxyglutarates. *Trends in Cancer*, **4**, 151–165.
- Yecies, D. *et al.* (2010) Acquired resistance to ABT-737 in lymphoma cells that up-regulate MCL-1 and BFL-1. *Blood*, **115**, 3304–3313.
- Yi, G.Z. *et al.* (2019) Acquired temozolomide resistance in MGMT-deficient glioblastoma cells is associated with regulation of DNA repair by DHC2. *Brain*, **142**, 2352–2366.
- Yoshida, G.J. (2020) Applications of patient-derived tumor xenograft models and tumor organoids. *J. Hematol. Oncol.*, **13**, 1–16.
- Yu, K. *et al.* (2020) Surveying brain tumor heterogeneity by single-cell RNA-sequencing of multi-sector biopsies. *Natl. Sci. Rev.*, **7**, 1306–1318.
- Yu, M. *et al.* (2013) Reversal of ATP-binding cassette drug transporter activity to modulate chemoresistance: Why has it failed to provide clinical benefit? *Cancer Metastasis Rev.*, **32**, 211–227.
- Yuan, K. *et al.* (2018) Ccube: A fast and robust method for estimating cancer cell fractions. *bioRxiv*, 484402.
- Zack, T.I. *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, **45**, 1134–1140.
- Zussman, B.M. and Engh, J.A. (2015) Outcomes of the ACT III study: Rindopepimut (CDX-110) therapy for glioblastoma. *Neurosurgery*, **76**, N17.

# Chapter 2 – Simulation of Heterogeneous Tumour genomes and *in silico* WES Data Sets

The majority of the work presented in this chapter was originally published by Oxford University Press in Bioinformatics (Tanner *et al.*, 2019). The material is reproduced under the Creative Commons CC BY license.

## 2.1 Introduction

### 2.1.1 Overview

Analysis of intratumour heterogeneity is commonly achieved through mutation calling and subclonal deconvolution from whole-exome sequencing (WES). The different methods and pipelines for performing such analyses have been found to give highly conflicting results (Abécassis *et al.*, 2019; Andor *et al.*, 2016; Salcedo *et al.*, 2020; Bhandari *et al.*, 2018). Reliable benchmarking, using datasets with known ground truths, is therefore required to determine which are the most accurate. However, such studies have not been extensively carried out for these analyses. As discussed below, previous benchmarking of both mutation calling and subclonal deconvolution has suffered from various limitations, not least due to difficulties in generating realistic artificial sequencing datasets with known ground truths, on which to test the methods. I therefore aimed to carry out more reliable benchmarking, requiring the development of two new computational methods to generate artificial datasets. This chapter describes these methods, which simulate tumour genomes and create artificial WES datasets from them, both of which significantly improve on existing methods.

In the literature, different terms are used to distinguish mutations that occur in an individual's inherited germline genome and mutations that develop in somatic tissue throughout life, owing to their differing relevance in biology. Germline point and copy number mutations are generally referred to as 'single nucleotide polymorphisms' (SNPs) and 'copy number variants' (CNVs), whereas the equivalent somatic mutations are referred to as 'single nucleotide variants' (SNVs) and 'copy number alterations' (CNAs). In order to simplify explanations of the simulation methods I developed, in this chapter only, I refer to both germline and somatic point and copy number mutations as 'single nucleotide variants' (SNVs) and 'copy number variants' (CNVs), and specifically state when referring to only germline or somatic mutations. Similarly, I refer to either germline or somatic small insertions and deletions, less than 50 bases in length, as InDels. In addition, I refer to mutations whereby a whole chromosome or genome has been replicated, as aneuploid events.



### 2.1.2 Limitations of existing benchmarking studies

The reliability of subclonal deconvolution is dependent on both accurate mutation profiles, consisting of variant allele frequencies (VAFs) of SNVs/InDels, and copy numbers at those positions, as well as accurate algorithms to process these without significantly limiting assumptions (Abécassis *et al.*, 2019; Noorbakhsh *et al.*, 2018; Bhandari *et al.*, 2018; Andor *et al.*, 2016). Therefore, in addition to subclonal deconvolution methods, it is also important to benchmark the methods used to call somatic SNVs/InDels and CNVs for input into the analyses. Numerous such studies have previously been carried out for all three processes but all have certain limitations. Benchmarking requires a dataset for which the ground truth is known, or estimated. This is generally achieved through one of two ways; 1) Using real datasets for which mutation calls or cancer cell fractions (CCFs) have been validated through one of a number of approaches, but that do not provide certainty of the ground truths, or 2) using artificial simulated datasets that represent given known ground truths, but which require simplifying assumptions to make generating them feasible.

Studies assessing the performance of somatic SNV/InDel calling algorithms have demonstrated a substantial lack of consensus between them. A comparison of eight different algorithms found that, of 29634 total SNVs called from WES of five breast cancer tumours, only 1348 were called by five or more algorithms, with 22032 called by only one (Krøigård *et al.*, 2016). Another study found that, of 2035 total SNVs called by four callers from WES of a chronic myeloid leukaemia tumour, only 36 were called by all algorithms and 1757 were called by only one, with many of these being assigned a high probability score. In the same study, when comparing matched germline-germline data generated by randomly splitting a single BAM file into two (as opposed to matched germline-tumour data as used in a real analysis), all four algorithms reported between 5-11 false positive SNVs, most of which had probability scores of greater than 0.90 (Roberts *et al.*, 2013).

Benchmarking of SNV/InDel calling methods using simulated data has generally involved spiking in variants at varying frequencies into aligned real sequencing reads (Hansen *et al.*, 2013; Cibulskis *et al.*, 2013; Xu *et al.*, 2014; Lai *et al.*, 2016; Christoforides *et al.*, 2013; Bian *et al.*, 2018; Detering *et al.*, 2019; Narzisi *et al.*, 2018). This, however, negates the effect of alignment errors and scores, as mutations are always inserted at the correct corresponding position of the reference genome. Other studies have avoided this issue through *in silico* sequencing of simulated genomes, though these are reliant on accurate error models for replicating realistic sequencing artefact distributions (Bohnert *et al.*, 2017; Stead *et al.*, 2013; Detering *et al.*, 2019). Alternatively, benchmarking approaches using real data, do so by validating variants either through resequencing, combining variant calling methods to create high confidence calls, or manual visual curation (Cibulskis *et al.*, 2013; Hansen *et al.*, 2013; Spinella *et al.*, 2016; Xu *et al.*, 2014; Krøigård *et al.*, 2016; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes, 2020; Narzisi *et al.*, 2018), but these approaches do not provide certainty of the ground truths, particularly in determining missed variants. The above

approaches, involving simulated and real data, were both used in one of the largest variant calling benchmarking studies, the TCGA-ICGC DREAM-3 Somatic Mutation Calling Challenge, which employs crowdsourcing to compare callers, and has been used by developers to test and fine tune their own methods (Ewing *et al.*, 2015). Other approaches involve mixing two germline sequencing datasets in varying proportions, where one represents the tumour fraction, and the other represents normal germline contamination (Chen *et al.*, 2020; Kim *et al.*, 2018; Bohnert *et al.*, 2017), with some making use of datasets from projects where germline variants have been characterised through pedigree analysis (Eberle *et al.*, 2017). This could bias against some variant callers, such as MuTect2 (Benjamin *et al.*, 2019), that take into account known positions of germline polymorphisms when distinguishing between somatic and germline variants. Another limitation of most of these studies, is that they use the hg19 human reference genome, which was last patched in 2013 and results in less accurate read alignment with around 5% fewer SNVs called than when using the same pipeline with the more recent hg38 genome (Pan *et al.*, 2019). Furthermore, the use of differing parameters or application of filters makes judging the performance of variant calling algorithms from existing studies difficult. For example, VarScan's default minimum VAF required to call a variant is 0.2 (v2.2.2 and earlier) (Koboldt *et al.*, 2012), resulting in very poor performance in detecting low frequency variants in studies that benchmarked callers with default parameters. As a result of these numerous sources of biases, SNV/InDel benchmarking studies provide highly conflicting results, leaving researchers with uncertainty as to the most suitable methods to use.

Previous studies on somatic CNV calling methods have shown drastically varying results, both between different callers and between different studies. Benchmarking has often been carried out using real data with CNVs validated through more reliable methods than WES, such as WGS or SNP arrays (Shen and Seshan, 2016; Nam *et al.*, 2016; Zare *et al.*, 2017; Favero *et al.*, 2015), though these are still susceptible to significant errors (Zhang *et al.*, 2019; Pitea *et al.*, 2018; Pinto *et al.*, 2011; Telenti *et al.*, 2016). Alternatively, studies have used simulated datasets, although these suffer from either issues with incorporating mutations into aligned real reads, as mentioned above, or from unrealistic *in silico* WES sequencing of simulated genome sequences. This latter issue results from the fact that the only available method that aims to model WES in a more sophisticated way than simply limiting *in silico* WGS to exome regions, Wessim (Kim *et al.*, 2013), still has significant limitations. Most importantly, through testing I found that it is unable to model the effect of amplification CNVs, creating the same number of reads in these regions as those in diploid regions. It is, however, able to model deletions. This is apparent from going through the program's code, but it is not stated in the documentation. Wessim has therefore been used in CNV calling benchmarking studies. An example of such a study tested somatic CNV calling methods on Wessim simulated datasets, and real TCGA endometrial carcinoma data validated with SNP array (Rieber *et al.*, 2017). This showed poor to good sensitivity for most methods in detecting deletions from both datasets, and good sensitivities for detecting amplifications from the TCGA data, but found sensitivities of

(practically) 0 for detecting amplifications from the Wessim dataset. This led the authors to speculate that the difference in sensitivities for amplifications was due to the real data containing more high-copy number amplifications which are easier to detect, whereas instead, I believe it just reflects Wessim not modelling amplifications. Another benchmarking study used VarSimLab (Hosny, 2017) to simulate data, which incorporates mutations into a reference genome, followed by *in silico* sequencing using a BED file (adapted to take into account CNVs shifting positions) to limit this to exome regions. As discussed in section 2.2.2, this approach to *in silico* WES does not realistically model real WES data. An additional limitation of existing CNV benchmarking studies is that few have focussed on methods that estimate allele-specific CNVs, where major and minor copy number states are given and which are often required for subclonal deconvolution. Also, few have focussed on those that detect subclonal CNVs, also required for some subclonal detection methods. Benchmarking that specifically assesses the accuracy of these processes, and with simulated datasets that realistically model WES, is therefore necessary.

Benchmarking of subclonal deconvolution methods has been carried out through a number of different approaches. Those using real data have usually involved single-cell sequencing to determine a ground truth (Jiang *et al.*, 2016; Roth *et al.*, 2014; Abécassis *et al.*, 2019), although this has not been applied on a large scale and may be affected by sampling bias or single-cell sequencing inaccuracies (Lähnemann *et al.*, 2020). An alternative approach has been to simulate mutation call sets that fit given ground truths, with errors added from simple distributions to represent imperfect mutation calling (Dentro *et al.*, 2020; El-Kebir *et al.*, 2016; Jiang *et al.*, 2016; Li and Li, 2014; Cun *et al.*, 2018; Körber *et al.*, 2019). This is unlikely to realistically capture the inaccuracies seen in real mutation call sets, which are influenced by both the mutation callers and aligner used (Alioto *et al.*, 2015). In addition, assumptions are made to simplify the simulation of call sets, meaning the complexity of real tumours is not fully modelled. Such assumptions include not modelling CNVs at all (Miura *et al.*, 2018), subclonal mutations only able to lie on one copy of a chromosome (Dentro *et al.*, 2020), or no more than two copy number states per region. These simplifications may be especially problematic given that most of the benchmarking using this approach has been carried out by the authors of the subclonal deconvolution methods, who therefore risk inadvertently fitting the data to their methods.

A third approach, and the one that I carry out in the next chapter, is to use mutation call sets from running mutation calling on simulated sequencing datasets that represent tumours with known ground truths. This has the benefit that it is closer to the real process than simply simulating mutation calls, and likely better captures the mutation calling errors specific to each caller. A major study applying such benchmarking is the “ICGC-TCGA DREAM Somatic Mutation Calling - Tumour Heterogeneity Challenge” (Salcedo *et al.*, 2020). This employs crowd sourcing to benchmark subclonal deconvolution methods from a set of simulated sequencing datasets. However, as yet, only the results of the in-house benchmarking has been published, consisting of only two subclonal deconvolution methods (PhyloWGS (Deshwar *et al.*, 2015) and DPCLust (Nik-Zainal *et al.*, 2012)), with one CNV caller (Battenburg (Nik-Zainal *et al.*, 2012)) and four SNV

callers, which do not include the most commonly used method, MuTect2 (Benjamin *et al.*, 2019). In addition, the simulated dataset is of WGS, which has different challenges for analysis in comparison to WES, and therefore the accuracies of the benchmarked methods are likely not transferable when used for WES data. As I demonstrate in section 2.2.2, existing *in silico* sequencing methods for WES, of which I am aware, are unable to generate realistic datasets suitable for benchmarking subclonal deconvolution methods. Furthermore, the method used to simulate the sequencing datasets in the DREAM challenge study, as well as all other somatic genome simulation methods that I identified, have drawbacks that limit their suitability for use in benchmarking subclonal deconvolution methods. The following section describes these in more detail.

### 2.1.3 Limitations of existing somatic genome simulation programs

Numerous somatic genome simulation programs exist, and these generally fall into two categories; 1) Those that take an input genome sequence (generally a reference genome) and incorporate mutations into it for each subclone in a tumour, which then require *in silico* sequencing to generate artificial sequencing reads from them, and 2) those that take an alignment of real sequencing reads, with mutations incorporated either as point variants into the read sequences, or as copy number variants by adjusting local read coverages along the genome through downsampling.

Somatic genomes show high intratumour heterogeneity, and contain extensive variation and rearrangement. In order to benchmark methods that aim to uncover these events, it's important to include this complexity in the test datasets. I identified 5 scenarios found in real tumours, that challenge mutation calling and subclonal deconvolution methods, but which existing somatic genome simulation programs lack the ability to model fully. These include:

- i. **Multi-level subclone phylogenies:** (Figure 6A) Subclones in tumours have complex phylogenetic architectures (Sottoriva *et al.*, 2013; McPherson *et al.*, 2016; Watkins and Schwarz, 2018). Current simulation tools create: no subclones (Hosny, 2017; Mu *et al.*, 2015), single layer phylogenies (Qin *et al.*, 2015; Xia *et al.*, 2017), or hierarchical structures but with no access to intermediate level genomes (Ivakhno *et al.*, 2017). More complex phylogenies can be created through iterative running of these tools, however this creates issues with keeping track of variant positions with respect to a stable reference. Xome-Blender (Semeraro *et al.*, 2018), SVEngine (Xia *et al.*, 2018) and the ICGC-TCGA DREAM BAMSurgeon wrap around script (Salcedo *et al.*, 2020) are exceptions to this, being able to create complex phylogenies, but instead succumb to other points mentioned below.
- ii. **Individual chromosome and whole-genome aneuploidy:** Both individual chromosome and whole-genome aneuploid events are common in cancer (Baysan *et al.*, 2017; Ben-David and Amon, 2020) but are not included in some existing tumour genome simulators. It's possible to get around this with some

programs by adding in CNVs that span entire chromosomes, although, depending on the program, this could then cause issues with having other overlapping CNVs on the same chromosome.

- iii. **Overlapping copy number variants (CNVs):** (Figure 6B) Given that tumours contain numerous CNVs, often reaching 10s of megabases in length (Krijgsman *et al.*, 2014; Tan *et al.*, 2014) it is likely that many will overlap, either i) nested within the same copy of a chromosome, ii) partially or fully on different copies within the same cell, or iii) partially or fully on copies in separate subclones. Many existing simulation tools do not allow for this.
  
- iv. **Mutations occurring in a flexible order:** (Figure 6C) Real somatic genomes acquire different types of variants in a flexible and varied order. As a result, a single nucleotide variant (SNV) may appear in only one, or in multiple copies of a replicated region, depending on whether it occurred before or after a copy number variant (CNV) or aneuploid event. This is referred to as the multiplicity of a variant and estimating it correctly for each variant is key to the accuracy of subclonal deconvolution methods. However, many existing simulation programs incorporate different types of variants in separate stages. For example, Pysim-sv (Xia *et al.*, 2017) generates all aneuploid events prior to SNVs, and all SNVs/InDels prior to CNVs. Therefore, aneuploid copies of a chromosome in a clone won't share any common variants, and SNVs will always be present in every copy of an overlapping CNV region on a chromosome.
  
- v. **Distinct germline and somatic variants:** The majority of SNVs and InDels in human germline genomes are at known polymorphic loci (1000 Genomes Project Consortium *et al.*, 2015; Shen *et al.*, 2013), recorded in dbSNP (Sherry *et al.*, 2001). Some somatic SNV/InDel callers make use of this information in determining the confidence with which an apparently tumour specific variant is assigned as somatic (Fan *et al.*, 2016; Cibulskis *et al.*, 2013). Such callers will be biased against when applied to simulated germline genomes without a proportion of variants at polymorphic loci. Likewise, the approach used by Xome-Blender (Semeraro *et al.*, 2018), which simulates tumour genome sequencing reads by re-assigning true germline variants as somatic, would result in biased metrics for SNV/InDel calling. Other simulation tools do not include somatic SNVs/InDels at all (Qin *et al.*, 2015; Xia *et al.*, 2018).

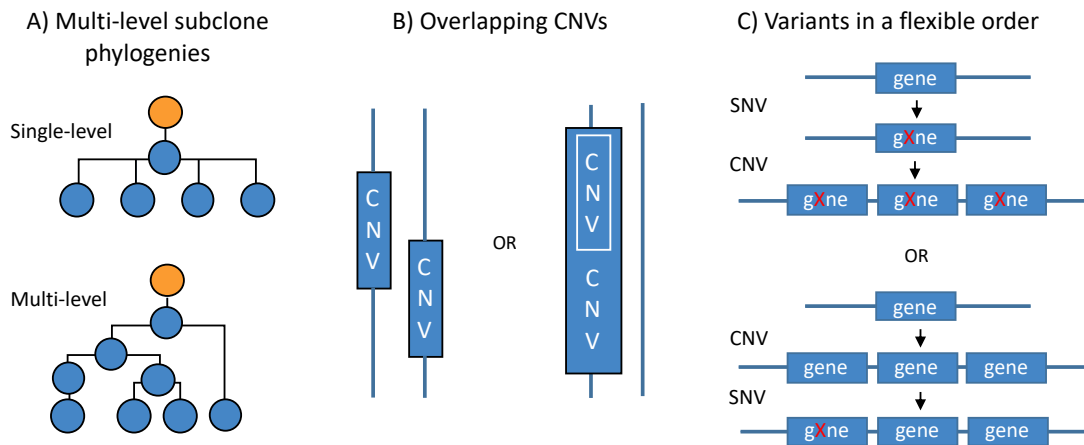


Figure 6. Illustration of certain features lacking from many somatic genome simulation programs.

A) The ability to create multilevel phylogenies with complex relationships between subclones, instead of just a single level of subclones. B) Overlapping CNVs, occurring on either separate copies of the same chromosome (in the same or different cells), or on the same copy, with partial or full overlaps. C) Variants occurring in a flexible order, so that an SNV or InDel may be present on either one or multiple copies of a region, depending on whether it came before or after a CNV or aneuploid event.

Some somatic genome simulators incorporate somatic variants directly into real sequencing data, as opposed to generating genome sequences (Semeraro *et al.*, 2018; Ivakhno *et al.*, 2017; Ewing *et al.*, 2015; Salcedo *et al.*, 2020). One such method is BAMSurgeon (Ewing *et al.*, 2015). The ICGC-TCGA-DREAM team have used a wrapper script for this method to create datasets for crowd-sourced benchmarking of subclonal deconvolution methods in their ‘Somatic Mutation Calling Challenge --Tumor Heterogeneity and Evolution’. The wrapper script adds features such as variant phasing, subclonality, and CNVs (Salcedo *et al.*, 2020). While this avoids the need for *in silico* sequencing, inaccuracies in read alignment are not reflected in the incorporated mutations. Additionally, the effect that variants have on WES library preparation during probe hybridisation, is not taken into account. Further problems are introduced by being limited in the copy number of simulated CNVs by the coverage depth of the inputted data.

Table 2. Features modelled by existing somatic simulation tools. ‘X?’ indicates that the feature is not mentioned in the accompanying documentation and I could find no evidence of it within the program, so is likely not to be included. The ability to simulate aneuploid events has not been indicated here as it is not easy to identify to what extent this is possible in many programs. VarSimLab, which was previously mentioned, is not included as it does not create distinct germline and somatic genomes.

Feature	SCNVSim (Qin <i>et al.</i> , 2015)	VarSim (Mu <i>et al.</i> , 2015)	tHapMix (Ivakhno <i>et al.</i> , 2017)	Pysim-sv (Xia <i>et al.</i> , 2017)	Xome-Blender (Semeraro <i>et al.</i> , 2018)	SVEngine (Xia <i>et al.</i> , 2018)	BAM Surgeon wrapper (Salcedo <i>et al.</i> , 2020)
Multi-level phylogenies	X	X	X	X	✓	✓	✓
Flexible variant order	X	X?	X?	X	X?	X?	✓
Overlapping CNVs	X?	X?	X?	X?	X?	X?	X?
Distinct germline and somatic SNVs/InDels	X	✓	✓	✓	X	X?	✓
Generates genome sequences	✓	✓	X	✓	X	✓	X

None of the existing somatic genome simulators identified are able to model all the scenarios required to reliably benchmark subclonal deconvolution pipelines (Table 2). I therefore aimed to create a new program for somatic genome simulation, and also improve *in silico* WES, so that a suitable dataset can be created for more reliable benchmarking of mutation calling and subclonal deconvolution pipelines.

## 2.2 Methods and Results

### 2.2.1 Creation of a program for simulating realistically complex tumour genomes

#### 2.2.1.1 HeteroGenesis Overview

In order to simulate genomes for sufficiently complex tumours for use in benchmarking, I required a new somatic genome simulator. I therefore developed HeteroGenesis, a program for generating heterogeneous tumour genomes, that overcomes the above limitations of previous methods.

HeteroGenesis is written in Python3, and is publicly available at

<https://github.com/GeorgetteTanner/HeteroGenesis>.

### 2.2.1.2 HeteroGenesis Workflow

HeteroGenesis consists of three consecutive modules (Figure 7):

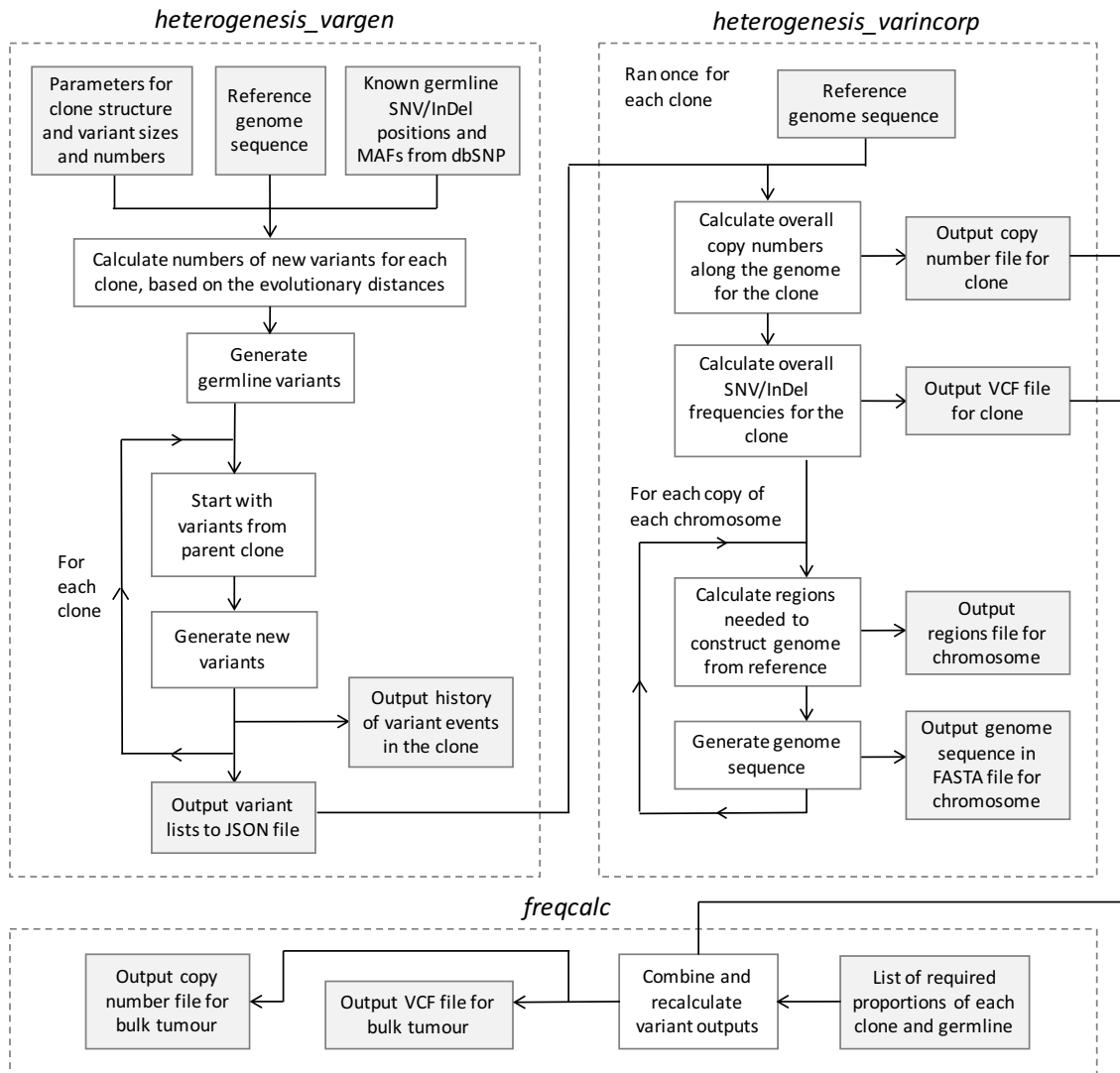


Figure 7. The workflow of HeteroGenesis. *heterogenesis\_vargen* first generates lists of mutations for the germline and each somatic clone in the tumour. *heterogenesis\_varincorp* then incorporates these mutations into a reference genome and calculates variant frequencies and copy numbers along the genome for a given clone. *freqcalc* can then be used to calculate overall bulk tumour mutation profiles.

#### 2.2.1.2.1 heterogenesis\_vargen

*heterogenesis\_vargen* generates lists of mutations (SNVs, InDels, CNVs and aneuploid events) to be incorporated into the genomes for each clone in a tumour, along with a matched germline. It takes as input: i) a reference FASTA genome sequence, ii) an optional variant call format (VCF) file containing known germline SNV and InDel locations and their minor allele frequencies (MAFs) from dbSNP, and iii) a JSON file containing a set of parameters. It outputs a JSON file with lists of mutations for each clone in the simulated



tumour (herein also referring to the matched germline, which is considered the germline 'clone'), as well as files containing the order that mutations occurred. The user is able to define: i) the subclonal structure, ii) the number of somatic aneuploid events, iii) rates of SNVs and InDels, iv) the length distributions of InDels, and v) the number and length distributions of CNVs. Separate parameter values are set for germline and total somatic variants. Users can also choose whether, and to what extent, to incorporate known germline SNVs/InDels into the simulated germline genome, weighted by MAF, or to sample all, or a proportion of germline or somatic mutations from user provided predefined list.

The clone structure of a tumour is defined by giving, for each clone ( $C_i$ ), its direct parent clone and a value representing the evolutionary distance from it ( $D_i$ ). These values are used to determine the set of mutations that a clone initially inherits, as well as the proportions of the total somatic mutations, ( $T$ ), that are assigned as new mutations in a clone, reflecting how far it has evolved from its parent. Therefore the number of new somatic mutations in a clone,  $C_i$ , is defined by  $T \frac{D_i}{\sum_1^n D_i}$ . This allows the user full control over the subclone phylogenetic architecture of a tumour (Figure 15).

CNV (>50 bp) and InDel ( $\leq 50$  bp) lengths follow scaled log normal distributions, which have been observed in real data from both our and other groups (Droop *et al.*, 2018; Krijgsman *et al.*, 2014), with user defined parameters for the mean and variance of the underlying normal distribution, and a scaling factor. Each copy of DNA in a CNV has an equal chance of being inserted in the forward or reverse direction. All default values for mutation parameters are chosen to reflect estimates from real human germline (Mills *et al.*, 2006; 1000 Genomes Project Consortium *et al.*, 2015; Durbin *et al.*, 2010; Shen *et al.*, 2013) and tumour genomes (specifically from glioblastoma) (Baysan *et al.*, 2017; Hu *et al.*, 2013; Kandoth *et al.*, 2013; Krijgsman *et al.*, 2014; Xi *et al.*, 2011; Mills *et al.*, 2006; Mullaney *et al.*, 2010; Droop *et al.*, 2018). Rearrangements are not simulated, although CNVs (in either forward and reverse order) replicate a similar challenge for analyses aiming to detect structural rearrangements and break points.

The program first determines the total numbers of each type of somatic mutations required in the final tumour and randomly splits these between somatic clones, based on the evolutionary distances between them from the provided parameters. Mutations are then generated for each clone, starting with the germline clone. Each clone is initiated with all the mutations inherited from its parent clone (with the root clone inheriting mutations from the germline clone) and new mutations are then added in a random order with respect to mutation type. Only the following are disallowed for pragmatic reasons: i) SNVs and InDels cannot occur more than once at a base on the same copy of a chromosome in a clone, ii) CNVs or InDel deletions cannot partially overlap on the same copy of a chromosome in a clone (fully overlapping on the same chromosome, or partially/fully overlapping on different copies of a chromosome, can occur), and iii) no mutation can occur within a deleted region, even if there are additional copies of the region on the chromosome (from a CNV) that haven't been deleted (though these may contain mutations that precede

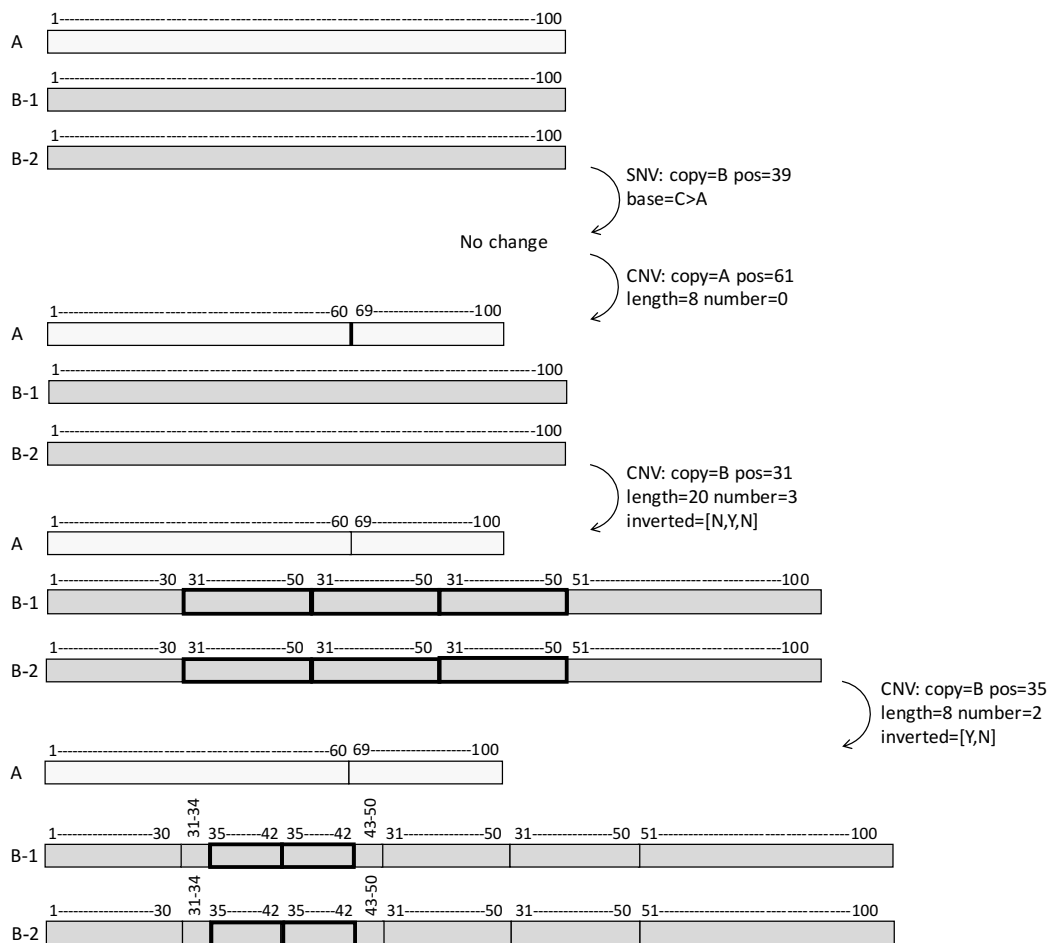
the deletion). Chromosomes are selected at random for placing mutations, taking length into account, except for aneuploid events where all chromosomes are selected with equal probability. Mutations are initially placed on either of two sets of chromosomes, thereby simulating a diploid genome. However, after an aneuploid replication event has occurred, additional copies of that chromosome, containing the same set of existing mutations, are then available for further mutations to be incorporated. Similarly, when a deletion aneuploid event has occurred, the deleted chromosome is no longer available for mutation placement and is not written to the outputs.

*heterogenesis\_vargen* takes 6hrs and 5GB RAM on a single thread to run under default parameters, which includes a germline and 2 somatic clones.

#### 2.2.1.2.2 *heterogenesis\_varincorp*

*heterogenesis\_varincorp* is run separately for each clone. It takes the lists of mutations generated by *heterogenesis\_vargen* and incorporates them into a reference genome sequence, as well as calculating copy numbers and mutation frequencies along the genome. This is done by sequentially using the mutations in the list to update three items:

- i. **cnblocks:** (Figure 8) Lists of chromosomal regions (herein referred to as blocks) for each copy of a chromosome, used to calculate copy numbers. Each list is initiated with a single block equal to the length of the chromosome. Blocks are updated each time a new CNV is incorporated by splitting them if they cross the CNV breakpoints and either replicating all blocks between the breakpoints (CNV replication) or removing them (CNV deletion). As direction is not relevant to copy number calculations, inversion information is ignored. After all CNVs have been incorporated, the number of blocks in all copies of a chromosome that correspond to each region are combined. This gives the overall copy number status along each chromosome, which is then written to a tab-delimited file. Deletion InDels (<50bp) are not taken into account.



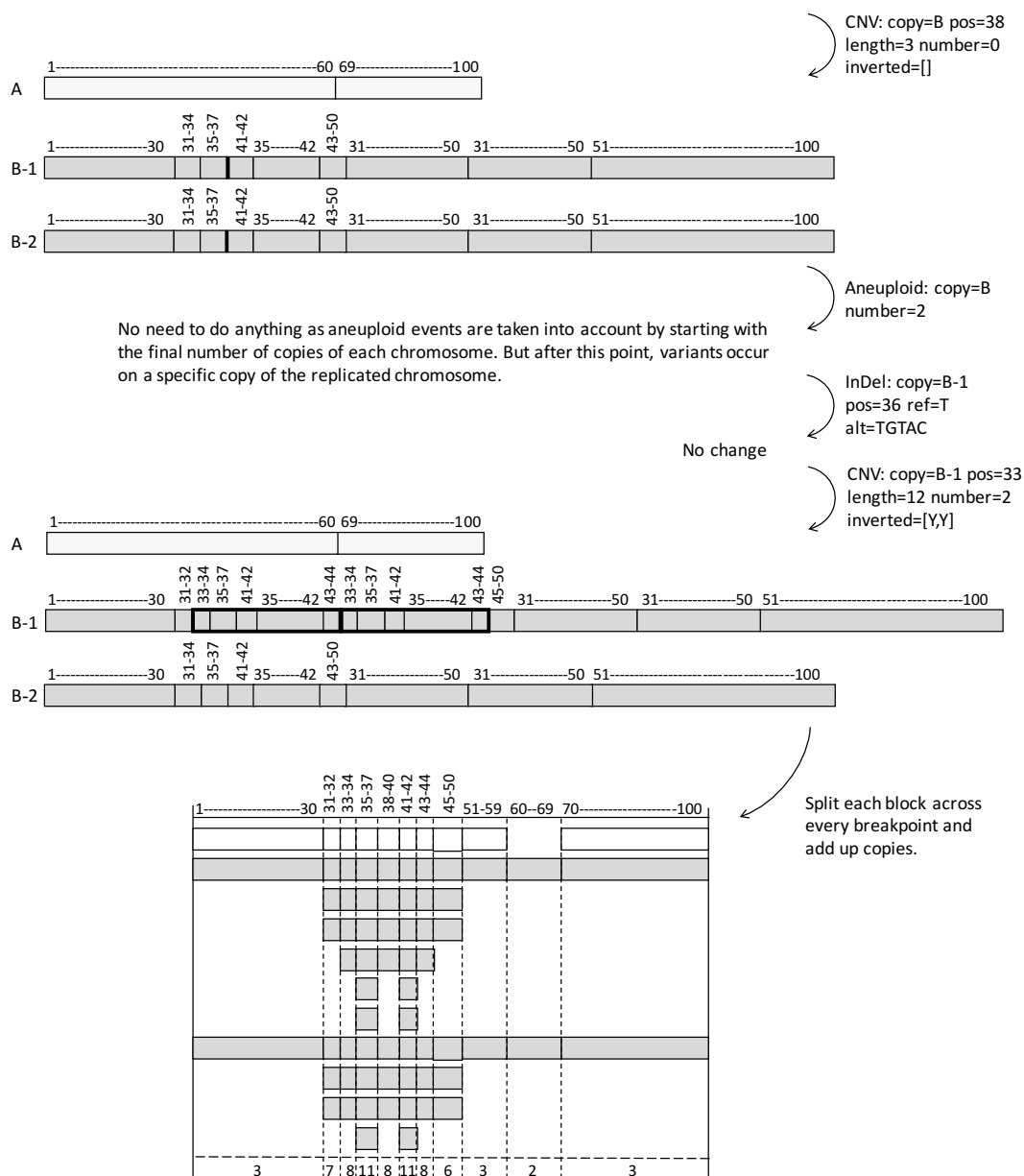
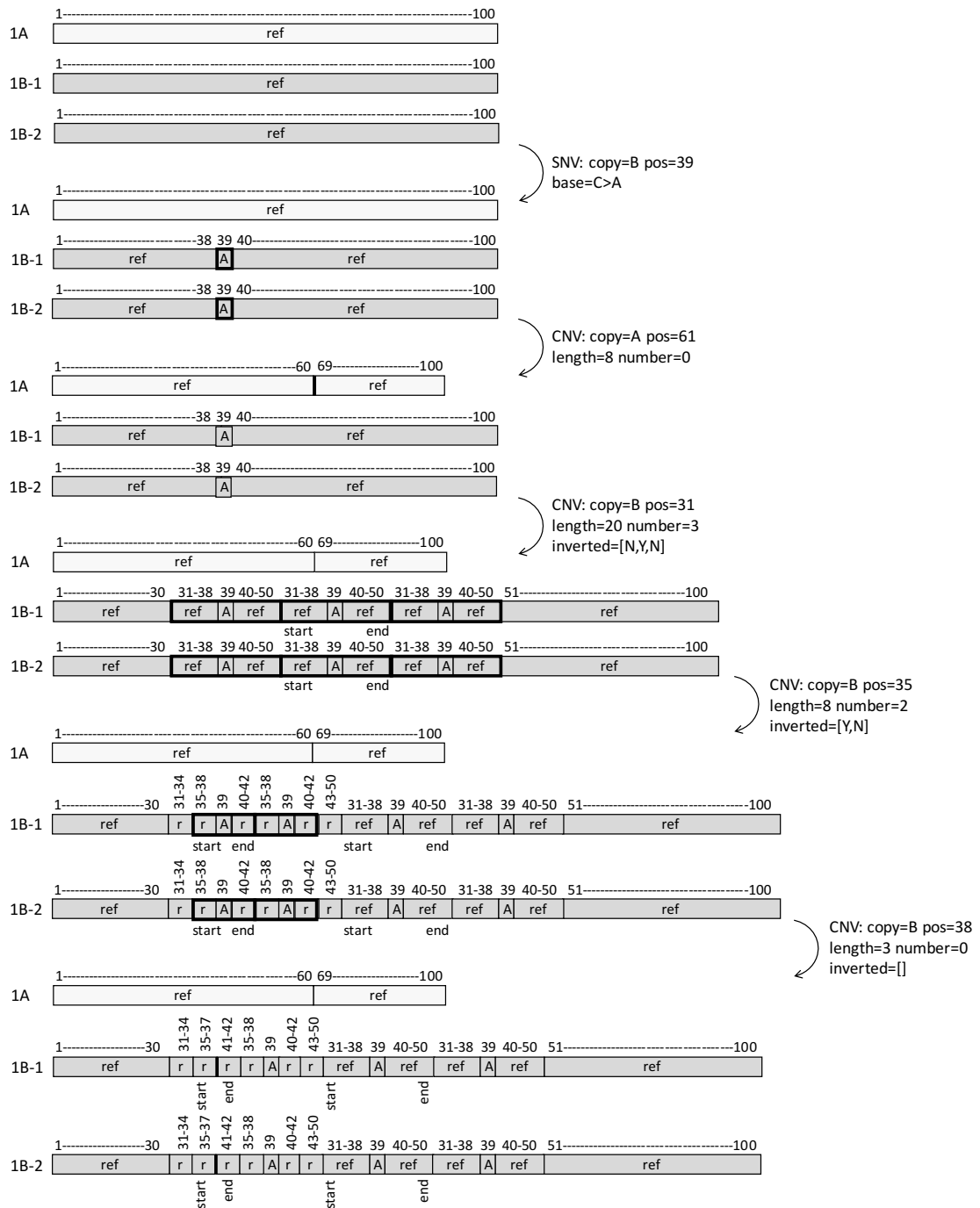


Figure 8. An illustration of how cnblocks is updated and used to calculate copy number along a genome given a list of mutations. This example represents a diploid genome with a single chromosome, 100bp in length. Initially, chromosomes are represented as a single block corresponding to the full length of the chromosome. The B chromosome copy undergoes a duplication aneuploid event at one point and is therefore represented by two blocks at the start. The blocks then undergo sequential modifications to reflect inputted CNVs and aneuploid events, with bold boxes and lines indicating the latest CNVs incorporated. All other mutations are ignored in this process. Mutations affecting the B chromosome prior to the aneuploid event are applied to both sets of the blocks, whereas, mutations occurring after the aneuploid event are applied only one set of blocks. Once all CNVs are incorporated, identical copies of each region are added up to indicate copy numbers along the genome.

ii. **allblocks:** (Figure 9) Analogous to cnblocks, but also includes blocks representing SNVs and InDels, and flags for starts and ends of inverted regions are recorded. The allblocks list is used to generate the simulated genome sequence by acting as a blueprint. For each block, the genome sequence is extended with either the corresponding reference sequence between the block's start and end positions, or the alternate allele sequence. When an inversion start flag appears, the succeeding sequence is held separately until an end flag appears, at which point the held sequence is inverted, translated into the complimentary sequence, and added onto the main sequence, or to a previously held sequence if there is an overlapping inverted sequence. After all blocks for a chromosome have been passed, the sequence is written to a FASTA file.



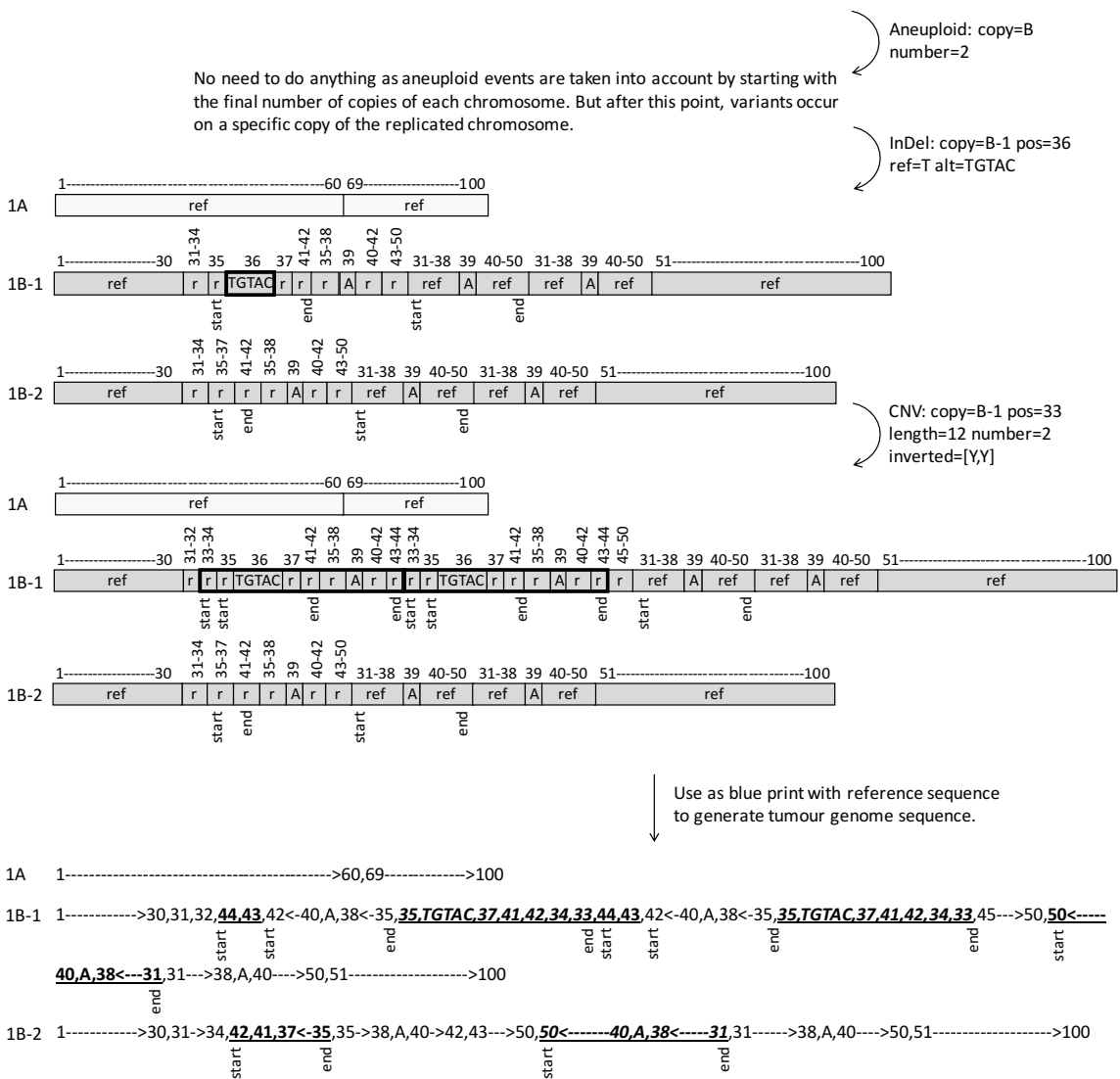


Figure 9. An illustration of how allblocks is updated and used to generate a FASTA sequence given a list of mutations. This example represents a diploid genome with a single chromosome, 100bp in length. Initially, chromosomes are represented as a single block corresponding to the full length of the chromosome. The B chromosome copy undergoes a duplication aneuploid event at one point and is therefore represented by two blocks at the start. The blocks then undergo sequential modifications to reflect all inputted mutations, with bold boxes and lines indicating the latest mutation. 'start' and 'end' flags indicate where inversions occur. Mutations affecting the B chromosome prior to the aneuploid event are applied to both sets of the blocks, whereas, mutations occurring after the aneuploid event are applied only one set of blocks. After all mutations are incorporated, the DNA sequences for the simulated chromosomes are generated from the reference genome using the blocks as a blueprint; 'ref'/'r' indicates to get the sequence between the block's start and end positions from the reference sequence, and eg. 'TGTAC' indicates to incorporate those bases into the simulated sequence. Underlined bold regions indicate those that have been inverted and where the complementary base sequence will instead

be incorporated into the simulated sequence. Regions that have been inverted twice, from overlapping inversions, are incorporated as normal in the forward direction.

- iii. **vcfcounts:** (Figure 10) Lists of incorporated SNVs and InDels for each copy of a chromosome, with the number of occurrences recorded for each. Each SNV/InDel also has information recorded on the position of CNVs that overlap them. This enables calculation of how many occurrences of an SNV/InDel to replicate/remove based on whether a new CNV falls within or around a previous CNV. Once all mutations have been incorporated, the vcfcounts list for all copies of a chromosome are combined, with numbers of occurrences for shared SNV/InDels added together. The overall copy number at each SNV/InDel position is taken from the combined cnblocks list and used with the total number of occurrences to calculate VAFs. These are then written to a VCF file.

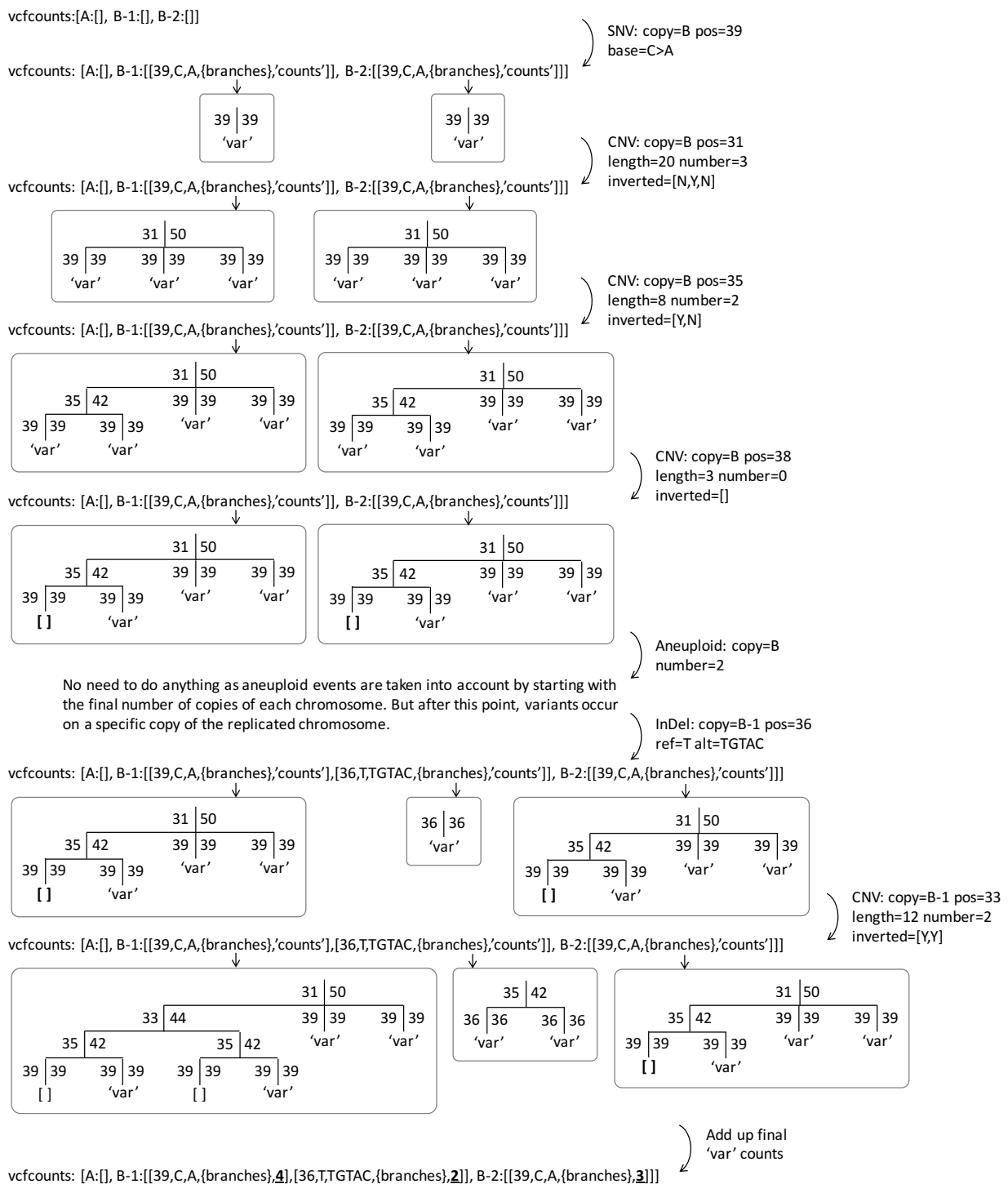


Figure 10. An illustration of how vcfcouunts is updated and used to calculate VAFs given a list of mutations. The tree diagrams represent the information that is contained in the 'branches' slot for each variant listed in vcfcouunts. Downward lines represent a copy of the region, with values giving the start and end positions. Each succeeding level shows either the presence of one variant ('var'), an absence of a variant from a deletion ('[ ]'), or the CNVs contained within the above region.



*heterogenesis\_varincorp* takes 6hr and 8GB RAM on a single thread to run 'clone1' of the output from *heterogenesis\_vargen* ran under default parameters. Clones and chromosomes may be run separately across different threads.

#### 2.2.1.2.3 *freqcalc*

*freqcalc* is provided as an accessory tool within HeteroGenesis and is used to calculate overall bulk tumour variant profiles that reflect user defined proportions of each clone in a sample. Furthermore, this allows users to assess how different sampling approaches affect analysis methods, as it can generate 1) samples with normal cell contamination, and 2) multiple different longitudinal, multi-region, or metastatic samples with varying proportions of clones in each. The ability to create such samples is important, as both tumour purity and the number of samples available from a tumour, are both factors that have a strong influence on the accuracy of methods for analysing intratumour heterogeneity (Bhandari *et al.*, 2018; Siegmund and Shibata, 2016; Watkins and Schwarz, 2018; Mahlokozera *et al.*, 2018).

*freqcalc* takes the VCF and copy number outputs for each clone from *heterogenesis\_varincorp*, along with a file specifying proportions of each somatic clone and the germline in a sample. It then calculates and outputs equivalent information for a bulk sample that contains the given clone proportions.

#### 2.2.1.3 How HeteroGenesis models scenarios to recreate tumour complexity.

- I. **Multi-level subclone phylogenies:** With HeteroGenesis, the user has full control over the subclonal architecture of tumours by defining two parameters per clone: parent clone and the evolutionary distance from it. Varied and complex evolutionary trajectories can therefore be modelled.
- II. **Individual chromosome and whole-genome aneuploidy:** HeteroGenesis simulates a user defined number of aneuploid events, with a user defined probability that each will be for a single chromosome or the whole genome. New copies of chromosomes inherit the existing variants on the parent chromosome, and then acquire further variants unique to each copy.
- III. **Overlapping copy number variants (CNVs):** Overlapping CNVs are made possible in HeteroGenesis by splitting the genome at every breakpoint into blocks that can be sequentially replicated or removed with each CNV.
- IV. **Mutations occurring in a flexible order:** In order to accommodate flexible orders of mutations by HeteroGenesis, the number of occurrences of each SNV and InDel in a genome are calculated from the CNVs and aneuploid events that occur subsequently over each variant. This requires keeping

track of CNV break points to determine whether a new CNV falls within or around an existing CNV, and therefore how many existing copies should be multiplied by the new CNV copy number.

- V. **Distinct germline and somatic variants:** With HeteroGenesis, the user has the option to take a proportion of the germline SNVs and InDels from known variants in dbSNP and, unlike any previous method, weights them by their frequency in the population.

#### 2.2.1.4 Modifications since publication

Following a request from other researchers interested in using HeteroGenesis, in a more recent version of HeteroGenesis since the publication describing its use (Tanner *et al.*, 2019), I updated the code to allow users to provide lists of given mutations for *heterogenesis\_vargen* to sample from. This allows a specified proportion of each type of mutation to be incorporate into the simulation, meaning that, for example, a proportion of CNVs or SNVs/InDels can be taken from the COSMIC database of known cancer mutations (Tate *et al.*, 2019).

#### 2.2.2 Optimisation of *In silico* whole-exome sequencing

##### 2.2.2.1 Wessim

In order to create WES datasets for use in benchmarking, reads need to be generated from the simulated genomes through *in silico* sequencing. This is a process that aims to simulate real DNA sequencing on a computer by sampling sections of a genome sequence and incorporating errors at rates estimated to occur from real DNA sequencers.

Many tools have been developed to simulate WGS (Zhao *et al.*, 2016), and while it is possible to simply perform *in silico* WGS and limit it to exome regions with a BED file, this requires tracking alterations to exon positions from CNVs and indels, which can be especially challenging when CNVs overlap, and it also does not realistically model the exon capture process. Alternatively, Wessim (Kim *et al.*, 2013), has been specifically developed for WES. This method was developed from GemSim (McElroy *et al.*, 2012), an *in silico* WGS simulator that creates a sequencing error model of insertions, deletions and transitions through training on real data. Wessim built on this by creating reads in user defined regions for WES. These regions can be defined in one of two ways; i) a BED file containing exome capture target regions (ideal target mode), as described above, or ii) a Blat (Kent, 2002) alignment of exome capture probes to a genome (probe hybridization mode), giving a more realistic distribution of reads along exome regions compared to the first method. The ideal target mode approach would not work with a genome where exon regions have moved positions, as is the case in one with CNA. It also wouldn't be able to model higher read numbers in amplified regions. Therefore, I chose to explore the probe hybridisation mode.

### 2.2.2.2 GemSIM error model

#### 2.2.2.2.1 Error model creation

In order to reliably assess the performance of variant calling, *in silico* sequencing reads need to have realistic base quality scores for both true and false base calls. This requires an accurate error model of sequencing performance in order to guide error incorporation.

Wessim uses error models created by GemSIM. The pre-made models available with the GemSIM package were for older sequencing machines, so I created a new model of a more up to date sequencing machine. For this, I trained GemSIM on publicly available Illumina HiSeq 2000, 101bp, paired-end, WGS data of NA12877 (SRA accession no. ERR194146), an individual in the Illumina Platinum Genomes project (Eberle *et al.*, 2017), which I had cleaned with Cutadapt (Martin, 2011), aligned to the hg19 reference genome with BWA-MEM (Li, 2013), realigned with GATK v3.8 (Van der Auwera *et al.*, 2013), downsampled, and limited to chromosome 1 reads only. A list of genomic sites to exclude in the model was taken from high confidence variant calls for NA12877, also provided by the Illumina Platinum Genomes project (Eberle *et al.*, 2017), that had been determined through a combination of methods including taking into account a 17 member pedigree.

I edited the code of GemSIM to accept a file for excluded sites, instead of comma separated list in the command line, and also to use a python dictionary instead of a list to store them, in order to reduce run time. In hindsight, these modifications may have been required because GemSIM was likely designed to be trained on sequencing of small genomes, such as viral ones, which have fewer polymorphic sites. However, as genomes from different organisms have unique characteristics, for example GC content, it was important to model the specific sequencing error profiles of human datasets.

#### 2.2.2.2.2 Error model validation

The resulting error model had substitution, insertion and deletion rates of  $4.1 \times 10^{-3}$ ,  $6.9 \times 10^{-5}$  and  $8.1 \times 10^{-5}$  errors/base, respectively. While previous reports on the error rates of Illumina HiSeq 2000, and other Illumina methods, vary substantially (Laehnemann *et al.*, 2016; Li *et al.*, 2009; Aguirre de Cárcer *et al.*, 2014; Pfeiffer *et al.*, 2018; Schirmer *et al.*, 2015), the error model I created with GemSIM fell within this range. I further investigated the validity of the model by analysing the base quality score distributions for true and false called bases from Wessim. To determine what a realistic distribution would be for each, I used three sources of evidence: 1) Specifications of HiSeq2500 machines state that >80% of bases should have quality scores of above 30 for paired-end 100bp datasets. 2) When looking at the quality scores of bases in overlap regions of paired reads that differ in the base call, and assuming the base that matches the reference genome is a true call and the base that disagrees is a sequencing error (and not a polymerase chain

reaction (PCR) error), I was able to plot distributions for true, false and total quality score distributions (Figure 11). 3) The sequencing dataset that I trained the GemSIM error model on has pedigree information on SNP positions. By ignoring these positions I could assume that any other base that disagrees with the reference genome is likely a false call and any that agrees is a true call (Figure 12). These three sources were in agreement with each other.

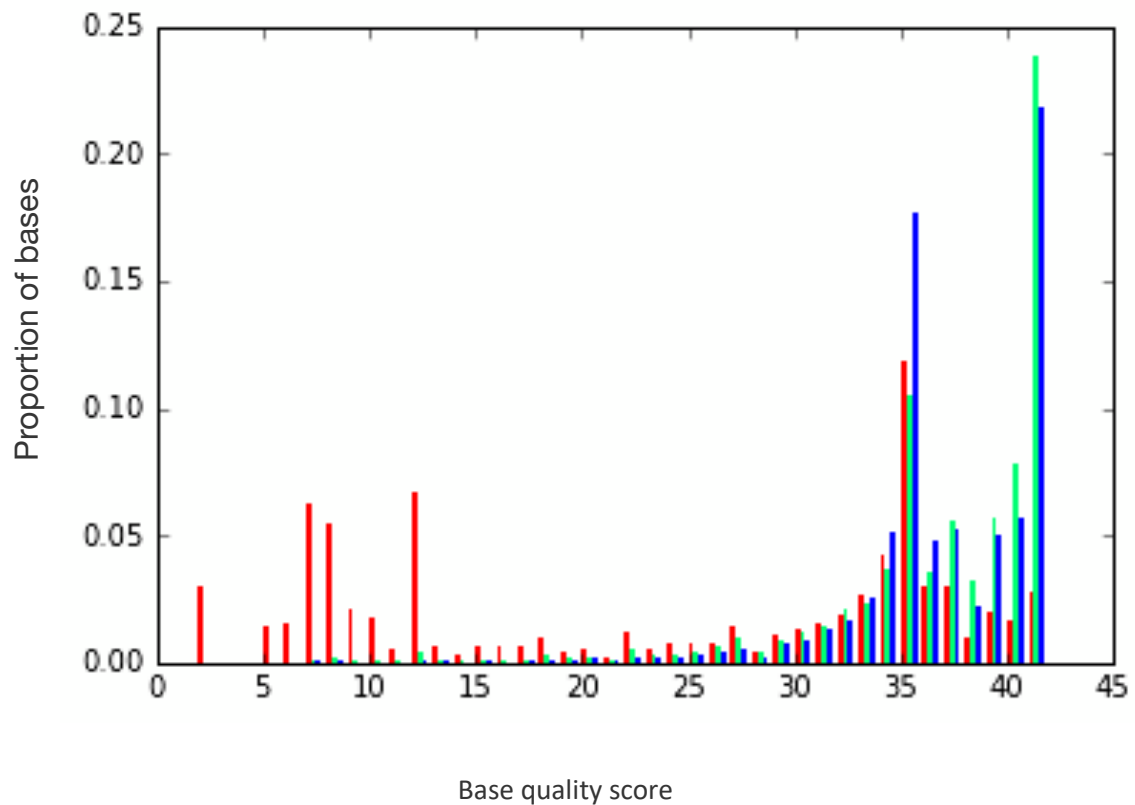


Figure 11. Quality score distributions of bases in overlap regions of paired reads. Red) bases that differed between forward and reverse reads, and which had the lower qscore of the pair, ie. likely to be errors. Green) the higher of the pair. Blue) all bases in overlap regions.

The quality scores for Wessim data followed a very similar distribution to the real data for true calls. However, whilst a similar pattern to real data is also observed for false calls, there is a higher proportion of quality scores above 30 for false calls in Wessim data; ~20% > 30 in real data, ~60% > 30 in Wessim (Figure 12). This indicates that the error model does not fully represent the base quality score distributions of real data, however the Wessim data still shows the same overall patterns, particularly for true bases.

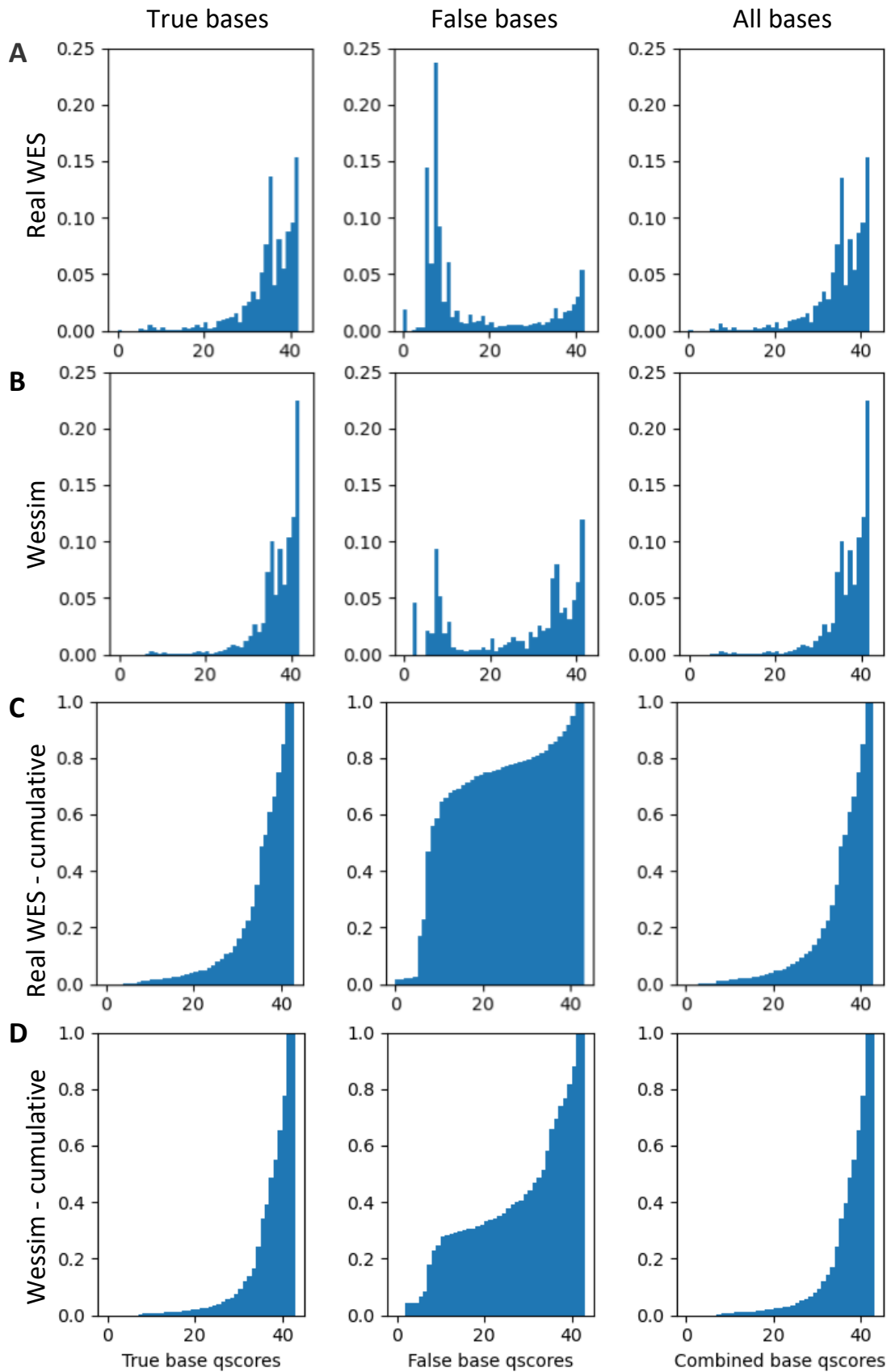


Figure 12. Distributions of base quality scores for sequencing reads from A+C) real WES data, and B+D) w-Wessim simulated data. The datasets were filtered for a minimum read mapping quality of 20. C+D show the distributions with a cumulative scale.

### 2.2.2.3 Wessim in probe hybridisation mode

In probe hybridisation mode, Wessim aims to mimic the exome capture step during sequencing library preparation. This is achieved through the use of Blat (Kent, 2002) alignments of hybridisation probes to a genome in order to define regions for sequencing. This means that sequenced regions will be in phase with target regions, regardless of any changes in length that occur in the genome from variants.

To test Wessim, I first created a Blat alignment of the probe sequences for the Agilent SureSelect Human All Exon V4+UTRs kit, the most recent Agilent kit for which probe sequences have been released, to a modified version of the hg38 reference genome which I had inserted a CNV with a copy number of 3 at position chr22:24,917,701-24,926,065. I then ran Wessim with this Blat alignment and cleaned the resulting reads, followed by aligning them to the hg38 genome with BWA-MEM (Li, 2013). However, when visualising the aligned reads in IGV (Thorvaldsdóttir *et al.*, 2013) I discovered that Wessim was not able to model the effect of the CNV I had inserted into the reference genome (Figure 13B and F). This is because the program selects probes at random for *in silico* hybridisation each time it creates a read, regardless of how many times they're listed to align to the genome by Blat. Therefore, the same number of reads is generated for a region regardless of copy number (other than for regions where the copy number is 0). In addition, the read distribution was visibly different to real sequencing reads, particularly with regards to off target reads (Figure 13B and F).

### 2.2.2.4 w-Wessim

To address the above limitations, I modified Wessim to create weighted-Wessim (w-Wessim) and combined it with an altered protocol. Together this allows 4 advantages over the original Wessim method:

- i. **CNVs Modeled.** In order to overcome the issue of Wessim not modeling CNVs in genomes, I modified the program's code to weight probe selection by the number of times each probe aligns to a genome. This resulted in read numbers across bases reflecting the copy number of the region (Figure 13C and G).
- ii. **Realistic read distributions.** The standard protocol for Wessim results in unrealistic read coverages, both across target regions and also due to a lack of off-target reads. In real sequencing data a large proportion of reads align to off-target regions. The Agilent SureSelect Human All Exon V5+UTRs kit is estimated by the manufactures to capture reads with only approximately 65% aligning to target regions and 77% aligning  $\pm$  100bp from targets. Additionally, I found that three WES datasets from the NCBI Sequence Read Archive, created independently with this kit and sequenced by an Illumina HiSeq 2500, had 56.5%-61.0% of bases aligning on, and 67.9%-77.3% bases aligning on or near target regions (Figure 14). Furthermore, background levels of off target reads are seen between the

larger (and generally on target) peaks when visualising alignments of these real reads (Figure 13A and E). However, when using a Blat alignment of probes from the Agilent SureSelect Human All Exon V4+UTRs kit (the most recent kit for which the sequences have been made available), w-Wessim and Wessim resulted in very high proportions of bases aligning near to, or on, target regions; 90.6% and 90.0% on target and 99.6% and 98.1% on or near target for w-Wessim and Wessim respectively. In addition, the mode coverages for the three real WES datasets, when subsampled down to 70 million reads, were 8x, 28x and 29x, whereas the mode coverages for the same number of reads generated by w-Wessim and Wessim, was 66x and 80x, respectively (Figure 14).

I initially attempted to overcome the lack of off target reads by altering the Blat parameters for much less stringent alignment so that regions of the genome that have a less perfect match to the probes would still be captured, allowing a more relaxed distribution of reads. However, this resulted in an even less realistic distribution with very high coverage across parts of some exomes.

I attempted another approach by utilising real WES reads to guide a more realistic distribution. For this, a Blat alignment of real WES reads from the Sequence Read Archive (SRR2103613 - frozen normal adult male lung, 99bp paired-end, 61.0% and 75.8% of bases on and on or near target respectively) was used instead of capture kit probe sequences. This dataset had the median percentage of on target reads of the three datasets I identified that were created with the Agilent V5+UTRs kit and Illumina HiSeq 2500.  $\sim 1 \times 10^8$  real reads were used, which allows for sufficient variation in regions for read generation by w-Wessim for at least 250x coverage, whilst still being practically possible computationally. Initially, too many off target reads were created by w-Wessim using this method, so I filtered the real reads used in the Blat alignment for those with a mapping score of  $\geq 60$  when aligned by BWA-MEM(Li, 2013) to the hg38 reference genome. I then further optimised the approach by sequentially testing a range of values for Blat parameters to increase alignment stringencies. I selected minidentity=95 and minscore=95 as this combination allowed for the most realistic distribution of reads created by w-Wessim; 61.1% on target and 79.7% on or near target, and with a mode coverage of 28x for 70 million reads (Figure 14). These settings are also likely to reflect the stringency of real exon capture hybridisation in the lab, thereby resulting in more realistic modelling of the affect that variants have on exon capture.

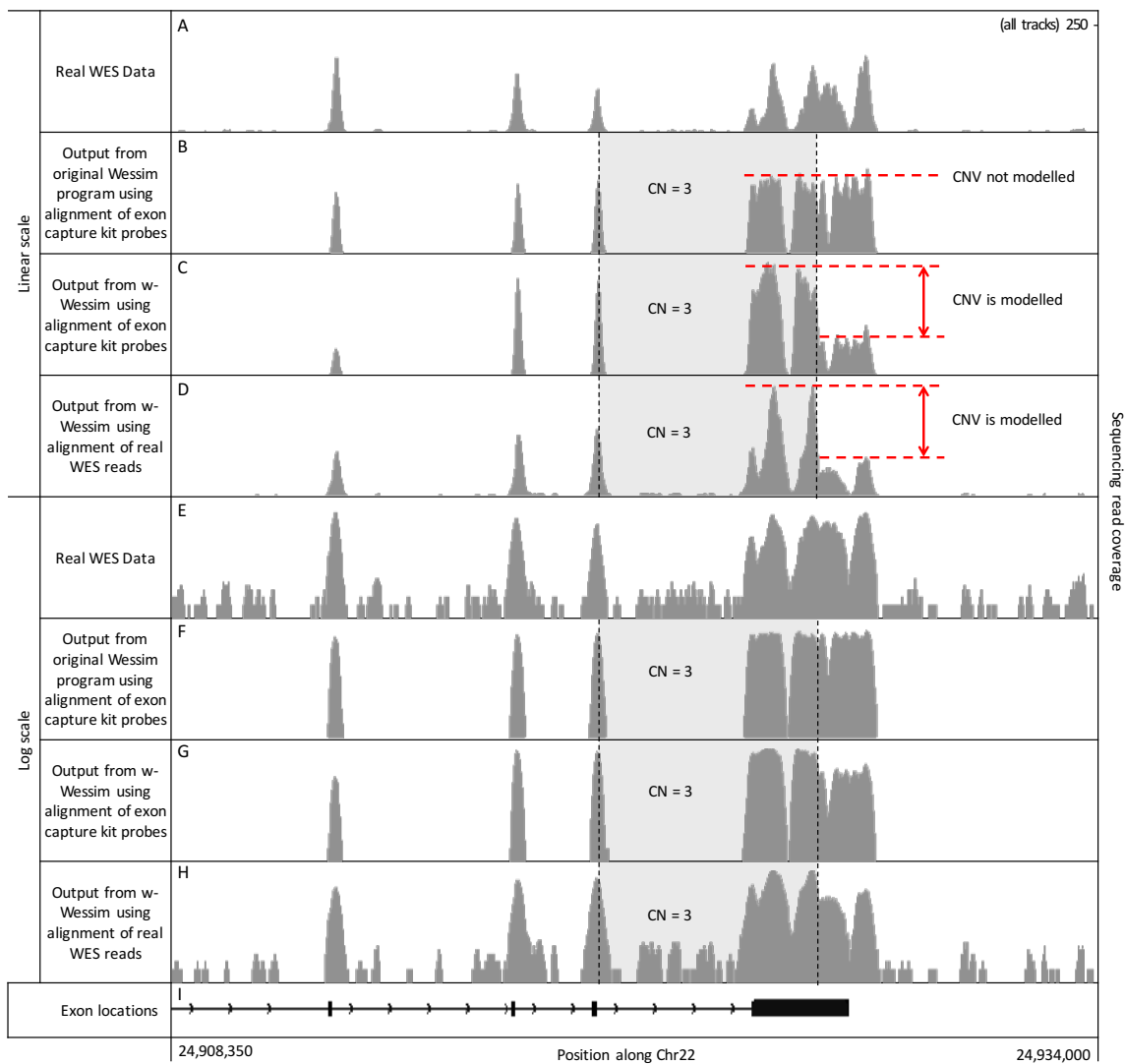


Figure 13. Distributions of real and simulated WES reads along a region of chromosome 22, with linear scales of coverage depth (A to D, enabling copy numbers to become apparent) and log scales (E to H, enabling off target coverage to become apparent). Simulated data was generated from the hg38 human reference genome that had a CNV with a copy number of 3 inserted at position chr22:24,917,701-24,926,065. A+E) Real reads from the SRR2103613 data set. B+F) Reads generated by the original Wessim program using the recommended protocol with a BLAT alignment of Agilent SureSelect Human All Exon V4+UTRs kit probe sequences. C+G) Reads generated by w-Wessim using the original Wessim recommended protocol with a BLAT alignment of Agilent SureSelect Human All Exon V4+UTRs kit probe sequences. D+H) Reads generated by w-Wessim using our modified protocol with a BLAT alignment of real reads from the SRR2103613 data set. I) Position of exon and intron locations, shown as boxes and lines respectively.



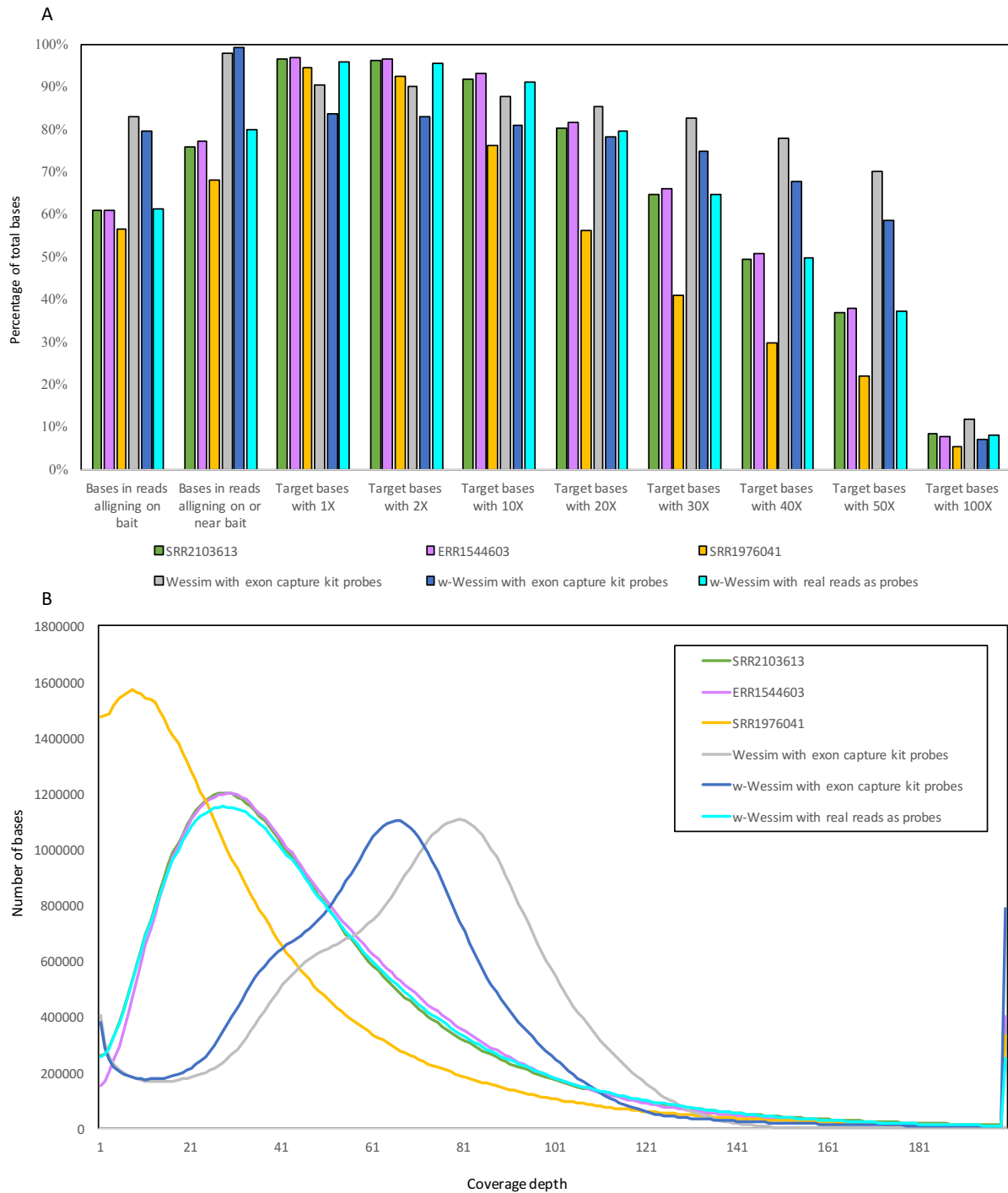


Figure 14. Coverage metrics for three publically available real WES datasets (named by their SRA accession number) created with the Agilent SureSelect Human All Exon V5+UTRs kit, and three simulation methods that use either the Agilent SureSelect Human All Exon V4 kit probe sequences or real reads from the SRR2103613 dataset as probes. All datasets contained 70 million reads. (WES metrics were calculated using Picard HsMetrics (Broad Institute) with the 'Covered.bed' target files downloaded from the Agilent website (<https://earray.chem.agilent.com/sureselect/index.htm>). The percentages of bases aligning to target regions was calculated by dividing the number of aligned bases (with a

mapping quality > 0) in bait regions, by the total number of aligned bases (also with a mapping quality > 0). “Bait”, not “target”, values from HsMetrics were used for on/off target calculations as these do not exclude low quality reads. “Target” values were used for coverage metrics as “bait” values were not available.)

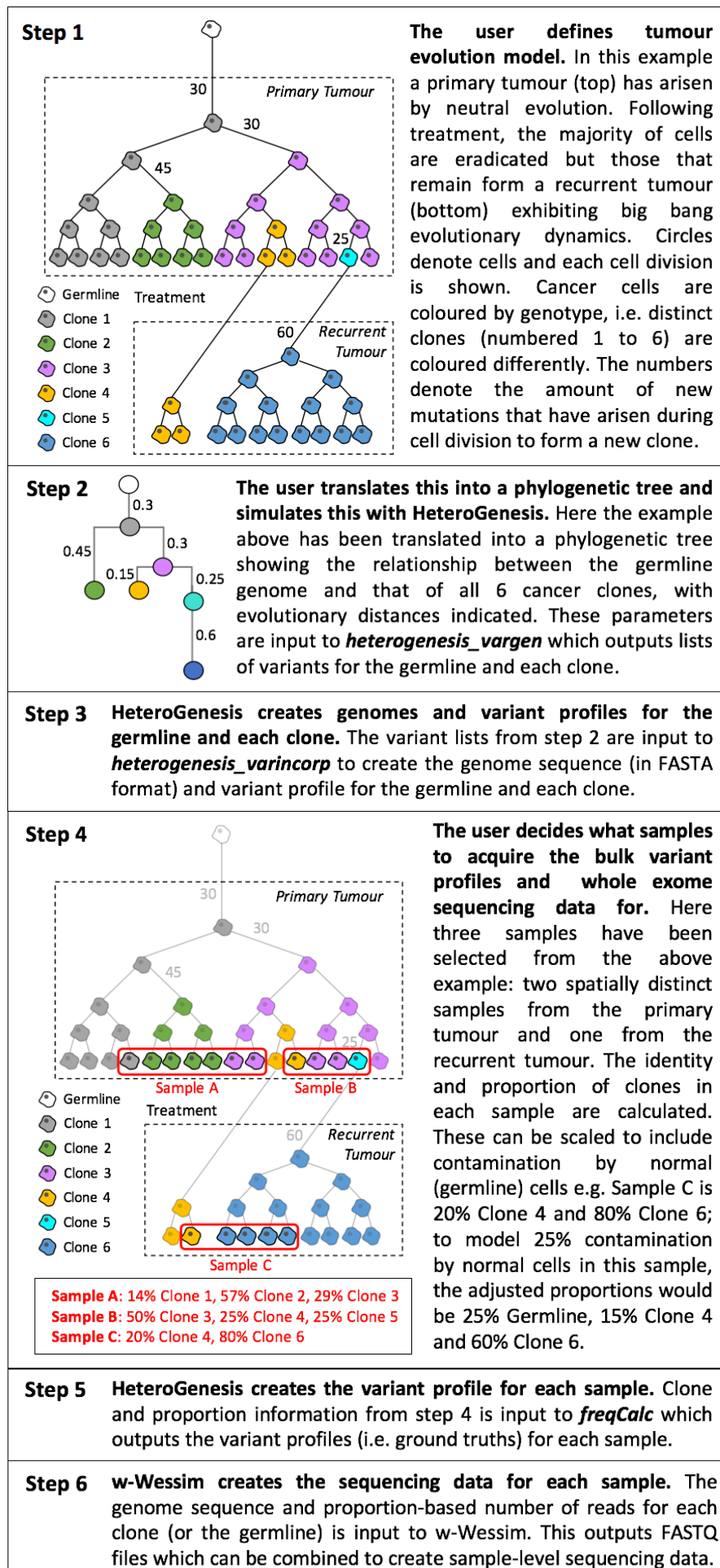
- iii. **Realistic read length distribution.** Wessim allows the user to choose a fixed length of reads to be created. However, when processing real sequencing data it’s common to trim a variable number of bases from the ends of reads to remove sequencing errors, which occur at higher frequencies at ends of reads, and also adapter sequences. This means the reads used for further analysis are of varying lengths, which could affect alignment performance. I therefore wanted to create artificial reads that have the same length distributions as trimmed datasets. Whilst it may have been feasible to create full length reads and then trim them in the same way as real data, this would have involved training an error model on aligned full length reads, which would have been more likely to be misaligned and therefore result in an overestimation of sequencing errors in the model created. Instead I allowed w-Wessim to take read lengths from the distribution in the GemSIM error model, thereby simulating pre-trimmed reads. This was achieved using code taken from GemSIM.
  
- iv. **Generation of sequencing fragments with a length distribution that can fall below read lengths.** I wanted to use w-Wessim to model formaldehyde-fixed, paraffin embedded samples, and these generally contain short fragment lengths. These fragment lengths can sometimes fall below the read lengths in the GemSIM error model distribution. Whilst Wessim generally handles such situations by limiting the fragment length, I found that it fails specifically when a fragment length is shorter than the read length and a deletion occurs at the last position of the read. I adjusted the code in w-Wessim to handle such scenarios.

With the full set of  $1 \times 10^8$  real read probes used in the example, this would take ~700h/threads for the BLAT alignment (which can be ran across multiple nodes in separate runs by splitting the read number, probes or genome sequence) and require around 9h and 72GB RAM to generate  $1 \times 10^7$  pairs of reads on a single thread by w-Wessim.

### 2.2.3 Demonstration of the combined use of HeteroGenesis with w-Wessim

HeteroGenesis and w-Wessim can be used in combination to create multi-sample bulk WES data for an artificial tumour with multiple subclones. Figure 15 provides an overview of the steps involved in achieving this.

Figure 15. A stepwise example showing how HeteroGenesis and w-Wessim can be used to simulate spatiotemporal sampling from heterogeneous tumours.



To demonstrate the use of both HeteroGenesis and w-Wessim, I used HeteroGenesis to simulate an example tumour with 8 clones and then sequenced it with w-Wessim, creating  $1 \times 10^8$  read pairs per clone. I aligned these to the hg38 reference genome with BWA-MEM (Li, 2013) and visualised the alignments in IGV (Thorvaldsdóttir *et al.*, 2013).

I first showed that the CNV modelling was appropriate and reflected what HeteroGenesis indicated was incorporated into the tumour genomes. For example, Figure 16 lists three CNV events that are recorded by HeteroGenesis to have occurred within the region chr3:183260000-184110000. The read counts in six regions, separated by CNV break points (as stated by HeteroGenesis), is given for both the germline and clone1 of the tumour. These counts, when normalised by the stated copy number for each region, were very similar between the germline and clone1 in all six regions, thereby indicating that the resulting read counts appropriately reflect the CNV events stated to have occurred by HeteroGenesis.

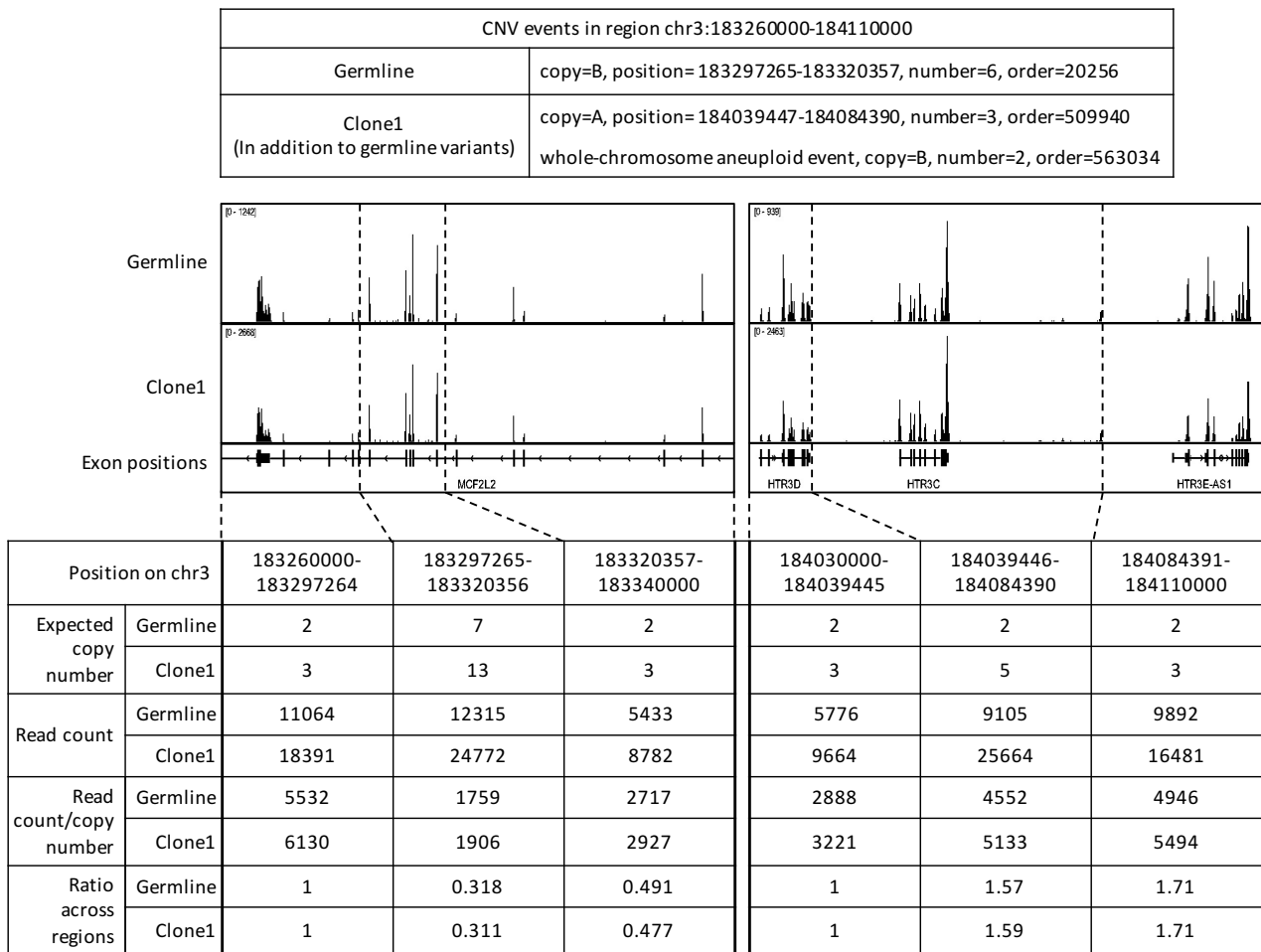


Figure 16. Sequencing datasets for the germline and clone1 of a tumour simulated by HeteroGenesis and *in silico* sequenced by w-Wessim, viewed on IGV. Read counts for each region are divided by the expected copy number and then divided by the number of reads in the first region to get the ratio of reads between regions. Equal ratios across regions, between the germline and clone1 samples indicate appropriate modelling of copy numbers.

I next demonstrated that point mutation locations and VAFs reflected what was stated to be incorporated by HeteroGenesis. Figure 17 lists all point mutations and CNVs that occurred within the region ch3:183260000-184110000. For each of the four point mutations, the expected and observed VAFs are given and found to be similar, thereby indicating appropriate modelling of both point mutations and CNVs.

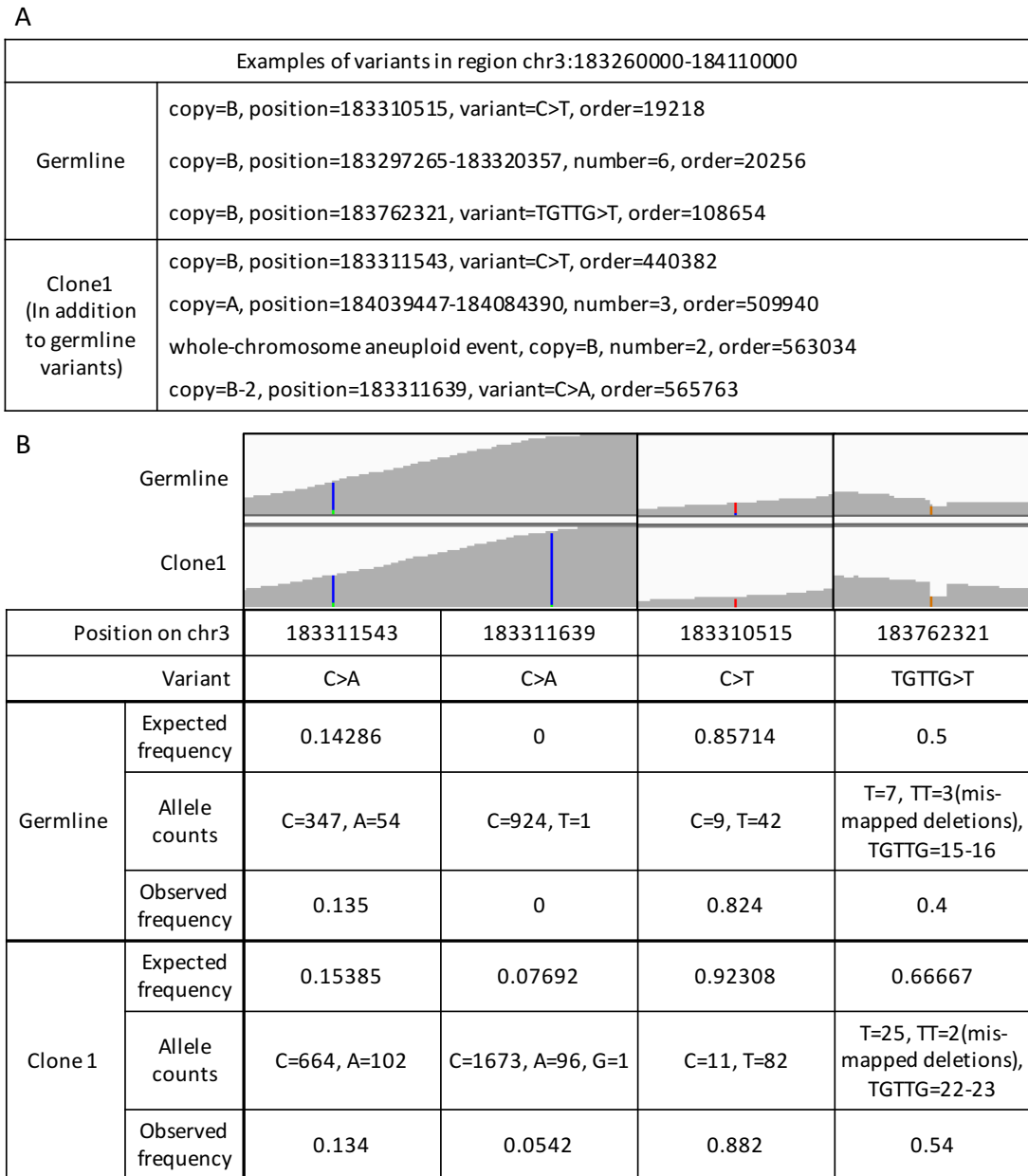


Figure 17. Variants in sequencing datasets for the germline and clone1 of a tumour simulated by HeteroGenesis and *in silico* sequenced by w-Wessim. A) Examples of variants and CNVs occurring in the ch3:183260000-184110000 region, listed in the order they occur. B) Details of observed and expected allele counts in the sequencing reads for each variant. Similar observed and expected frequencies indicate appropriate modelling. The read alignment panel is created from viewing reads in IGV.

## 2.3 Discussion

In this study, I aimed to develop methods that would allow me to simulate realistically complex WES datasets for heterogeneous tumours that are suitable for benchmarking mutation calling and subclonal deconvolution methods. I achieved this by creating two programs, HeteroGenesis and w-Wessim.

HeteroGenesis simulates genome sequences for each clone in a tumour and calculates mutation profiles for individual clones as well as user defined bulk samples. SNVs, InDels, CNVs, and aneuploid events are incorporated with few restrictions into clones with fully customisable phylogenetic relationships, leading to highly complex tumours. This overcomes numerous limitations of previous somatic genome simulation methods that makes them unsuitable for benchmarking methods for analysing sequencing data from heterogeneous tumours.

HeteroGenesis does have some minor limitations; Firstly, it cannot model partially overlapping CNVs on the same chromosome (but does model fully overlapping on the same copy, or partially overlapping in a different copy). Including this feature would increase the complexity of the program substantially, but with little to no gain for its application in benchmarking subclonal deconvolution pipelines. Secondly, structural variants that don't alter copy numbers, such as translocations, are not modelled, as these are not relevant to subclonal deconvolution. Still, the program can be used for testing methods for detecting break points since break points occur at CNV locations. In cases where researchers wish to benchmark methods for detecting translocations, other simulators that focus on these variants are available (Qin *et al.*, 2015; Xia *et al.*, 2018, 2017).

I created HeteroGenesis with a top down approach, where the resulting tumour is fully characterised by input parameters. An alternative approach, which HeteroGenesis could be adapted to achieve, is to allow the tumour to randomly evolve overtime in a way that takes into account the effects of each simulated mutation. For example, if a mutation affects a DNA repair gene, the subsequent mutation rate in that clone, and all daughter clones, could be increased. Such a model may allow insights into the mathematics of factors that influence tumour progression.

w-Wessim features important improvements to the only existing *in silico* sequencing method specifically for WES. The modifications, combined with an altered protocol that uses real WES reads, allow w-Wessim to model the effect of CNVs and to create more realistic WES data than the original Wessim method. A shortcoming of w-Wessim is that the reads I generated from it using the GemSIM error model, had base quality scores that, for error positions, were higher on average than in real WES data. Additionally, w-Wessim is unable to model some sources of strand-bias, such as those arising from PCR errors during library preparation (although other sources such as alignment and post-alignment processing will still be included in the resulting datasets) (Guo *et al.*, 2012). These limitations could be overcome in the future by

combining the WES specific features of w-Wessim with an alternative and up to date *in silico* WGS method, such as InSilicoSeq (Gourlé *et al.*, 2019) or ReSeq (Schmeing and Robinson, 2020), and making use of their error modelling ability. Alternatively, w-Wessim could be modified to override the error model distribution for error bases, and instead force a given distribution, parametrised to bring it in-line with that of real data. I also plan to investigate whether downsampling the real WES reads, used as probes for w-Wessim, significantly impacts the distribution of simulated reads, in order to find a way to reduce the computational resources required.

Nonetheless, the current w-Wessim approach still allows for datasets that are suitable for achieving my aims. The differences in quality score distributions should have no effect on CNV calling, particularly as the alignment step is not dependant on base quality scores. True positive calls in SNV calling are also unlikely to be affected, as the quality score distribution for true sequenced bases was highly realistic. Furthermore, assessment of certain features of SNV calling, such as VAF estimation and sensitivity at lower frequencies, which are relevant to subclonal deconvolution, is unlikely to be affected. Potential effects on numbers of false positive calls will be investigated in the benchmarking.

By combining HeteroGensis with w-Wessim, I have shown how realistically complex WES datasets can be created. Whilst the primary aim of developing these programs was to allow me to create datasets that were suitable for benchmarking mutation calling and subclonal deconvolution methods, they will also likely be of use to other researchers for developing new analysis methods and carrying out future benchmarking studies.

## 2.4 References

- 1000 Genomes Project Consortium, T. 1000 G.P. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
- Abécassis, J. *et al.* (2019) Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data. *PLoS One*, **14**.
- Aguirre de Cárcer, D. *et al.* (2014) Evaluation of viral genome assembly and diversity estimation in deep metagenomes. *BMC Genomics*, **15**, 989.
- Alioto, T.S. *et al.* (2015) A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.*, **6**, 10001.
- Andor, N. *et al.* (2016) Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.*, **22**, 105–13.
- Van der Auwera, G.A. *et al.* (2013) From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.*
- Baysan, M. *et al.* (2017) Detailed longitudinal sampling of glioma stem cells *in situ* reveals Chr7 gain and Chr10 loss as repeated events in primary tumor formation and recurrence. *Int. J. Cancer*, **141**, 2002–2013.
- Ben-David, U. and Amon, A. (2020) Context is everything: aneuploidy in cancer. *Nat. Rev. Genet.*, **21**, 44–62.
- Benjamin, D. *et al.* (2019) Calling Somatic SNVs and Indels with Mutect2.
- Bhandari, V. *et al.* (2018) Quantifying the Influence of Mutation Detection on Tumour Subclonal Reconstruction. *bioRxiv*, 418780.
- Bian, X. *et al.* (2018) Comparing the performance of selected variant callers using synthetic data and

- genome segmentation. *BMC Bioinformatics*, **19**, 1–11.
- Bohnert,R. *et al.* (2017) Comprehensive benchmarking of SNV callers for highly admixed tumor data. *PLoS One*, **12**, e0186175.
- Broad Institute Picard Tools - By Broad Institute.
- Chen,Z. *et al.* (2020) Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci. Rep.*, **10**, 1–9.
- Christoforides,A. *et al.* (2013) Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC Genomics*, **14**, 302.
- Cibulskis,K. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.
- Cun,Y. *et al.* (2018) Copy-number analysis and inference of subclonal populations in cancer genomes using ScIust. *Nat. Protoc.*, **13**, 1488–1501.
- Dentro,S.C. *et al.* (2020) Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *bioRxiv*, 312041.
- Deshwar,A.G. *et al.* (2015) PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.*, **16**.
- Detering,H. *et al.* (2019) Accuracy of somatic variant detection in multiregional tumor sequencing data. *bioRxiv*, 655605.
- Droop,A. *et al.* (2018) How to analyse the spatiotemporal tumour samples needed to investigate cancer evolution: A case study using paired primary and recurrent glioblastoma. *Int. J. Cancer*, **142**, 1620–1626.
- Durbin,R.M. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
- Eberle,M.A. *et al.* (2017) A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.*, **27**, 157–164.
- El-Kebir,M. *et al.* (2016) Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. *Cell Syst.*, **3**, 43–53.
- Ewing,A.D. *et al.* (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods*, **12**, 623–630.
- Fan,Y. *et al.* (2016) MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.*, **17**, 178.
- Favero,F. *et al.* (2015) Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, **26**, 64–70.
- Gerlinger,M. *et al.* (2014) Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.*, **46**, 225–233.
- Gourlé,H. *et al.* (2019) Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*, **35**, 521–522.
- Guo,Y. *et al.* (2012) The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics*, **13**, 666.
- Hansen,N.F. *et al.* (2013) Shimmer: detection of genetic alterations in tumors using next-generation sequence data. *Bioinformatics*, **29**, 1498–503.
- Hosny,A. (2017) NabaviLab/VarSimLab.
- Hu,Y. *et al.* (2013) Tumor-Specific Chromosome Mis-Segregation Controls Cancer Plasticity by Maintaining Tumor Heterogeneity. *PLoS One*, **8**, e80898.
- ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**, 82–93.
- Ivakhno,S. *et al.* (2017) tHapMix: simulating tumour samples through haplotype mixtures. *Bioinformatics*, **33**, 280–282.
- Jiang,Y. *et al.* (2016) Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl. Acad. Sci.*, **113**, E5528–E5537.
- Kandoth,C. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333.
- Kent,W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res.*, **12**, 656–64.
- Kim,S. *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.



- Kim,S. *et al.* (2013) Wessim: a whole-exome sequencing simulator based on in silico exome capture. *Bioinformatics*, **29**, 1076–1077.
- Koboldt,D.C. *et al.* (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
- Körber,V. *et al.* (2019) Evolutionary Trajectories of IDH WT Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis. *Cancer Cell*, **35**, 692–704.e12.
- Krijgsman,O. *et al.* (2014) Focal chromosomal copy number aberrations in cancer—Needles in a genome haystack. *Biochim. Biophys. Acta - Mol. Cell Res.*, **1843**, 2698–2704.
- Krøigård,A.B. *et al.* (2016) Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLoS One*, **11**, e0151664.
- Laehnemann,D. *et al.* (2016) Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Brief. Bioinform.*, **17**, 154–179.
- Lähnemann,D. *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome Biol.*, **21**, 1–35.
- Lai,Z. *et al.* (2016) VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.*, **44**, e108.
- Li,B. and Li,J.Z. (2014) A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol.*, **15**, 473.
- Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*.
- Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–9.
- Mahlokozera,T. *et al.* (2018) Biological and therapeutic implications of multisector sequencing in newly diagnosed glioblastoma. *Neuro. Oncol.*, **20**, 472–483.
- Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.
- McElroy,K.E. *et al.* (2012) GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*, **13**, 74.
- McPherson,A. *et al.* (2016) Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.*, **48**, 758–767.
- Mills,R.E. *et al.* (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.*, **16**, 1182–90.
- Miura,S. *et al.* (2018) Predicting clone genotypes from tumor bulk sequencing of multiple samples. *Bioinformatics*, **34**, 4017–4026.
- Mu,J.C. *et al.* (2015) VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics*, **31**, 1469–1471.
- Mullaney,J.M. *et al.* (2010) Small insertions and deletions (INDELS) in human genomes. *Hum. Mol. Genet.*, **19**, R131-6.
- Nam,J.-Y. *et al.* (2016) Evaluation of somatic copy number estimation tools for whole-exome sequencing data. *Brief. Bioinform.*, **17**, 185–192.
- Narzisi,G. *et al.* (2018) Genome-wide somatic variant calling using localized colored de Bruijn graphs. *Commun. Biol.*, **1**, 1–9.
- Nik-Zainal,S. *et al.* (2012) The life history of 21 breast cancers. *Cell*, **149**, 994–1007.
- Noorbakhsh,J. *et al.* (2018) Distribution-based measures of tumor heterogeneity are sensitive to mutation calling and lack strong clinical predictive power. *Sci. Rep.*, **8**.
- Pan,B. *et al.* (2019) Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics*, **20**, 101.
- Pfeiffer,F. *et al.* (2018) Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.*, **8**, 10950.
- Pinto,D. *et al.* (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.*, **29**, 512–520.
- Pitea,A. *et al.* (2018) Copy number aberrations from Affymetrix SNP 6.0 genotyping data-how accurate are commonly used prediction approaches? *Brief. Bioinform.*
- Qin,M. *et al.* (2015) SCNVSim: somatic copy number variation and structure variation simulator. *BMC Bioinformatics*, **16**, 66.
- Rieber,N. *et al.* (2017) Reliability of algorithmic somatic copy number alteration detection from targeted capture data. *Bioinformatics*, **33**, 2791–2798.

- Roberts,N.D. *et al.* (2013) A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics*, **29**, 2223–2230.
- Roth,A. *et al.* (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396–8.
- Salcedo,A. *et al.* (2020) A community effort to create standards for evaluating tumor subclonal reconstruction. *Nat. Biotechnol.*, **38**, 97–107.
- Schirmer,M. *et al.* (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.*, **43**, e37–e37.
- Schmeing,S. and Robinson,M. (2020) ReSeq simulates realistic Illumina high-throughput sequencing data. *bioRxiv*, 2020.07.17.209072.
- Semeraro,R. *et al.* (2018) Xome-Blender: A novel cancer genome simulator. *PLoS One*, **13**, e0194472.
- Shen,H. *et al.* (2013) Comprehensive Characterization of Human Genome Variation by High Coverage Whole-Genome Sequencing of Forty Four Caucasians. *PLoS One*, **8**, e59494.
- Shen,R. and Seshan,V.E. (2016) FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.*, **44**, e131–e131.
- Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–11.
- Siegmund,K. and Shibata,D. (2016) At least two well-spaced samples are needed to genotype a solid tumor. *BMC Cancer*, **16**, 250.
- Sottoriva,A. *et al.* (2013) Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 4009–4014.
- Spinella,J.-F. *et al.* (2016) SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC Genomics*, **17**, 912.
- Stead,L.F. *et al.* (2013) Accurately Identifying Low-Allelic Fraction Variants in Single Samples with Next-Generation Sequencing: Applications in Tumor Subclone Resolution. *Hum. Mutat.*, **34**, 1432–1438.
- Tan,R. *et al.* (2014) An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. *Hum. Mutat.*, **35**, 899–907.
- Tanner,G. *et al.* (2019) Simulation of heterogeneous tumour genomes with HeteroGenesis and *in silico* whole exome sequencing. *Bioinformatics*.
- Tate,J.G. *et al.* (2019) COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.*, **47**, D941–D947.
- Telenti,A. *et al.* (2016) Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, 11901–11906.
- Thorvaldsdóttir,H. *et al.* (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
- Watkins,T.B.K. and Schwarz,R.F. (2018) Phylogenetic Quantification of Intratumor Heterogeneity. *Cold Spring Harb. Perspect. Med.*, **8**, a028316.
- Xi,R. *et al.* (2011) Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, E1128–36.
- Xia,L.C. *et al.* (2018) SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. *Gigascience*, **7**.
- Xia,Y. *et al.* (2017) Pysim-sv: a package for simulating structural variation data with GC-biases. *BMC Bioinformatics*, **18**, 53.
- Xu,H. *et al.* (2014) Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics*, **15**, 244.
- Zare,F. *et al.* (2017) An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*, **18**, 286.
- Zhang,L. *et al.* (2019) Comprehensively benchmarking applications for detecting copy number variation. *PLOS Comput. Biol.*, **15**, e1007069.
- Zhao,M. *et al.* (2016) Systematic review of next-generation sequencing simulators: computational tools, features and perspectives. *Brief. Funct. Genomics*, **24**, elw012.

# Chapter 3 – Benchmarking of mutation calling and subclonal deconvolution methods

## 3.1 Introduction

### 3.1.1 Overview

Characterising intratumour heterogeneity (ITH) is important for investigations into tumour evolution and response to treatment. A common approach is to estimate cancer cell fractions (CCFs) of mutations from bulk sequencing data, or to go further and attempt to delineate the subclonal architectures by assigning mutations into distinct clones. However, as discussed in Chapter 2, there has been a lack of reliable studies benchmarking methods for performing such analyses, particularly involving full pipelines. Therefore, in this chapter, I carry out benchmarking of such analyses that use whole exome sequencing (WES) of a single bulk tumour sample with a matched normal, to allow me to identify the most suitable approach for analysing our real in house GBM datasets. This involves testing full pipelines, from *in silico* WES datasets created with w-Wessim and HeteroGenesis, through to subclonal deconvolution methods, and also includes assessment of somatic variant calling and copy number aberration (CNA) calling methods used to provide inputs for CCF estimation.

Genome sequencing is an increasingly common approach for investigating genetic mutations. In particular, whole-genome sequencing (WGS) has the advantage that it can detect variants located anywhere in the genome. Although, due to the high cost involved at greater sequencing depths, whole-exome sequencing (WES) is widely used as an alternative by focusing on the 1-2% of the genome that codes for proteins (Sakharkar *et al.*, 2004). This enables higher sequencing depths to be reached, within a given budget, of the regions usually of most interest to researchers; due to their ability to alter the sequence, and potentially function of proteins, mutations in coding regions are more likely (though not exclusively) to drive selection than non-coding mutations (Rheinbay *et al.*, 2020). This makes WES a popular choice for investigating ITH. Another consideration for researchers and clinicians is how many samples to sequence from a tumour. Whilst a higher number of multi-region samples greatly improves the accuracy of analyses (Bhandari *et al.*, 2018; Siegmund and Shibata, 2016; Watkins and Schwarz, 2018; Sun *et al.*, 2017; Chkhaidze *et al.*, 2019), financial constraints or a limited amount of material mean that often only one sample per tumour is sequenced. Therefore, in this study I focus on subclonal deconvolution pipelines suitable for this scenario, where a single tumour sample, and matched normal, are whole-exome sequenced to typical depths of 60x-250x (Figure 18).

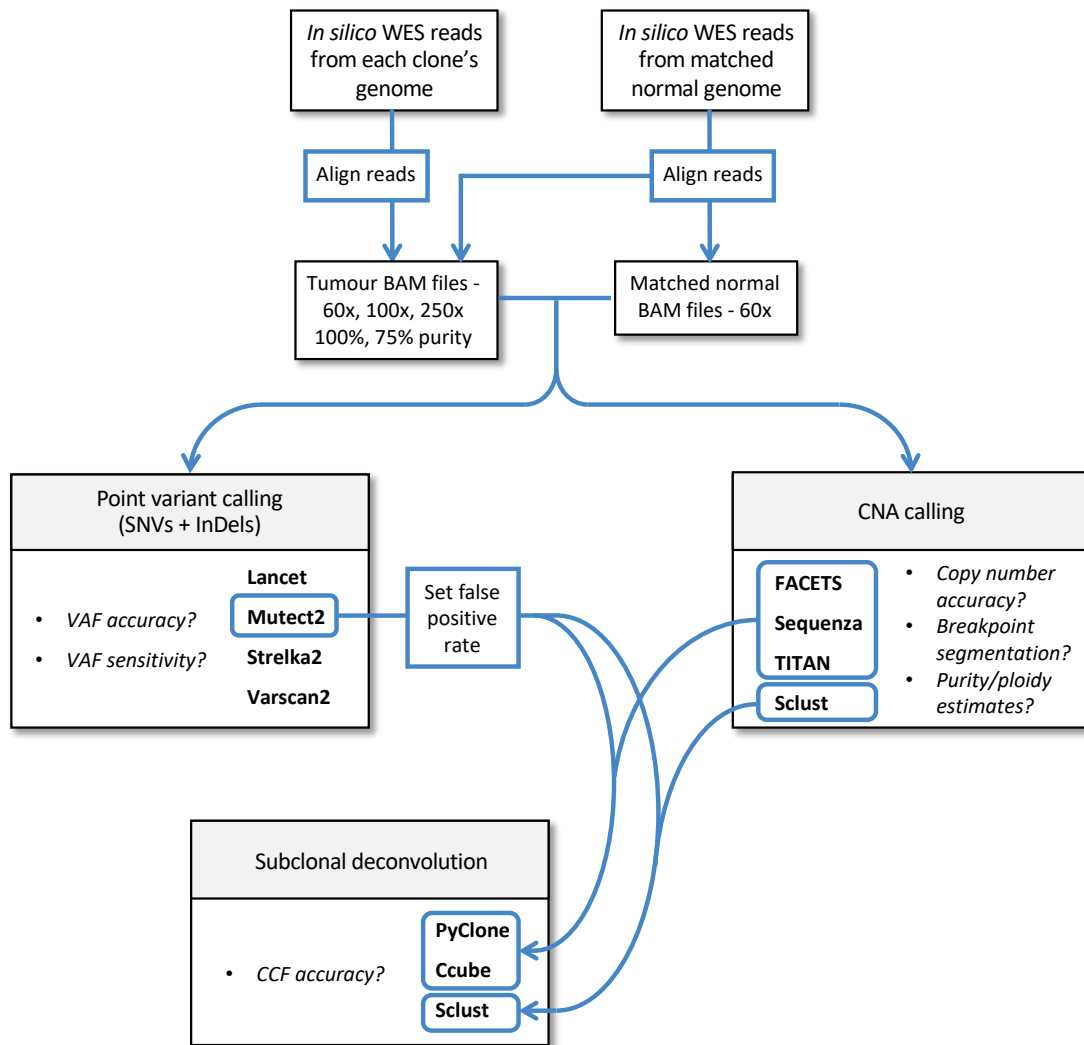


Figure 18. Flow diagram of the benchmarking process.

### 3.1.2 Variant calling

Variant calling, as referred to in this study, is the process of identifying both somatic SNVs and short InDels from sequencing data. Calling such variants involves identifying differences in the alleles between genomes in a tumour sample and matched normal sample, whilst attempting to ignore artefacts that result from DNA damage, polymerase chain reaction (PCR) errors, sequencing errors, or misalignment of reads. Many machine learning and probabilistic approaches exist for this, and are often equally applicable to both WGS and WES data (Xu, 2018). In order to detect low frequency clones via subclonal deconvolution, it's necessary to be able to detect variants that have very low variant allele frequencies (VAFs). This can make the distinction between true mutations and artefacts difficult. False positive calls may arise from base mismatches due to sequencing errors, PCR errors during library preparation, DNA damage (Chen *et al.*, 2017) or due to read misalignment, with the latter potentially resulting in multiple reads with mismatches at the same location. In addition, many tumour samples have been formalin-fixed and paraffin-embedded

(FFPE), causing extensive DNA damage throughout the genome, as well as limiting the availability of viable DNA for sequencing (Sikorsky *et al.*, 2007; Wong *et al.*, 2014; Wang *et al.*, 2015).

Variant callers aim to avoid these artefacts by applying one of a number of probabilistic methods in combination with sets of adjustable criteria that aim to distinguish true from false signals. This often involves features of the data such as, i) base quality score, indicating the confidence of signal from the sequencing machine, ii) read mapping score, indicating how likely a read is likely to be correctly aligned, iii) coverage across a variant, which determines the amount of evidence for a variant, iv) VAF, v) number of alternate allele supporting reads, and vi) strand-bias. Strand-bias describes scenarios where alternate allele supporting reads are significantly more prevalent on one strand. These can result from artefacts during library preparation, sequencing, alignment, or post alignment processing steps (Guo, Li, *et al.*, 2012), and therefore often lead to false positive calls that need to be removed. However, strand bias can also be a result of natural sampling variation or an effect of targeted capture where both reference and variant supporting reads map only to one strand, and in these cases variants should not be filtered out (Guo, Long, *et al.*, 2012). Filtering variants based on fixed thresholds of strand-bias removes many true positives, and many callers instead rely on more complex strand-bias filtering methods (Mutect2, Strelka2, VarScan2), or recommend not filtering based on strand-bias (Strelka2, VarScan2), particularly as newer sequencing preparations have less pronounced strand-bias effects. Incorrect read alignment is another source of error. Even when reads are mapped to the correct general location, some bases in the read may be aligned to incorrect positions, particularly if a variant is near the start or end of a read. To overcome this, many variant callers perform local assembly and realignment using information across all reads in the area so that potential variants are taken into account. Alternatively, this process can be carried out prior to variant calling using additional methods (Van der Auwera *et al.*, 2013; Mose *et al.*, 2019).

I aim to benchmark variant callers by assessing their performance in terms of features that are important to subclonal deconvolution. These include the sensitivities of callers across different VAFs, as well as the accuracies of VAF estimates. For this, I select four variant callers that are either commonly used, performed well in previous benchmarking, or are novel:

- **Mutect2** (Benjamin *et al.*, 2019). Mutect2 is part of the genome analysis tool kit (GATK) and is possibly the most commonly used somatic variant caller. It first generates candidate haplotypes by performing local assembly using de Bruijn-like graphs out of kmerized reads, with an adaptive pruning model to prioritise the most likely candidates while accounting for local depth and observed sequencing error rates. Reads are then locally realigned and a pair hidden Markov model assigns a likelihood for each read being derived from candidate haplotypes, as opposed to a result of a sequencing error. A log odds ratio (LOD) is then defined for each variant by comparing evidence for the model with and without the variant haplotype. FilterMutectCalls is a recommended filter for

Mutect2 call sets, which filters variants with the aim of removing artefacts resulting from errors such as those during sample preparation and alignment. Using annotations provided by the Mutect2 output, and several different error models, it filters on a threshold estimated to maximise the harmonic mean of recall and precision.

- **VarScan2** (Koboldt *et al.*, 2012): VarScan2 takes a less complex approach than the other included methods, using a heuristic process. First, candidate positions are called from both the tumour and normal samples. If these two match then they're compared to the reference to identify germline variants. If they don't match, then a Fisher's Exact test determines the significance of the difference, with those passing the probability threshold being classed as somatic variants. fpcfilter then employs numerous hard filters to reduce false positives resulting from artefacts.
- **Lancet** (Narzisi *et al.*, 2018). A relatively new method for variant calling, using localised coloured de Bruijn graphs for local assembly of reads. Unlike other methods that use local assembly, Lancet jointly analyses tumour and normal reads together, thereby increasing sensitivity, particularly for indels. A Fisher's Exact test is used to score and filter variants.
- **Strelka2** (Kim *et al.*, 2018): Also a relatively novel method, and builds on the commonly used original version, Strelka (Saunders *et al.*, 2012). Strelka2 aims to reduce runtime compared with other variant callers, through a number of modified approaches. It first estimates model parameters using only a small number of reads, saving time compared to using the whole dataset. It then employs a tiered method for calling variants, whereby a simple model based on input read alignments is used for less challenging positions, and local assembly and realignment is reserved for those more challenging. Finally, Strelka2 calculates a single empirical variant score to allow more informative prioritisation, using a pre-trained random forest model to aggregate information from numerous call-quality features.

### 3.1.3 CNA calling

Somatic CNA calling aims to identify regions that differ in copy number between genomes from a tumour sample and a matched normal sample, to determine alterations specific to the tumour. A number of different approaches have been developed to achieve this from genome sequencing data (Zhao *et al.*, 2013). Read depth methods compare numbers of aligned reads in a region between tumour and normal samples, with higher copy numbers resulting in increased read depth. Paired end, split read and *de novo* assembly-based methods aim to more precisely determine where break points are in a genome from reads, or inserts located across them, allowing copy numbers to be inferred at base level precision. Combinatorial methods incorporate multiple of the above approaches to improve accuracy. A limitation of WES is that

often break points are not sequenced, with only those that fall in exon regions able to be precisely detected. Therefore, read depth methods are the standard approach when calling CNAs from WES.

CNA detection can be impaired by numerous factors and, when performed on WES, is particularly susceptible to noise and systematic error. This largely results from variations in probe hybridisation affinities along the genome during the exon capture step (Zhao *et al.*, 2013). Other sources, common to both WGS and WES, include misalignment of reads, particularly in repeat genomic regions, and variation of GC content within a genomic region. Guanine base pairs to cytosine more strongly than adenine pairs with thymine, resulting in less fragmentation in GC rich regions and, therefore, may appear less in the pool of fragment sizes needed for sequencing (Benjamini and Speed, 2012). However, these can be largely accounted for during a normalisation step, and by comparing tumour samples against a matched normal, which will have similar patterns of bias (Figure 19).

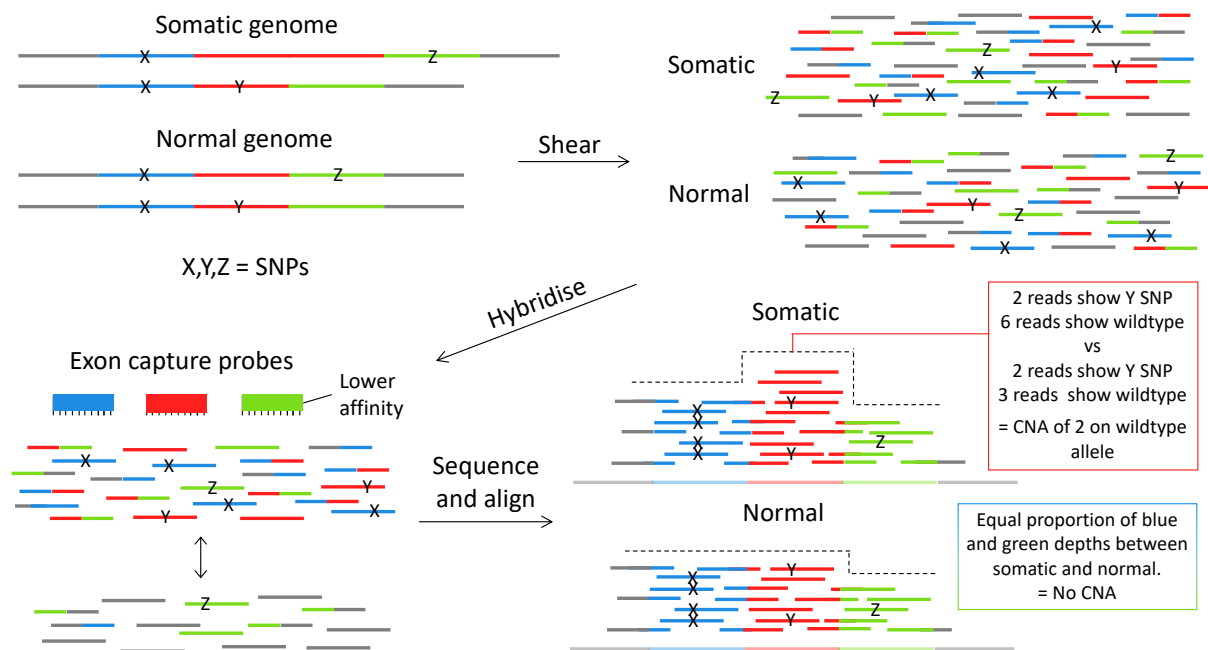


Figure 19. Diagram of WES with allele specific CNA calling. The different colours represent different regions of normal and somatic genomes, which may contain X, Y, or Z SNPs on one or both alleles. During exon capture hybridisation, the probes enrich the pool of sheared sequences for exon regions, though some probes (green) will be less efficient than others. By comparing matched normal and somatic samples, this bias is accounted for, and only regions with true CNAs identified, which can be assigned to a specific allele based on the proportions of SNP vs wildtype supporting reads.

Most read-depth based CNA calling methods follow a similar workflow. Firstly, read depths are normalised to account for variations such as mapability and GC-content, as well as differences between the overall sequencing depths of the tumour and normal samples. Next, the log-2 ratio (logR) between the normal and tumour total read depth is calculated for windows, with deviations from 1 suggesting a CNA. Segmentation is then carried out by one of a number of approaches that aim to remove noise in the copy number predictions along the genome through grouping similar logR regions into bins, thereby identifying breakpoint locations. Modelling then enables estimation of purity, ploidy and copy number values from the segmented logR bins. Some methods extend this to include modelling of subclonal CNAs, occurring in only a portion of the sample, and which can be incorporated into many subclonal deconvolution methods.

In addition to total copy numbers estimates, many subclonal deconvolution methods require allele specific copy number information, whereby estimates are provided for maternal and paternal, or major and minor, alleles separately. For this, B-allele frequencies (BAFs), which describe the fraction of reads covering a heterozygous SNP position that support the alternate allele, are used to appropriately split total read depths into allele-specific read depths. This allows estimation of allele specific copy numbers and detection of regions with loss of heterozygosity (Figure 19).

The method used to call CNAs can have a large influence on the accuracy of subclonal deconvolution (Bhandari *et al.*, 2018; Andor *et al.*, 2016). Therefore, I investigate the performance of four somatic CNA callers, and their impact on CCF estimation from subclonal deconvolution methods. These were chosen either due to being widely used, or because they were newer methods with the potential to improve on previous ones.

- **Sequenza** (Favero *et al.*, 2015): This method is one of the most popular for somatic CNA calling. Segmentation of allele-specific logRs at heterozygous SNP positions is achieved using a penalized least squares regression approach through the copynumber package (Nilsen *et al.*, 2012). A probabilistic model is then used to predict purity and ploidy through a grid-based search over reasonable values. The model parameters are estimated using a maximum *a posteriori* approach, with prior probabilities that prefer diploid states over others. Subclonal CNA events are not predicted using this method.
- **Sclust** (Cun *et al.*, 2018). Sclust performs both subclonal somatic CNA calling and subclonal deconvolution. First, a coarse segmentation method, similar to circular binary segmentation (Venkatraman and Olshen, 2007), is applied to allele-specific logRs, allowing initial estimates of purity and ploidy to be computed via a conditional maximum likelihood approach. A more sensitive segmentation and smoothing step is then performed, using a method similar to that of SegSeq (Chiang *et al.*, 2009). This provides segments for estimating CNA calls and updating purity and



ploidy values. The sensitive segmentation and CNA estimation steps are then repeated. An advantage of Sclust combining CNA calling with subclonal deconvolution is that tumour purity can instead be estimated from variant clusters when samples have insufficient numbers of CNAs.

- **TITAN** (Ha *et al.*, 2014): This method is another very popular somatic CNA caller. A two-factor hidden Markov model is used to cluster heterozygous SNP positions with similar frequencies, using total read depths, allele-specific read depths, and logRs. This jointly enables segmentation, identification of subclonal populations, and estimation of purity and ploidy. The model is run multiple times with varying numbers of clones, and then identifies the optimal number of clones via the best fitting model.
- **FACETS** (Shen and Seshan, 2016): Unlike most allele-specific CNA calling methods that perform segmentation using only heterozygous SNP positions, FACETS combines this with segmentation of total read count logRs at pseudo-SNP positions, thereby avoiding lack of signal in regions with few true heterozygous SNPs. This segmentation is achieved using a joint non-parametric approach based on a Hotelling  $T^2$  statistic, which has improved power over segmenting total and allele-specific logRs separately. Segments are then clustered into groups with the same copy number, and CNA calls are calculated from these using a Gaussian-non-central  $\chi^2$  mixture model, while factoring in estimates for purity and ploidy. Another unique aspect introduced by FACETS is the use of log-odds ratios between tumour and normal variant read counts at heterozygous SNP positions, instead of B allele frequencies, which are biased due to differences in mapping affinity for variant and reference reads.

### 3.1.4 Subclonal deconvolution

Subclonal deconvolution is the process of delineating sub-populations of cells by their distinct genotypes, from within bulk tumour samples, (Figure 4). It first requires estimation of the proportions of cancer cells carrying different somatic mutation, with a cancer cell fraction (CCF) estimated for each. A variety of mutation types can be used for this, however, CCF estimation is most commonly, and often most powerfully, performed on somatic point variants. Such an approach utilises VAFs, whilst accounting for CNAs across variant positions, normal cell contamination, and technical noise. While most recent methods do account for CNAs, many do not, and instead require either prior correction for CNAs or masking of variants in these regions. The latter risks removing important driver variants such as *EGFR* mutations in glioblastoma (GBM), which occur in a commonly amplified region of chromosome 7 (Miura *et al.*, 2018). Such methods are therefore not included in this study.

CCF estimation from VAFs is a challenging process, and is affected by the accuracy of both variant and CNA calling methods used to generate inputs, the accuracy of purity estimates from either CNA callers or histology, and the number of samples per tumour (Noorbakhsh *et al.*, 2018; Abécassis *et al.*, 2019; Bhandari *et al.*, 2018; Andor *et al.*, 2016). The approach requires predicting the number of copies of DNA in a cell that each variant lies on, termed the multiplicity. However, cancers have a tendency to undergo extensive structural rearrangements, causing numerous gains and losses that, depending on both timing and which alleles are affected, may or may not affect variants in those regions. Therefore, predicting the multiplicity and CCF from each variant from its VAF, even before considering the inaccuracies of CNA calling methods, involves a high level of uncertainty and requires potentially unrealistic limiting assumptions.

Once CCFs are estimated, the next step is to deconvolute the variants into distinct subclones, allowing refinement of individual variant frequencies and informing on co-occurring variants within the same subclones. This is often achieved via clustering, which relies on subclones having undergone some level of selection, so that their CCFs increase and become distinct from background mutations having accumulated as a result of neutral evolution. It is worth noting that clusters do not always define subclones, as some variants will be found in multiple subclones and therefore appear in a different cluster to those unique to one subclone (Figure 4). Other subclones may share the same frequency and therefore variants unique to each will appear in the same cluster. Such scenarios may be resolved by multi-sample analyses, so that changes in CCFs between samples are taken into consideration to allow better separation of distinct clusters. Other methods aim to improve identification of subclones and resolution of their boundaries, through clustering variants based on their mutational signatures in combination with CCFs, whilst also informing on evolutionary trajectories of mutational signature activity (Rubanova *et al.*, 2020).

Many subclonal deconvolution methods include a further step to infer the evolutionary relationships between subclones by fitting a phylogenetic tree to the CCFs (Kuipers *et al.*, 2017; Schwartz and Schäffer, 2017). Alternatively, some infer trees directly from VAFs and CNAs, thereby providing an alternative approach to CCF estimation. These require the use of limiting assumptions to make the state-space of possible phylogenetic trees manageable, though risk introducing errors. Many methods include the “infinite-sites” assumption, that states only one variant may occur at any one site and that variants do not revert (El-Kebir *et al.*, 2016; Jiang *et al.*, 2016). CNAs are handled in similar ways, though, due to the fact that CNAs over a region are likely to develop independently in separate cells, the assumptions are more relaxed; SPRUCE, for example, assumes variants may change state, or multiplicity, numerous times as a result of amplifications or deletions, but never to the same state more than once (El-Kebir *et al.*, 2016). However, this overlooks common scenarios, such as when a position gains a CNA that is later reverted by another. Canopy avoids this by considering CNA break points, allowing any state changes resulting from CNAs with differing break points or copy number, but those with the same occur only once (Jiang *et al.*,

2016). In this benchmarking study, I focus only on clustering approaches, as phylogenetic approaches often require multiple samples, result in numerous solutions, or emphasise manual curation of inputs.

- **PyClone** (Roth *et al.*, 2014). Possibly the most commonly used method for performing subclonal deconvolution, PyClone uses a hierarchical Bayes statistical model to estimate variant multiplicities and then cluster CCFs into an unfixed number of clones. The method is potentially limited compared to others included, by the assumption that all cells containing a variant have the same CNA genotypes. Additionally, it does not accept estimates of the cellular frequencies of subclonal CNAs as input, and instead considers all inputted CNAs as total copy numbers across all subclones. Due to this, I only include clonal CNAs when using it with calls from FACETS and TITAN.
- **Ccube** (Yuan *et al.*, 2018). This method was found to perform best in a recent benchmarking study using simulated call sets from distributions (Dentro *et al.*, 2020), and it's therefore of interest to determine if it performs similarly well on the more complex datasets used in this study. Unlike methods that prefix multiplicities prior to clustering, Ccube estimates them during, under the assumption that multiple variants will share the same CCF. Clustering is achieved via a Bayesian mixture model approach, with the number of clones defined by a truncated Dirichlet Process. The method is able to take into account subclonal CNAs from two copy number populations.
- **Sclust** (Cun *et al.*, 2018). This method performs both CNA calling and subclonal deconvolution. After CNAs are identified for both clonal and subclonal populations (as described above), the most likely variant multiplicities are identified, given the available copy number states. Clusters are next determined by deconvoluting intrinsic sampling noise from the unknown distribution of subclonal populations, using smoothing splines. Variants are then assigned to the most likely cluster.

## 3.2 Results

### 3.2.1 Creation of datasets for use in benchmarking

#### 3.2.1.1 Genome simulation

Using HeteroGenesis, I simulated genome sequences for clones and matched germlines for each of nine tumours. These consisted of three replicates (R1, R2, R3) from each of three parameter sets (S1, S2, S3), that varied in numbers of somatic mutations (Figure 20). Mutation frequencies, along with those for parameters consistent between all sets, were chosen to reflect values estimated in real germline and GBM tumours (Table 3). In particular, the highest density variant set is included to represent tumours having undergone i) hypermutation, which can result from chemotherapeutic agents such as temozolomide, and ii) chromothripsis, a term describing chromosomal shattering, resulting in tens to thousands of chromosomal

rearrangements and CNAs in a minority of samples across tumour types (Zack *et al.*, 2013; Rode *et al.*, 2016). The clonal architecture, consisting of 8 related clones (Figure 20), is representative of the complexity of scenarios seen in GBM previously (Körber *et al.*, 2019; Johnson *et al.*, 2014).

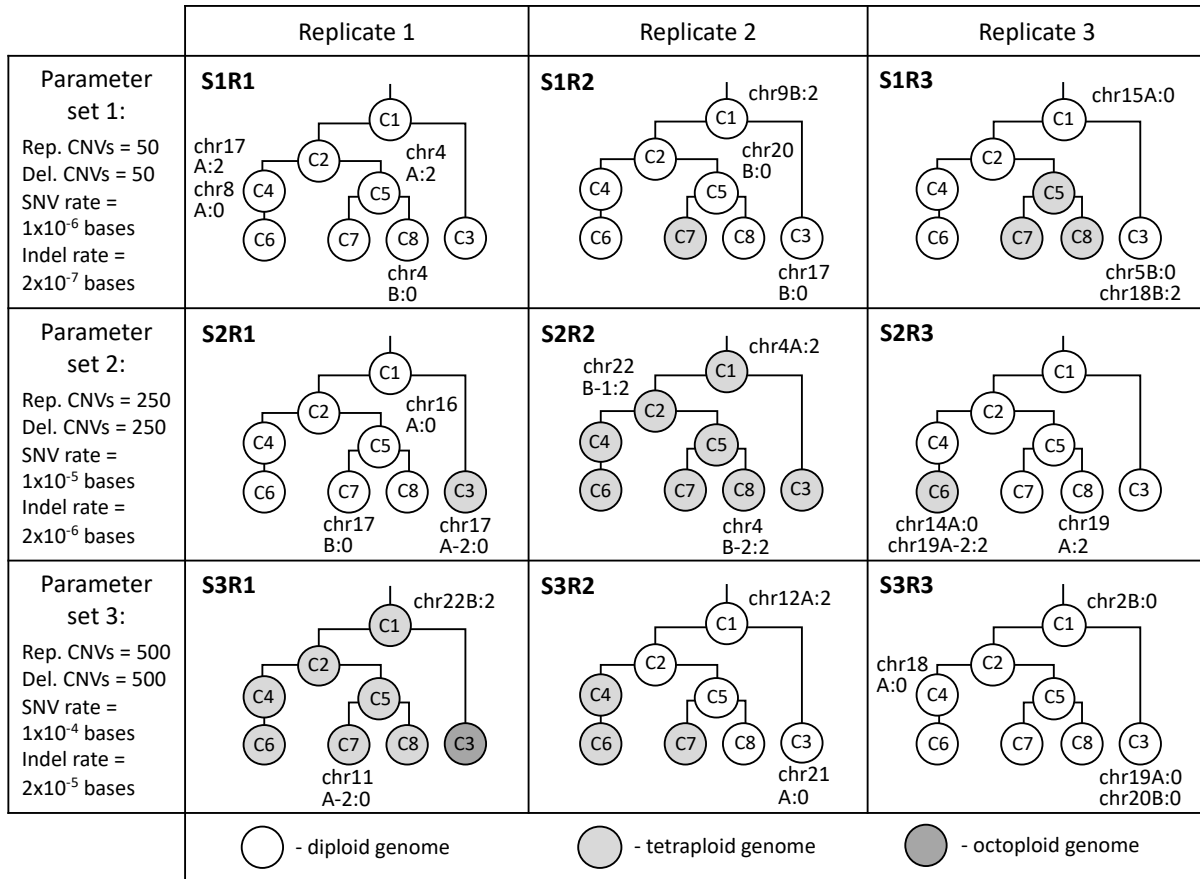


Figure 20. Tumour clonal architecture designs created by HeteroGenesis. Clones are named C1...C8. All relative distances of clones from their parent clone (or from the germline for C1) were set at 1, with the exception of C3 which had a relative distance of 3. Light grey circles indicate clones with tetraploid genomes, and dark grey indicates clones with octoploid genomes. Single chromosome aneuploid events are indicated next to the clone in which they first appear, but are also present in all daughter clones. Three different parameter sets were used to define total somatic mutation numbers in each tumour, with three replicates created for each set.

Table 3. Parameter values used with HeteroGenesis in simulating tumour genomes.

Parameter	Values	Reference
Germline SNV rate	$1.4 \times 10^{-3}$ errors/base	(1000 Genomes Project Consortium <i>et al.</i> , 2015; Shen <i>et al.</i> , 2013)
Germline InDel rate	$1.4 \times 10^{-4}$ errors/base	(Mills <i>et al.</i> , 2006; Shen <i>et al.</i> , 2013)
Germline CNVs per genome	360	(1000 Genomes Project Consortium <i>et al.</i> , 2015)
Somatic SNV rate	$1 \times 10^{-4}$ , $1 \times 10^{-5}$ , $1 \times 10^{-6}$ errors/base	(Kandoth <i>et al.</i> , 2013)
Somatic InDel rate	$2 \times 10^{-5}$ , $2 \times 10^{-6}$ , $2 \times 10^{-7}$ errors/base	(Mills <i>et al.</i> , 2006; Mullaney <i>et al.</i> , 2010)
Proportion of germline SNVs taken from dbSNP	0.9	(Durbin <i>et al.</i> , 2010; Shen <i>et al.</i> , 2013)
Proportion of germline indels taken from dbSNP	0.5	(Durbin <i>et al.</i> , 2010; Shen <i>et al.</i> , 2013)
Somatic CNAs per tumour (equal replications to deletions ratio)	100, 500, 1000	(Xi <i>et al.</i> , 2011; Droop <i>et al.</i> , 2018)
Aneuploid events per tumour	4	(Hu <i>et al.</i> , 2013; Baysan <i>et al.</i> , 2017)
Probability that an aneuploid event is a whole genome duplication	0.333	-
Germline CNV lengths distribution	Log-normal distribution with $\mu = -10$ and $\sigma = 3$ multiplied by $1 \times 10^6$ , $\geq 50b$ .	(Krijgsman <i>et al.</i> , 2014)
Somatic CNA lengths distribution	Log-normal distribution with $\mu = -1$ and $\sigma = 3$ multiplied by $1 \times 10^6$ , $\geq 50b$ .	(Krijgsman <i>et al.</i> , 2014; Droop <i>et al.</i> , 2018)
Germline and somatic indel lengths	Log-normal distribution with $\mu = -2$ and $\sigma = 2$ , $\leq 50b$ .	(Mills <i>et al.</i> , 2006)
CNV/CNA copy number	Log-normal distribution with $\mu = 1$ and $\sigma = 0.5$ , rounded down to an integer, $\neq 1$ .	(Droop <i>et al.</i> , 2018)

### 3.2.1.2 *In silico* sequencing and creation of BAM files.

To create WES datasets that represent bulk samples from each of the tumours, I first *in silico* sequenced each clone and germline in the tumours to a depth of 134x using w-Wessim, with the error model and real WES reads used as probes, as described in chapter 2. After aligning the reads to the hg38 reference genome, I subsampled and merged the resulting BAM files at varying proportions to create tumour samples with a mixture of clones, as indicated in Figure 21, and with varying sequencing depths and purities. GBM tumours typically have high purity, with an estimated median of 85% purity from samples in the Cancer Genome Atlas (Aran *et al.*, 2015). I therefore created samples with 100% and 75% purity, which represent easy and more challenging cases, respectively, for analysis methods to handle, and encompass the majority of GBM samples. Normal samples were created with 60x sequencing depths, whereas tumour samples were created with 60x, 100x, and 250x sequencing depths, thereby covering ranges commonly seen in WES.

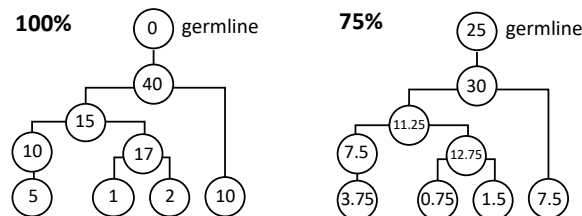


Figure 21. Clone proportions in 100% and 75% purity samples for each tumour. Circles correspond to clones represented in Figure 20, with numbers indicating the percentage of total bulk reads taken from a clone.

The freqcalc module of HeteroGenesis then allowed me to consolidate and recalculate variant profiles generated for the individual pure clones, into those that describe the heterogeneous tumour samples, thereby providing the ground truths for benchmarking.

InDel realignment was performed in pairs of tumour and corresponding normal samples with GATK (Van der Auwera *et al.*, 2013). This is likely to improve results for methods without their own inbuilt local realignment steps, such as VarScan2.

## 3.2.2 Benchmarking of mutation calling and CCF estimation methods

### 3.2.2.1 Variant calling

I first assessed the performance of variant calling methods. For this, I took a set3 tumour (S3R1\_75%) so as to have a large number of variants included, and at three sequencing depths (60x, 100x, and 250x). When choosing parameters to set for each method tested, it is not always possible or appropriate to match these between callers, nor would it be a fair comparison to use only default settings, which are sometimes

unsuitable for detecting variants with low VAFs. I therefore adjusted parameters only when necessary to allow detection of low frequency variants and to standardise the minimum required coverage across callers (Appendix Table 1). Some variant calling methods include a secondary filtering step, and so to determine the effectiveness of these, I compared call sets both pre and post filtering. Mutect2 was also ran with a third set up where the clustered\_events flag was ignored during filtering, due to an observation that it was removing many true positives.

Receiver operator curves (ROCs), which illustrate the trade-off between true positive rate against false positive rate, were plotted using each callers' outputted probability or score thresholds as parameters (Figure 23). To allow full ROCs to be generated, I needed to prevent the variant callers from throwing out candidate variants that do not reach the default score threshold. For Varcan2 and Lancet, I simply set the score threshold values to the minimum stringency. Strelka2 already outputs the full range of candidate variants, and instead flags those that pass the minimum threshold score. With Mutect2, variants are not called independently of each other, particularly with regards to filtering clusters of variants. Also, the additional filtering step estimates the optimum threshold based on modelling of the data, and therefore, lowering the threshold in the initial calling to create full ROCs is not appropriate and substantially affects the results. Therefore, I ran Mutect2 with default thresholds, resulting in partial ROCs (Appendix Table 1).

The runtimes of the different callers showed significant variation. VarScan2 had the shortest run time, followed by Strelka2 and then Mutect2. Lancet had the longest run time by far, 500 times greater than VarScan2 (Figure 22). It should be pointed out that, while VarScan2 had a very short run time, it is the only method that does not have its own local realignment step and therefore requires a relatively time consuming prior indel realignment step.

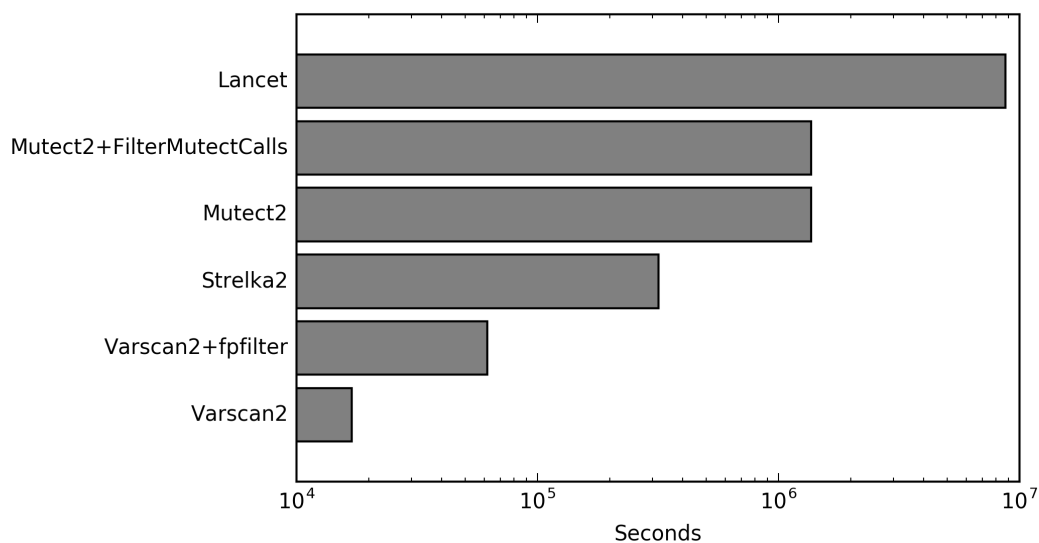


Figure 22. Runtimes of variant calling pipelines when used on the S3R1\_75%\_250x sample. Values include the time required to run additional programs necessary for each method, as indicated in Table 4. Runtimes for Lancet and Strelka2, which were run multi-threaded across 24 cores due to architecture time restraints, are multiplied by this number to indicate runtimes for a single core. Mutect2 and Lancet, were run on individual chromosomes separately, with runtimes summed for each. All pipelines were run on the same architecture.

Most variant callers showed very high false positive rates (Figure 23) and low precisions ( $\text{true positives}/(\text{true positives} + \text{false positives})$ ) and recall ( $\text{true positives}/(\text{true positives} + \text{false negatives})$ ) (Table 4).

Table 4. Precision and recall of the tested variant callers.

Method	60x		100x		250x	
	Precision	Recall	Precision	Recall	Precision	Recall
Strelka2	0.10	0.22	0.07	0.25	0.04	0.31
Lancet_1	0.39	0.16	0.39	0.18	0.58	0.21
Lancet_2	0.02	0.23	0.03	0.24	0.05	0.27
VarScan2	0.32	0.19	0.45	0.19	0.56	0.21
VarScan2_filtered	0.67	0.16	0.71	0.17	0.79	0.18
Mutect2	0.02	0.23	0.02	0.27	0.01	0.36
Mutect2_filtered	0.25	0.18	0.36	0.20	0.19	0.20
Mutect2_filtered_ic	0.24	0.20	0.23	0.23	0.14	0.29



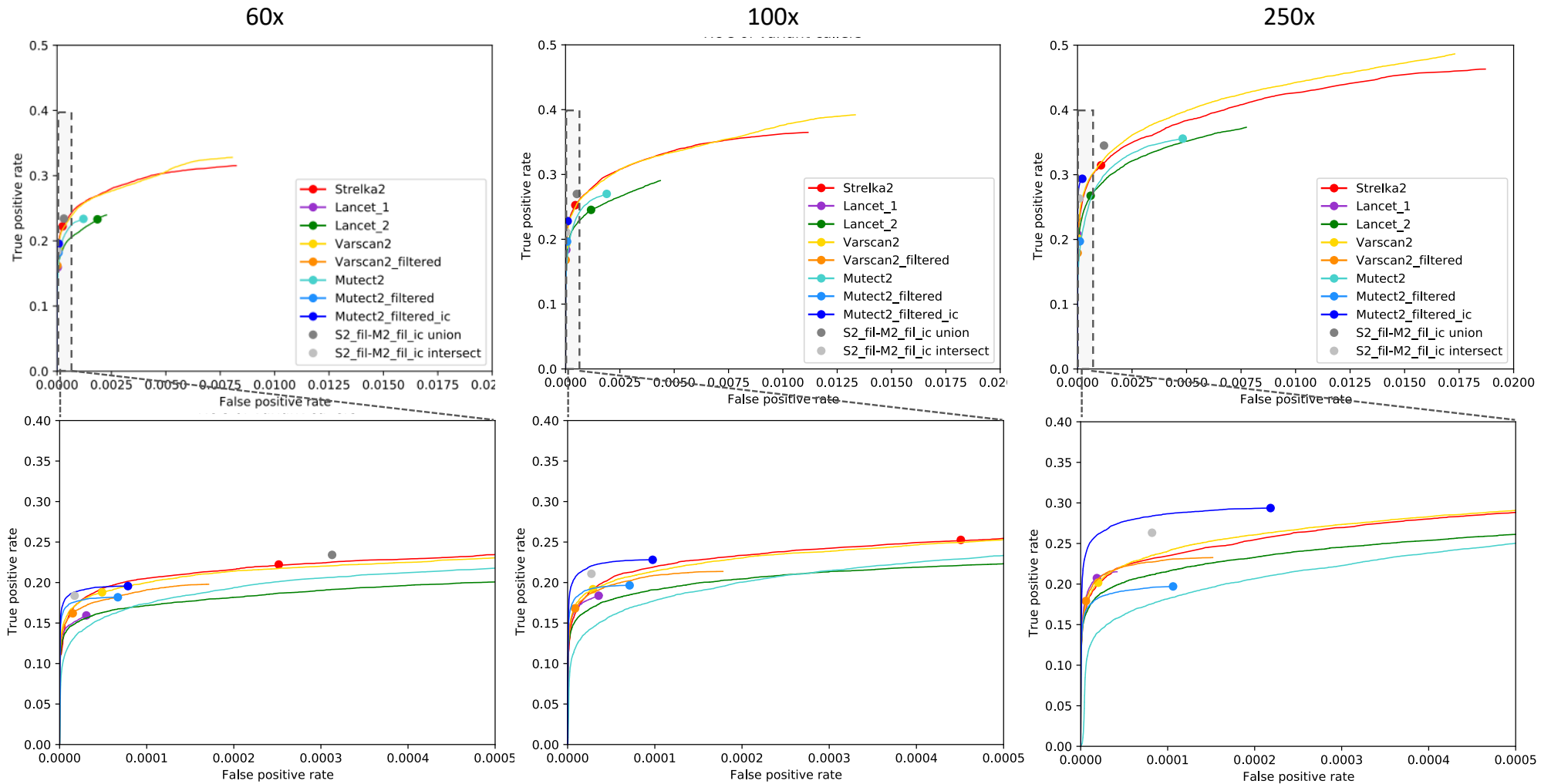


Figure 23. ROCs for each variant calling method, showing true positive rate (recall) against false positive rate, as a function of probability or score thresholds. Variant callers were used with sequencing datasets of S3R1\_B at 60x, 100x and 250x sequencing depths. Positions were limited to those with  $\geq 8$  reads in both the tumour and normal samples. Circles indicate the performance at the default threshold for each caller, or in the case of Strelka2, indicate the performance when variants are filtered (which is almost entirely based on a score threshold). Default values are; Strelka - aggregate variant score : SNVs=7, InDels=6; Lancet\_1/Lancet\_2 - Fisher's exact test score =5; Mutect2/Mutect2\_filtered(ic) - LOD=3; VarScan2 - P value=0.05. The performance when taking the union or intersect of calls from Strelka2 and Mutect2\_filtered\_ic is also plotted. Mutect2\_filtered\_ic: Mutect2\_filtered with the clustered\_events flag ignored. S2\_fil-M2\_fil\_ic: Union and intersect of Strelka2 and Mutect2\_filtered\_ic.

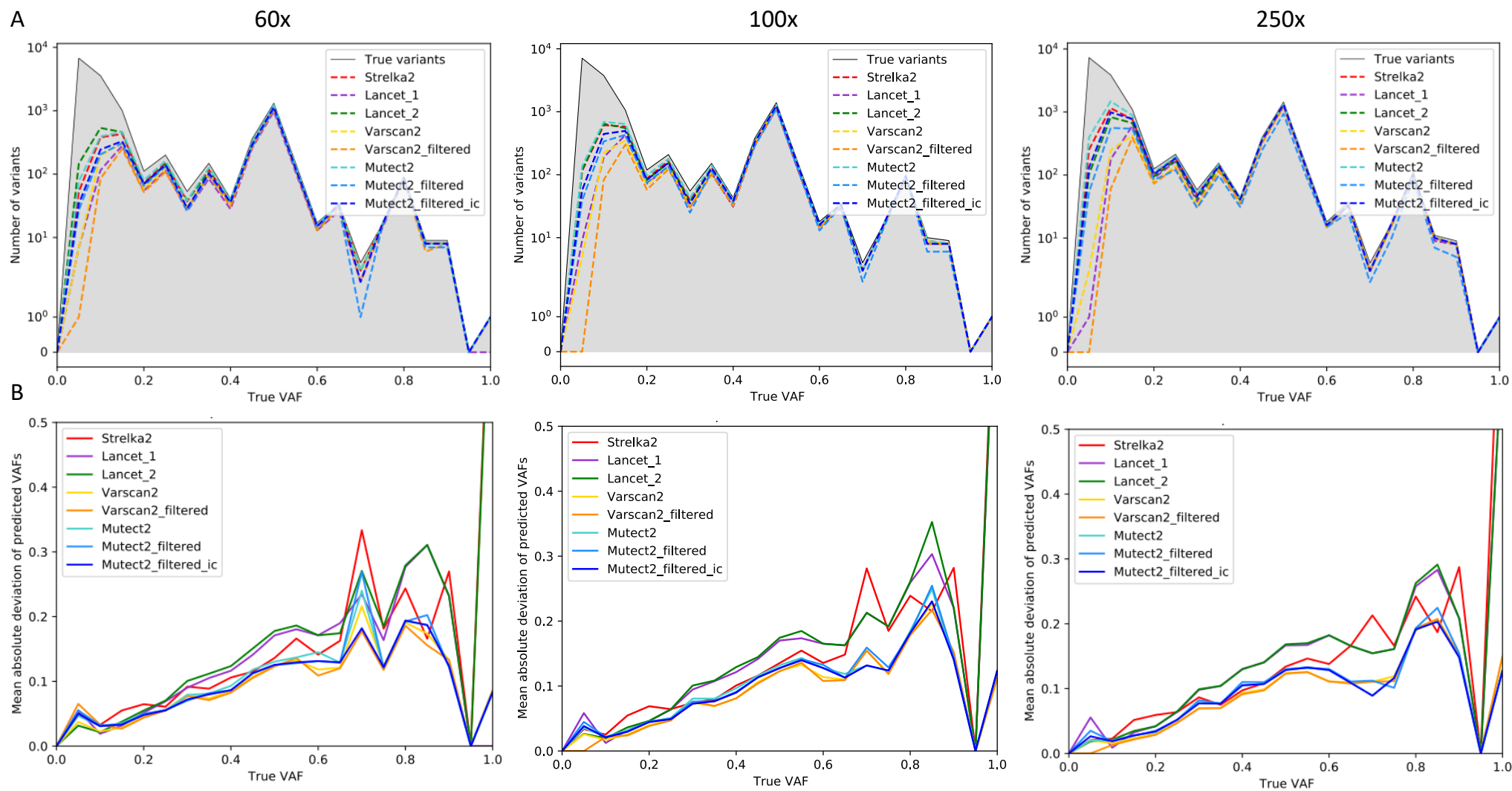


Figure 24. A) Numbers of true variants, on a log scale, called by each variant calling method at different true VAFs, from S3R1\_75% at 60x, 100x and 250x sequencing depths. The numbers of true variants are shown shaded in grey. B) Mean absolute difference of predicted VAFs from true VAFs, across different true VAFs, for each of the variant calling methods.

A proportion of false positives is expected due to the challenging task of distinguishing between true base changes occurring in only a small proportion of cells, and those resulting from artefacts or misalignments. However, this may have been exacerbated by the higher than expected quality scores for error bases in this dataset, introduced during *in silico* sequencing (as discussed in chapter 2). Varscan2\_filtered achieved good precision, up to 0.79, likely as a result of its hard cut-off for minimum VAFs. In contrast, Strelka2 and Mutect2\_filtered, which instead rely more on base quality scores instead of hard filtering of VAFs, had low precisions varying from 0.03-0.24. Estimates from other recent studies suggest precisions for Strelka2 and Mutect2 in the range of 0.4-1.00 (Benjamin *et al.*, 2019; Chen *et al.*, 2020; Bohnert *et al.*, 2017; Narzisi *et al.*, 2018; Detering *et al.*, 2019). It's possible that some of these studies overestimate precisions, due to the methods they use. For example, Detering *et al.* spike-in variants into BAM files to generate a test dataset with known ground truths. This does not reflect false positives that result from misalignment of variant supporting reads, as variants are inserted into the 'correct' position in the aligned BAM files; Benjamin *et al.* uses WGS sequencing for validation of variant calls from WES, whilst ignoring those where the WGS is underpowered to determine the variant allele presence, potentially missing false positives. Additionally true calls may be misclassified as true positives where misalignment cause the same false positives in both WES and WGS datasets; Bohnert *et al.* utilised mixtures of samples from the Genome in a Bottle consortium, which provide high confidence mutation calls using multiple technologies and pedigree information, but these still have a level of uncertainty. Nonetheless, true precisions are likely underestimated in the current study.

To address the likelihood of unrealistically high false positive proportions in the call sets, which may affect the performance of subclonal deconvolution methods, I subsample them to more commonly estimated proportions, reflecting precisions of 1.00, 0.9, and 0.5. This range of values allows for an assessment of how the precision of call sets affects the accuracy of subclonal deconvolution.

The recall (true positive rate) of all callers was low (Table 4, Figure 23), as expected due to the highly heterogeneous make-up of the samples (Figure 21). Most true variants were not detected by any caller, but this was heavily dependent on VAF (Figure 24A). Overall, Mutect2 called the most true variants, but there was a substantial decrease when FilterMutectCalls was applied. After looking into the reasons for this, I found that the clustered\_events flag was responsible for the highest number of filtered true positives, and also the highest proportion of filtered true positives relative to filtered false positives. When the filter was modified to ignore the flag, the true positive rate was drastically improved with relatively little increase in false positive rate (Figure 23). The flag responsible for removing the next highest relative proportion of true variants was "strand\_bias". As w-Wessim is unable to model some major sources of strand-bias affecting real data, such as PCR errors, it is not possible to fully assess the beneficial effect from the filter, but it is still interesting to note that it removes a substantial number of true positives.

After unfiltered Mutect2, Strelka2 had the next highest recall (Table4, Figure 23). In contrast, Lancet\_1, VarScan2, and VarScan2\_filtered had low recall and were less able to detect low frequency variants (Figure 24A). Lancet\_2 achieved better recall, but this came at a cost of much lower precision, particularly at higher sequencing depths (Figure 23). Two additional call-sets are analysed, consisting of the union and intersect of calls from Mutect2\_filtered\_ic and Strelka2, in order to investigate the benefits of ensemble variant calling approaches. Whilst the union calls did improve the true positive rate compared to either Mutect2\_filtered\_ic or Strelka2 alone, this came at a cost of an increased false positive rate. On the other hand, taking the intersect reduced both the true and false positive rates (Figure 23, Table 4).

In order to achieve accurate VAF and CCF estimates, variant callers must provide accurate numbers of variant and reference supporting reads, while discounting those that are likely incorrect base calls or mapped to the wrong position. I therefore assessed the accuracy of the read counts reported by each caller, through calculating VAFs from them (*variant supporting reads/total reads*) and comparing these to true VAFs. The callers generally showed similar accuracies, though Lancet showed comparatively poorer accuracy at higher VAFs (Figure 24B).

### 3.2.2.2 Copy number calling

Many previous benchmarking studies have assessed CNA calling through the proportion of CNAs called correctly as gain or loss, with a certain proportion of overlap between true and called CNAs (Nam *et al.*, 2016; Shen and Seshan, 2016; Zare *et al.*, 2017). This, however, is not appropriate in this study given the complexity of the simulated genomes, with overlapping CNAs at varying cellular frequencies, some of which may be grouped between or split into separate CNA calls. In addition, it's important for subclonal deconvolution that CNAs have accurate predictions of copy numbers (as opposed to just gain vs. loss), though some large scale errors, such as those resulting from false whole genome duplication ploidy predictions, are less problematic than other errors. I therefore assess CNA calling performance in three ways:

- 1) Purity and ploidy estimates from each CNA calling method, which must be accurate for reliable CNA calls, are directly assessed through correlation to true values.
- 2) Heatmaps of predicted and true total copy numbers along the genome are generated to allow a visual assessment of the relative performance of methods and the factors affecting each of them.
- 3) CNA call sets from the different callers are used with subclonal deconvolution methods, and the accuracy of the resulting CCF estimates are compared.

Compared to point variant calling, CNA calling performance is more likely to vary between samples due to the effects of purity and ploidy. Therefore, for CNA benchmarking, I used all nine simulated tumours, at

both 100% and 75% purity, and at three sequencing depths. Four callers were assessed, using default parameters in most cases.

I first assessed the accuracy of tumour ploidy and purity estimates from the callers, by quantifying the mean absolute difference (MADif) from true values (Figure 25, Table 5). Estimates were highly accurate for FACETS, TITAN, and Sclust. Sequenza’s estimates were also accurate in most cases, however it drastically underestimated purity and overestimated ploidy for many samples. As expected, none of the callers were able to detect whole genome duplications, and instead estimated them as being diploid. Sequencing depth was not found to affect ploidy or purity estimates (Figure 25, Table 5).

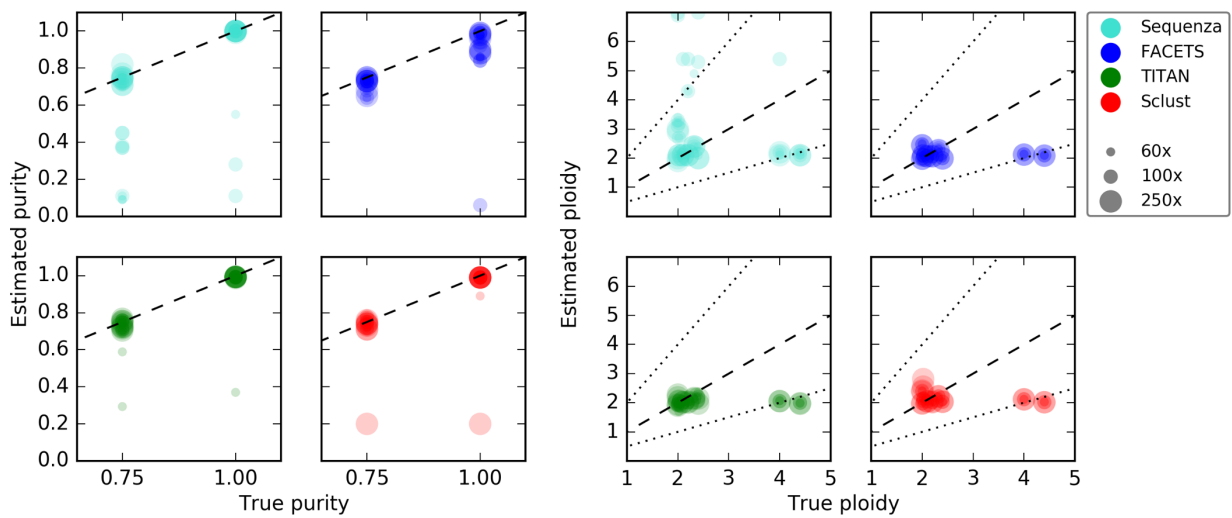


Figure 25. Accuracy of purity and ploidy estimates from the tested CNA callers. Dashed lines represent accurate ploidy and purity estimates, and dotted lines indicate ploidy estimates where the caller misclassified samples as tetraploid instead of diploid, or vice versa, but were otherwise accurate.

Table 5. Mean absolute difference of purity and ploidy estimates from true values, for each CNA caller and with different sequencing depths.

Feature	Depth	Sequenza	FACETS	TITAN	Sclust
Purity	60x	0.141	0.048	0.083	0.018
	100x	0.225	0.097	0.015	0.013
	250x	0.011	0.044	0.014	0.087
	Combined	0.126	0.063	0.037	0.039
Ploidy	60x	0.946	0.59	0.579	0.614
	100x	2.123	0.592	0.581	0.613
	250x	0.646	0.595	0.612	0.672
	Combined	1.238	0.592	0.591	0.633

Next, I generated heatmaps of amplifications and deletions along the genomes for predicted and true copy numbers for each CNA caller (Appendix Figure 1). These again showed that sequencing depth had little effect on performance. An exception to this was with Sequenza, which called many very short false CNAs in only the 250x samples, as a result of over-segmentation. Tumour purity also had little effect, with only a few examples of CNAs detected in the 100% sample but missed in the 75%. Conversely, Sclust detected more true CNAs in S2R1\_75% than in S2R1\_100%. As expected, all methods, and particularly Sclust, failed to detect CNAs present in very low frequency clones, where the total copy number remains close to two. Both TITAN and Sequenza falsely called large regions of heterozygosity on S1 samples, and some S2 samples in the case of Sequenza. None of the methods were able to detect subclonal whole genome duplications, with the exception of Sequenza which called them as clonal. Sequenza also frequently falsely predicted whole genome duplications or higher whole genome copy numbers.

### 3.2.2.3 Subclonal deconvolution

Using the Mutect2\_filtered\_ic variant calls, with controlled false positives, in combination with the CNA calls and purity estimates from Sequenza, FACETS or TITAN, I prepared inputs for subclonal deconvolution. I used these with two subclonal deconvolution methods; Ccube and Pyclone. A third method, Sclust, was also included, using the same Mutect2 calls, but with its own CNA and purity predictions. All methods were run with default values, with the exception of Sclust, which required a slightly larger smoothing parameter to allow completion for some runs, as indicated by the authors (Cun *et al.*, 2018).

Accuracy of subclonal deconvolution pipelines, and factors that affect them, was assessed by comparing the estimated variant CCFs with known true CCFs from the HeteroGenesis outputs. Whilst this doesn't assess how accurately the methods deconvolute variants into distinct clones, it instead focuses on the likely more critical and informative step of CCF estimation that provides the input into clone assignment. Furthermore, CCFs themselves are often used by researchers for investigating changes in cellular frequencies across time points (Barthel *et al.*, 2019).

True CCFs for the simulated tumours take a large number of values, partly due to the high number of clones ( $n=8$ ), but also from deletions that result in daughter clones not having the full set of variants of the parent. To assess the effect that different factors have on CCF estimation, comparisons between estimated and true CCFs are plotted for groups of combined samples from different scenarios. As a reference, an additional plot is shown where CCFs are estimated by simply doubling VAFs (to assume a purely heterozygous diploid genome) and dividing by true purity (which is often accurately estimated by CNA calling methods, and therefore not much of a biased advantage), and limiting to  $\leq 1$ . This represents a baseline for which CCF estimation should improve on. To achieve a suitable metric for comparison, I calculated the mean absolute difference (MADif) of true vs. estimated CCFs, across variants in all samples in

a group. Each sample was weighted equally so that those with the highest variant numbers did not dominate the results:

$$MADif = \frac{1}{n} \sum_{i=1}^n \frac{\sum_{j=1}^m |T_{S_{ij}} - E_{S_{ij}}|}{m}$$

where n is the number of samples (S) in a group, m is the number of called true variants, and T and E are the true and estimated CCF values for a variant.

I first determined the effect that tumour purity has on the accuracy of CCF estimates. Compared with 100% pure tumours, those with 75% purity resulted in slightly higher MADif values for all pipelines (Figure 26). However, there was also an increase in MADif for the 75% samples when using doubled VAFs divided by true purity, and as purity, ploidy and CNA calls did not appear to be negatively affected by lower purity, the increase may just reflect lower numbers of somatic reads in the 75% purity samples.

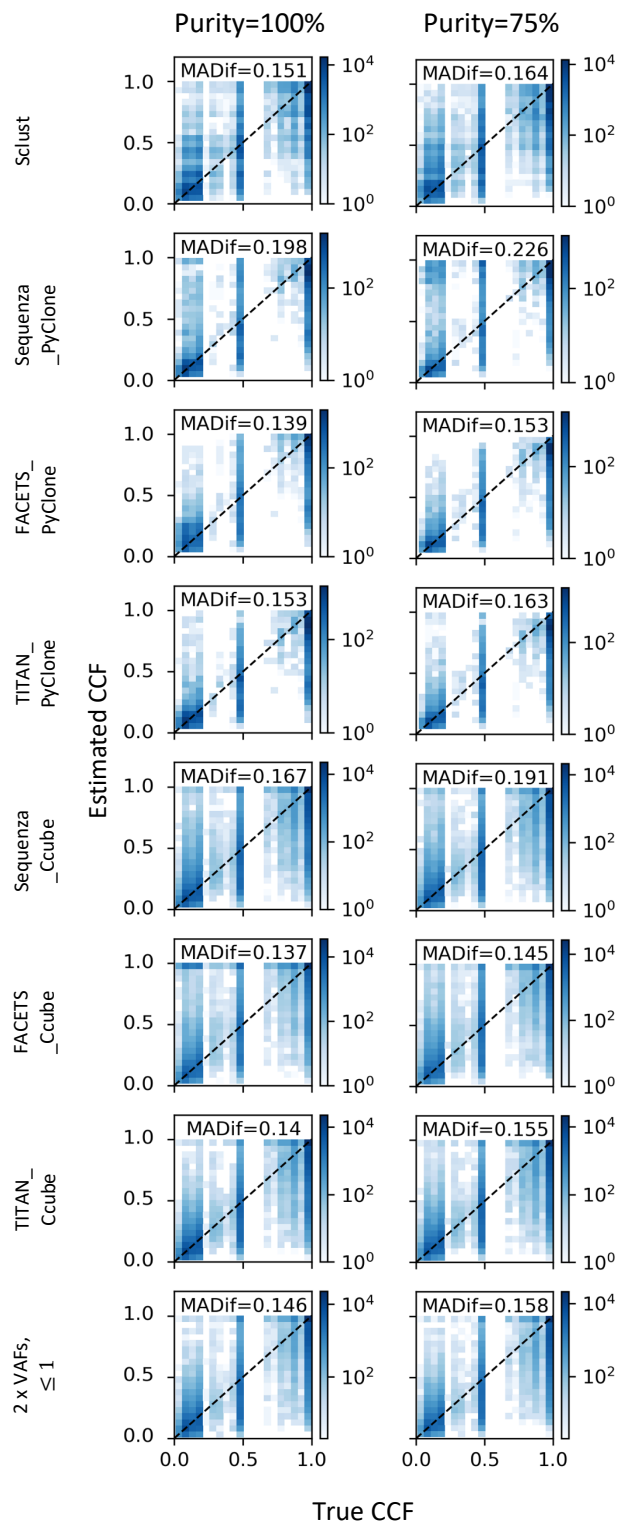


Figure 26. Estimated vs. true CCF values for all true positive Mutect2\_filtered\_ic called variants across 100% or 75% purity samples from seven different pipelines. The bottom panel shows estimated CCFs that are instead calculated by doubling VAFs and dividing by true purity, up to a maximum value of 1. Estimated CCFs in the Ccube pipelines are also limited to 1, as these can otherwise be higher due to modelling error.

Next, I looked at the effect call set precision had on the CCF estimates of true positives. Across all pipelines, there was no increase in MADif with increasing false positive proportion (Figure 27), suggesting that it's not important to stringently filter out false positives for CCF estimation, though it can't be ruled out that false positives have a more negative effect on the deconvolution clustering step.



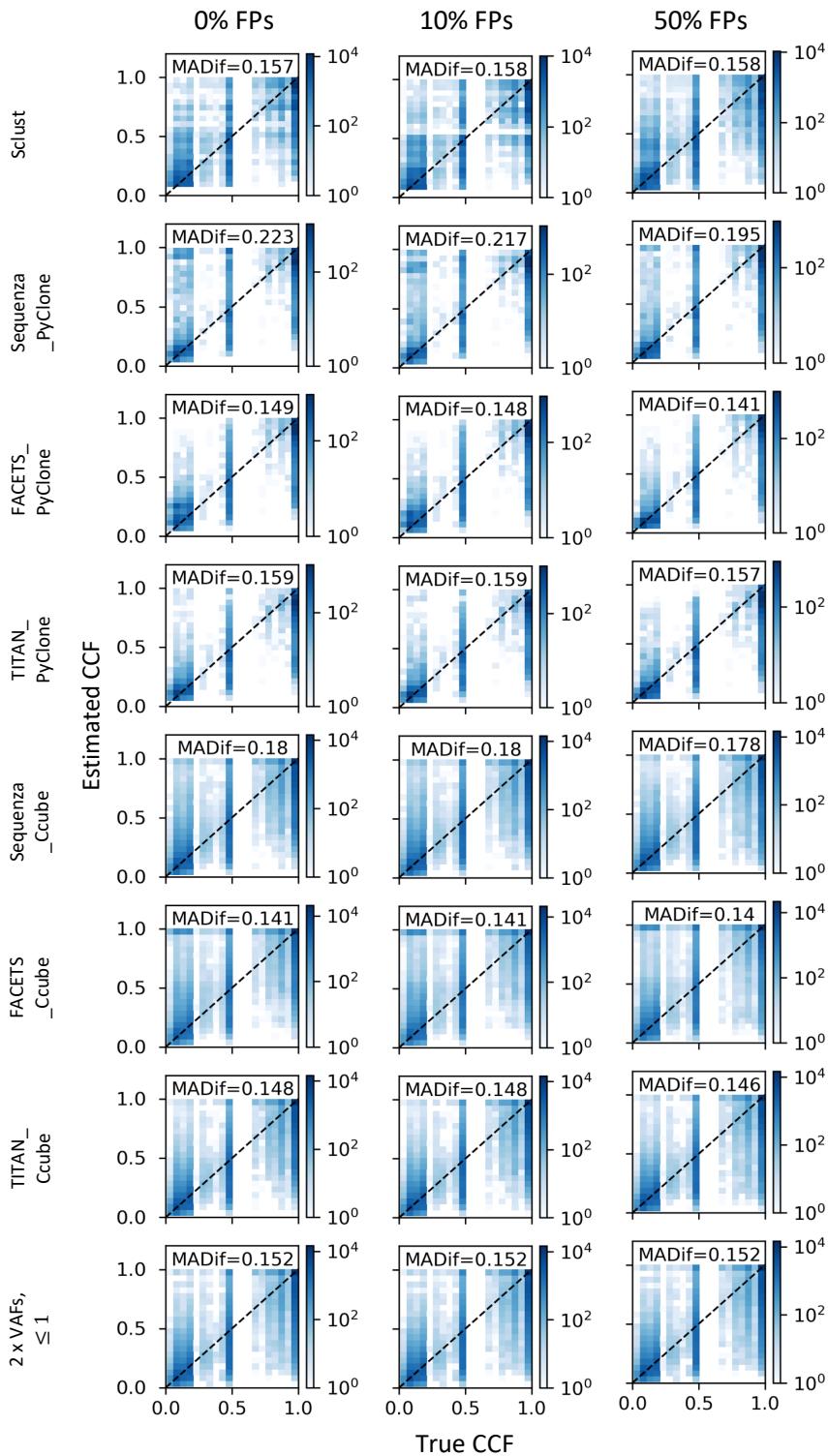


Figure 27. Estimated vs. true CCF values for all true positive Mutect2\_filtered\_ic called variants across all samples, from seven different pipelines, with precisions of 1.00, 0.9, and 0.5. The bottom panel shows estimated CCFs that are instead calculated by doubling VAFs and dividing by true purity, up to a maximum value of 1. Estimated CCFs in the Ccube pipelines are also limited to 1, as these can otherwise be higher due to modelling error.

I then assessed whether higher mutation densities affected the accuracy of CCF estimates, by comparing results between the three mutation parameter sets. No obvious differences were seen between the different parameter sets (Figure 28). It's likely that there exist opposing effects between higher numbers of CNAs leading to larger portions of the genomes with incorrect copy number calls, and higher CNA numbers providing more data points for estimating purity and ploidy, with pipelines being affected differently for each sample.

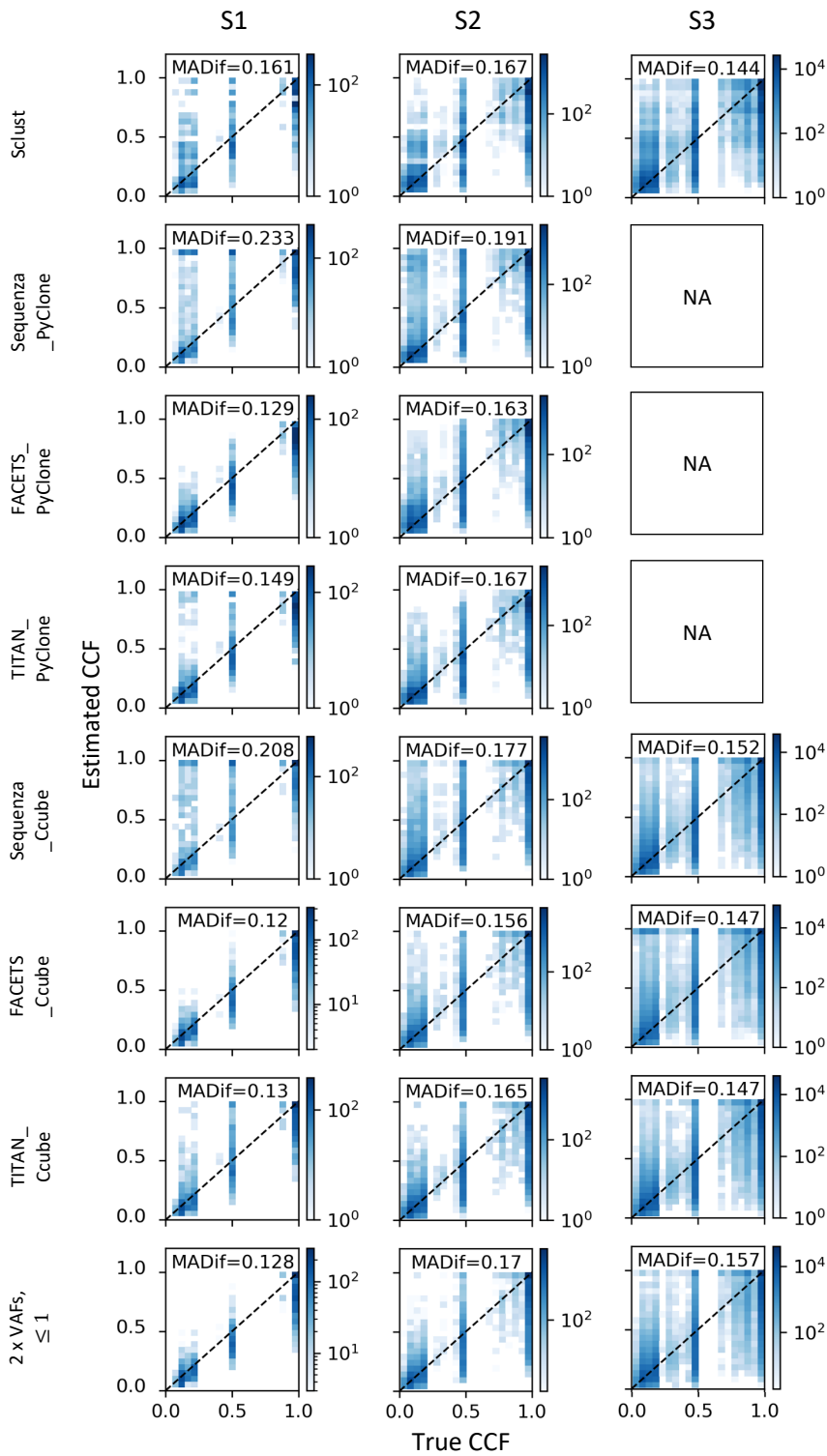


Figure 28. Estimated vs. true CCF values for all true positive Mutect2\_filtered\_ic called variants across samples with three different mutation frequencies, from seven different pipelines. The bottom panel shows estimated CCFs that are instead calculated by doubling VAFs and dividing by true purity, up to a maximum value of 1. Estimated CCFs in the Ccube pipelines are also limited to 1, as these can otherwise be higher due to modelling error.

Lastly, I investigated whether sequencing depth affected CCF accuracy. As expected, and likely due to the better resolution of VAFs, higher depth was found to substantially reduce MADif for all Ccube and PyClone pipelines, and to a lesser extent, for Sclust (Figure 29).

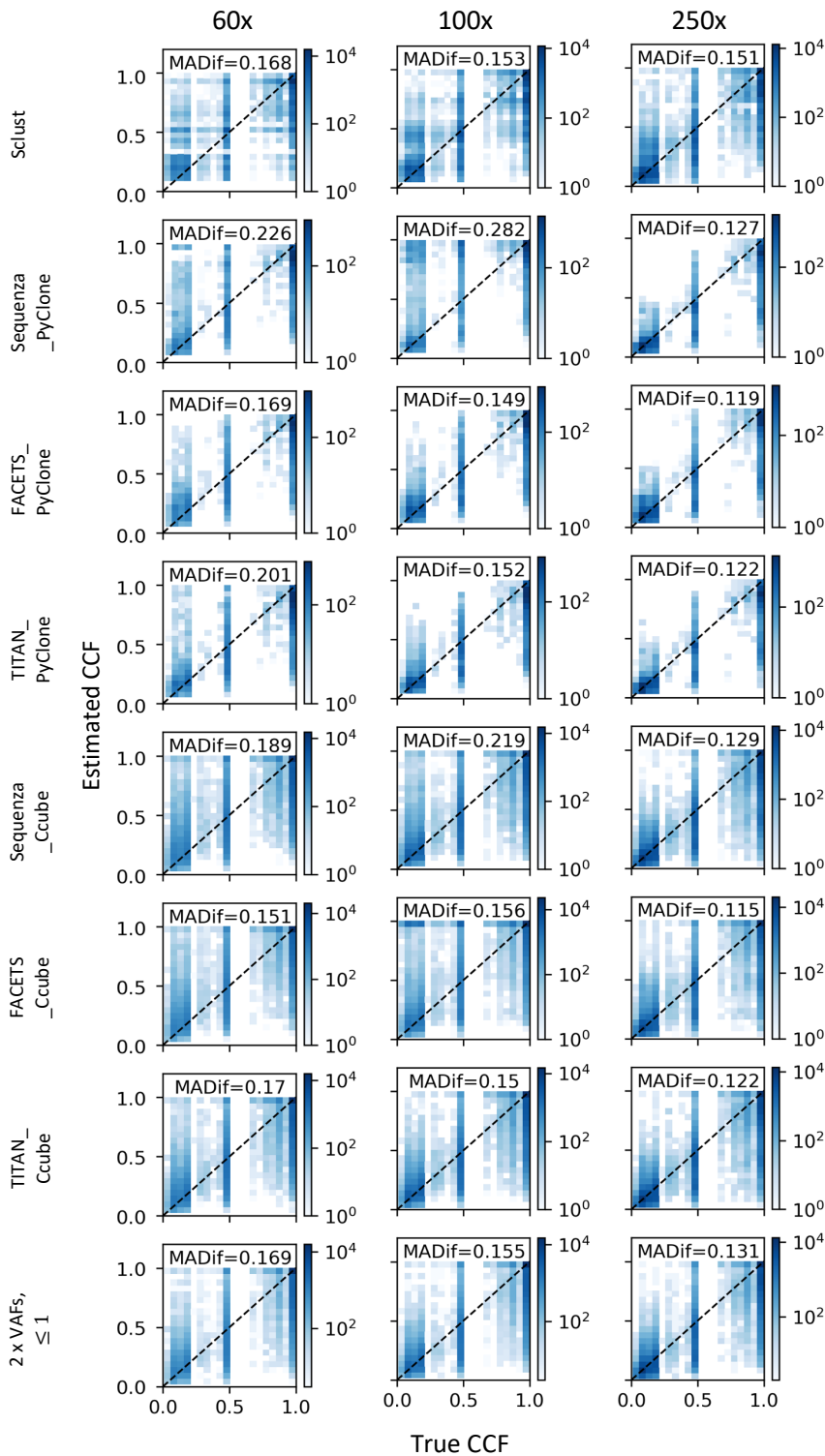


Figure 29. Estimated vs. true CCF values for all true positive Mutect2\_filtered\_ic called variants across all samples at 60x, 100x, or 250x sequencing depth, from seven different pipelines. The bottom panel shows estimated CCFs that are instead calculated by doubling VAFs and dividing by true purity, up to a maximum value of 1. Estimated CCFs in the Ccube pipelines are also limited to 1, as these can otherwise be higher due to modelling error.

Overall, MADif values were high for all pipelines, with CCF estimates being, on average, between 0.115 and 0.282 away from true values, and often not significantly better than just doubling VAFs and accounting for purity. The FACETS\_Ccube pipeline performed the best, reaching a MADif of 0.115 in the 250x samples, though this is only marginally better than when using doubled VAFs (MADif=0.131) (Figure 29). The FACETS\_Pyclone pipeline also achieved similar results, though PyClone was not able to produce results for

the samples with the highest number of point variants (Figure 28). TITAN\_Ccube, TITAN\_PyClone and ScIust performed similarly to each other on 100x samples but, whereas ScIust was not affected much by differences in sequencing depth, the MADif for the two TITAN pipelines increased and decreased significantly with the 60x and 250x samples, respectively, particularly with PyClone. These pipelines were all comparable to when using doubled VAFs. The two Sequenza pipelines achieved the poorest accuracies, particularly Sequenza\_PyClone with 60x and 100x samples, which achieved MADifs of 0.226 and 0.282, far worse than the MADifs when using doubled VAFs, of 0.169 and 0.155 (Figure 29).

## 3.3 Discussion

### 3.3.1 Overview

In this study, I benchmarked the performance of subclonal deconvolution pipelines for investigating ITH from single tumour and matched normal samples, using the novel simulation methods that I developed specifically for this purpose. This is an important goal, as subclonal deconvolution has not been extensively benchmarked previously, and the few studies that have include drawbacks. The “ICGC-TCGA DREAM Somatic Mutation Calling - Tumour Heterogeneity Challenge” study employs crowd sourcing to benchmark subclonal deconvolution pipelines, but this only includes data for WGS, not WES, simulates mutations by adding them to reads post alignment, and as yet has only published the in-house results, which do not include newer analysis methods (Salcedo *et al.*, 2020). Another important benchmarking study did so by generating mutation calls from simple distributions, and with simplifications such as subclonal variants must be carried by exactly 1 chromosome copy (Dentro *et al.*, 2020). To improve such benchmarking studies, we need to be able to model the extensive complexity of tumour genomes and generate realistic sequencing reads from them. In Chapter 2, I addressed this issue with the creation of HeteroGenesis and w-Wessim. These programs have allowed me to create test datasets that more closely represent the noise encountered in real analyses, allowing a better evaluation of the accuracy of subclonal deconvolution methods. This process also included benchmarking of somatic CNA calling, an analysis that has previously suffered from test datasets either lacking complexity or with incomplete known ground truths. The current study will therefore also be of use to researchers specifically interested in detecting somatic CNAs, particularly from highly heterogeneous tumours. While this study focusses only on WES analysis methods, the results of the subclonal deconvolution benchmarking will likely also be applicable to when using whole-genome sequencing data.

The key features of the WES datasets used in this study are i) the complexity of the underlying genomes and tumour subclonal architecture that the reads represent, and ii) the distributions of WES reads along the genome, which closely mirrors the noise and systemic bias of real WES data. These features are important for allowing a reliable representation of how CNA calling and subclonal deconvolution methods perform with real data. An aspect of the simulated datasets that doesn't so closely reflect real data, is the

higher than expected base quality score distribution, specifically for error bases. This likely resulted in the low precisions seen for variant callers. However, this is unlikely to substantially affect the benchmarking of CNA calling or subclonal deconvolution, for three reasons; 1) CNA calling is dependent only on read positions, which are unaffected by base quality scores as the alignment algorithm does not take these into account (Li, 2013). 2) The true sequenced bases in the simulated datasets do have a base quality score distribution that is an accurate representation of that in real data. Therefore, true positive variant calls are unlikely to be affected, and should be a good representation of the true variants called from real data. 3) The false positive variant calls, which increased in number as a result of the base quality scores for error bases being too high, was adjusted for by subsampling them to more realistic proportions. This resulted in call sets with precisions that covered those seen in previous studies. Overall, this means I can be confident that the inputs to the subclonal deconvolution methods are a good representation of those from real data.

### 3.3.2 Variant calling

The high base quality score distributions for error bases prevent accurate estimates of precisions for variant callers. Nonetheless, other observations can be drawn from the analysis. The filtering step for Mutect2 removes a very high number of true positives due to the 'clustered\_events' flag. Others have also noted removal of true positives with this flag, with some choosing to ignore the flag, as I did, or to combine a second caller to check the variants filtered by it (Letouzé *et al.*, 2017; Giroux Leprieur *et al.*, 2020; Tauriello *et al.*, 2018). It should be noted that in this study, it's possible that the flag's impact was exacerbated by higher rates of false positives being called. Strelka2 had a small increase in true positive rate compared to Mutect2\_filtered\_ic. It also had an approximately four-fold shorter run time per core, can be run multithreaded, and requires fewer steps. However, Strelka2 was also found to have a lower precision, compared to Mutect2. Furthermore, Mutect2 has additional features not utilised in this study, such as a read orientation artefacts filter and multi-sample mode, which will likely provide it with an advantage over other methods when used with FFPE data or multiple samples from the same tumour (Detering *et al.*, 2019). VarScan2 showed poor recall at lower VAFs, likely due to its hard filter for minimum VAF, at 0.05. It should be pointed out that, while Lancet did not perform as well in detecting combined SNVs and InDels as Mutect2 and Strelka2, its strengths are specifically focussed on InDels, and may perform comparatively better when assessed on these alone.

Ensemble callers (Rashid *et al.*, 2013; DePristo *et al.*, 2011; Wang *et al.*, 2020; Anzar *et al.*, 2019; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020), which combine the results of multiple SNV callers, are of potential use in generating a high confidence call set. I demonstrated in this study, through taking the union and intersect of the two best performing callers, that such methods risk increasing errors or decreasing the number true positives compared to when using single methods alone. This would be particularly problematic when poorer performing callers are included. Ensemble callers aim to overcome

these issues, either by using machine learning to create a classifier for filtering out false positives (DePristo *et al.*, 2011; Anzar *et al.*, 2019), or through a consensus approach (Wang *et al.*, 2020; Rashid *et al.*, 2013; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). However, this adds to an already lengthy process, and nonetheless, the best results will be achieved when only including callers that have been shown individually to perform well in benchmarking studies.

### 3.3.3 CNA calling and subclonal deconvolution

To investigate the performance of subclonal deconvolution methods, I assessed their ability to accurately estimate variant CCFs, from the Mutect2\_filtered\_ic call sets. Whilst the choice of subclonal deconvolution method had an effect on this, the results in this study suggest the choice of CNA caller is likely a larger factor, with the MADif metric varying more between CNA callers than between PyClone vs. Ccube with the same CNA caller. FACETS was found to be more sensitive in detecting CNAs than the other methods, resulting in improved accuracy for CCF estimation. This is likely due to its use of pseudo-SNPs during segmentation, which allows it to pick up on changes in regions not necessarily covering high numbers of SNPs. Additionally, thanks to a convenient wrap-around script (cnv\_facets (Beraldi)) from [https://github.com/dariober/cnv\\_facets](https://github.com/dariober/cnv_facets), FACETS was also the easiest of the tested CNA calling methods to install and run, with TITAN being the most complex. Ccube was shown to be the most accurate method at estimating CCFs, with PyClone performing the worst. PyClone was run using the clonal copy numbers from TITAN and FACETS, as it is unable to accept estimates of the cellular frequencies for subclonal CNAs, and instead treats all inputted copy numbers as the overall values across all subclones in the tumour. It's possible that PyClone's performance may improve by instead providing it with all CNAs estimated to be in greater than 50% of cells, or recalculating subclonal copy numbers to reflect the overall sample (assuming an otherwise diploid genome in the remaining cells). Nonetheless, Ccube also outperformed PyClone when using identical CNA calls from Sequenza, which does not provide subclonal CNAs. It should also be noted that parameters exist for some subclonal deconvolution methods that may improve their performance. For example, PyClone has parameters for increasing iterations or reducing runtime, which may be useful when only using a single tumour sample or when hundreds of variants are detected. However it is not practical to benchmark multiple parameter setups.

Recently, a large study by Dentre *et al.* developed a set of consensus approaches to performing both WGS CNA calling and subclonal deconvolution, and applied them to a set of 2,658 tumours (Dentre *et al.*, 2020). These approaches combined estimates from 6 CNA callers and 11 subclonal deconvolution methods. Individually, the CNA callers, which were for WGS and not WES tested in the current study, showed a good confidence consensus on 93% of tumours, whereas the subclonal deconvolution methods showed highly varied results. When benchmarking these, Dentre *et al.* found that the consensus approaches performed comparably to the best performing individual method, Ccube. Due to the way in which this benchmarking

was achieved, through using mutations generated from distributions and with assumptions that limit the complexity of the underlying ground truth, it was not known whether Ccube's higher accuracy would hold when used with more complex datasets in the current study. However, I found that Ccube did indeed outperform PyClone and Sclust, two methods also included by Dentre *et al.*, although none showed a high level of accuracy. The current study therefore confirms some of the results from Dentre *et al.* although, due to the more complex and realistic datasets, it is likely better able to assess the limitations of subclonal deconvolution methods.

The minimum mean difference of estimated CCFs from true values was 0.115, for FACETS\_Ccube with 250x samples, only marginally better than just doubling purity adjusted VAFs, with a MADif of 0.131. This is before considering the fact that only a small portion of the overall tumour is likely to be represented within a single sample (Bhandari *et al.*, 2018; Siegmund and Shibata, 2016; Watkins and Schwarz, 2018; Sun *et al.*, 2017; Chkhaidze *et al.*, 2019), and therefore, results from subclonal deconvolution of single samples should be interpreted cautiously. For example, a variant with a low VAF in a single primary tumour sample and high VAF in a single recurrent sample, does not necessarily imply that the cells containing the variant expanded in frequency over time. Instead, this could reflect the substantial inaccuracies in estimating CCFs from VAFs in one or both of the samples, or may reflect different proportions of variant containing cells being captured in the sampling from each tumour. Multi-sampling approaches are able to address these issues by looking across multiple regions of a tumour, thereby reducing the effects of sampling bias as well as likely improving CCF estimates. When such datasets are not available, looking for recurring patterns across large cohorts of patients may instead help to overcome the challenges of using single samples.

### 3.3.4 Future directions and conclusions

The benchmarking of subclonal deconvolution methods was carried out using inputs with subsampled false positives variant calls. This largely overcomes the issues resulting from the higher distribution for error base quality scores. Nonetheless, this analysis would benefit from adjusting the test datasets to bring the scores more in-line with those of real data, to increase confidence in the results. This could be achieved through modifications to the *in silico* sequencing method, as discussed in Chapter 2.

An important feature of tumour clones is that they continue to develop new mutations as the cells divide and grow. This results in a neutral tail of low VAFs, where the cumulative number of variants has a linear relationship with the inverse of their allele frequency (Williams *et al.*, 2016). The neutral tail is distinct from the major peak of VAFs present in every cell of the clone, and with most subclonal deconvolution methods, it may wrongly be identified as primary peaks from additional clones. A novel subclonal deconvolution method, MOBSTER, instead applies evolutionary modelling to identify and remove neutral tails from the data, allowing true clones to be determined more accurately (Caravagna *et al.*, 2020). A drawback of the MOBSTER method is that the required input is in the form of VAFs that have been normalised for purity,

and it does not take CNAs into account. It is possible to instead provide CCFs divided by two, meaning that other methods, such as those included in this study, could first be used to adjust for CNAs, although the authors stress the importance of accurate CCF input data. Neutral tails are not modelled in the datasets created in this study. In the future it may be possible to create genomes for each cell in a tumour and to *in silico* sequence each one using HeteroGenesis and w-Wessim, but this is currently not feasible given the high computational resources required. A more practical alternative for achieving neutral tails in the dataset, would be to spike variants into the reads from each clone, prior to mixing, in a way that models continued neutral evolution. Nonetheless, even without neutral tails, the datasets in this study allow for a more realistic assessment and comparison of subclonal deconvolution methods than previous studies.

Our group plans to expand the benchmarking in this study to test multi-sample subclonal deconvolution methods, including those that reconstruct phylogenetic trees of the relationships between subclones, such as Canopy (Jiang *et al.*, 2016), SPRUCE (El-Kebir *et al.*, 2016) and PhylogicNDT (Leshchiner *et al.*, 2018). This is a much needed analysis, and would fully utilise the unique capabilities of HeteroGenesis and its freqcalc module. The program supports the creation of multi-sample datasets by providing copy number and point variant profiles to reflect different bulk samples from the same tumour, each containing varying proportions of clones. Bulk sequencing datasets can then be created to reflect these profiles by merging reads from each clone in the correct proportions.

Regardless of the computational tools used, somatic variant and CNA calling, and therefore subclonal deconvolution, is ultimately limited by current sequencing technology and depth (Shi *et al.*, 2018). In the future it is likely that single cell analyses will overcome many of the issues with investigating ITH, allowing for more accurate estimation of variant frequencies and subclonal deconvolution of tumours. Currently these methods are expensive, low throughput (though increasing), and produce a substantial amount of noise due to drop-outs (Gawad *et al.*, 2016). Furthermore, archived material does not easily lend itself to single-cell analysis due to the damaging effects of freezing or FFPE (Guillaumet-Adkins *et al.*, 2017; Martelotto *et al.*, 2017). The benchmarking of subclonal deconvolution pipelines, carried out in this study, is therefore important to inform researchers on the most suitable, as well as the limitations of, available methods.

## 3.4 Methods

### 3.4.1 Creating datasets for use in benchmarking

HeteroGenesis was used to simulate nine sets of tumour genomes from three parameter sets, with values listed in Table 3. The hg38 human reference genome was used as the baseline sequence for simulating genomes from. Known SNPs and InDels were taken from dbsnp\_146.hg38.vcf. The 22 major autosomal chromosomes were included in the simulation, but the sex chromosomes were excluded to avoid issues



from using real WES reads from a male during sequencing with w-Wessim. HeteroGenesis's `freqcalc` command was used to calculate overall bulk mutation profiles from individual clone profiles, for samples described in Figure 21. Genome sequences from HeteroGenesis were *in silico* sequenced with w-Wessim, as described in Chapter 2. Fragment lengths were set at a mean of 170b  $\pm$ 35 sd, which reflects those in real WES data for a set of mixed formalin-fixed paraffin embedded (FFPE)/fresh frozen GBM samples and matched blood samples (Droop *et al.*, 2018). Read lengths were taken from the distribution in the error model to account for the trimmed low quality bases and clipped reads in the training data set. The simulated reads were then aligned to the hg38 reference genome with BWA-MEM (Li, 2013).

Per sample BAM files were created from the per clone files, as indicated in Figure 21, through downsampling and merging. Downsampling of BAM files was achieved using both SAMtools (Li *et al.*, 2009) and a custom script (<https://github.com/GeorgetteTanner/randomsplitbam>) that randomly splits reads in BAM files into two, with user defined proportions in each and with paired reads together. This ensures no overlap between the outputs, so as to avoid normal samples and corresponding tumour samples with normal contamination, containing identical reads. Such a scenario is unavoidable using SAMtools alone.

### 3.4.2 Variant calling

Commands and methods used to run the variant calling methods are described in the appendix.

Variants and background base positions were limited to those with  $\geq 8$  reads in both the tumour and normal samples. This was achieved using SAMtools `mpileup` (Li *et al.*, 2009) in combination with a custom script for the ground truth variants, and either via variant caller parameters (VarScan2 and Lancet) or custom filtering (Mutect2 and Strlka2) for the call sets.

A custom Python3 script was used to analyse the call sets, including the Scikit-learn package (Pedregosa *et al.*, 2011) for creating the ROCs.

### 3.4.3 CNA calling

Sequenza (bitbucket commit 059325a, sequenza-utils v3.0.0) was run with the provided pype pipeline. All default settings were used, including a binning size of 50, with a `gc_wiggle` file also with bin size of 50.

TITAN was run using the provided snakemake workflow with default parameters, with the following exceptions: Parameters were set for use with the hg38 human reference genome. A BED file of regions covered by the S04380219 Agilent SureSelect All Exon v5+UTR probes was provided to define target regions (`ichorCNA_exons`, `TitanCNA_chrs`). Chromosome X was removed from the list of chromosomes to include (`ichorCNA_chrs`). Initial normal contamination values was set to "`c(0,0.1,0.2,0.3,0.4,0.5,0.6)`"

(ichorCNA\_normal). The maximum number of clusters was set to 8 (TitanCNA\_maxNumClonalClusters). A statistical parameter was adjusted to accommodate WES datasets (TitanCNA\_alphaK: 2500).

FACETS (v0.5.14) was run with default parameters using the wrap around script, `cnv_facets` (v0.15.0) (Beraldi) from [https://github.com/dariober/cnv\\_facets](https://github.com/dariober/cnv_facets). A bed file specifying target exon regions was provided.

Sclust (v1.1) was run under default parameters and as described in (Cun *et al.*, 2018), with the exception of four samples (s1r3-A\_250x\_lim0.1, s2r1-A\_250x\_lim0.1, s1r2-B\_100x\_lim0.5, s2r1-A\_100x\_lim0.5) for which a slightly larger smoothing parameter (-lambda 1e-6.9) was required for completion of the cluster module, as recommended by the authors. “-part 1” was added to the bamprocess module command to indicate whole exome sequencing.

Heatmaps were generated using CNVkit (Talevich *et al.*, 2016), customised in order to create the required plots. Ground truth somatic copy number profiles were calculated by taking the ratios of somatic copy numbers to germline copy numbers, thereby normalising for the effect of germline CNVs.

### 3.4.4 Subclonal deconvolution

Purity estimates for all CCF estimation methods were taken from the same CNA caller used to provide copy number inputs. All were run with the assumption of an entirely heterozygous diploid germline sample.

PyClone (v0.13.1) was run under default parameters, through the ‘run\_analysis\_pipeline’ command to run the full workflow. As this method only considers total average copy number estimates, when used with outputs from FACETS and TITAN, only clonal CNAs were incorporated.

Ccube (v1.0) was run with the command, ‘RunCcubePipeline(ssm = data, numOfClusterPool = 1:10, numOfRepeat = 1, runAnalysis = T, runQC = T, multiCore = F)’, and providing either clonal copy number estimates from Sequenza or clonal/subclonal estimates from TITAN/FACETS.

Sclust (v1.1) – See section 3.4.3.1.

A custom Python3 script was used to analyse results.

## 3.5 Appendix

### 3.5.1 Variant calling commands

The following commands were used to run the somatic variant calling methods.

Lancet (v1.1):

- `lancet --min-strand-bias 1 --min-map-qual 15 --min-phred-fisher 0 --min-coverage-tumor 8 --min-coverage-normal 8 --min-alt-count-tumor 3 --ref hg38.fasta --reg {CHRO} --num-threads 24 --tumor {SAMPLE}.bam --normal normal_{SAMPLE}.bam > {SAMPLE}_{CHRO}_lancet_1.vcf`
- `lancet --min-strand-bias 0 --min-phred-fisher 0 --min-coverage-tumor 8 --min-coverage-normal 8 --min-vaf-tumor 0.01 --min-alt-count-tumor 2 --ref hg38.fasta --reg {CHRO} --num-threads 24 --tumor {SAMPLE}.bam --normal normal_{SAMPLE}.bam > {SAMPLE}_{CHRO}_lancet_2.vcf`
- `cat {SAMPLE}_chr*_lancet_1.vcf > {SAMPLE}_lancet_1.vcf`
- `cat {SAMPLE}_chr*_lancet_2.vcf > {SAMPLE}_lancet_2.vcf`

Mutect2 (GATK v4.1.2.0):

- `gatk Mutect2 --native-pair-hmm-threads 5 -R hg38.fasta -L {CHRO} -I {SAMPLE}.bam -tumor {SAMPLE} -I normal_{SAMPLE}.bam -normal normal_{SAMPLE} --germline-resource af-only-gnomad.hg38.vcf.gz -O {SAMPLE}_mutect2_{CHRO}.vcf.gz`
- `gunzip {SAMPLE}_mutect2_chr*.vcf.gz ; cat {SAMPLE}_mutect2_chr*.vcf > {SAMPLE}_mutect2.vcf ; bgzip {SAMPLE}_mutect2.vcf`
- `gatk MergeMutectStats -stats {SAMPLE}_mutect2_chr1.vcf.gz.stats -stats ... -stats {SAMPLE}_mutect2_chr22.vcf.gz.stats -O {SAMPLE}_mutect2.vcf.gz.stats`
- `gatk IndexFeatureFile -I {SAMPLE}_mutect2.vcf.gz`
- `gatk FilterMutectCalls -V {SAMPLE}_mutect2.vcf.gz -O {SAMPLE}_mutect2_filtered.vcf.gz -R hg38.fasta`

Strelka2 (v2.9.10):

- `configureStrelkaSomaticWorkflow.py --normalBam normal_{SAMPLE}.bam --tumorBam {SAMPLE}.bam --referenceFasta hg38.fasta --runDir ./ --callRegions callable.bed.gz --exome (callable.bed.gz contains the whole lengths of chromosomes 1-22, X, Y, and M) (configureStrelkaSomaticWorkflow.py.ini was unchanged)`
- `runWorkflow.py -m local -j 24`

VarScan2 (v2.4.4, samtools mpileup v1.9, htlib v1.9, bam-readcount v0.8.0):

- `samtools view -b -h -q 15 -o {SAMPLE}_q15.bam {SAMPLE}.bam ; samtools view -b -h -q 15 -o normal_{SAMPLE}_q15.bam normal_{SAMPLE}.bam`
- `samtools mpileup -B -f hg38.fasta normal_{SAMPLE}_q15.bam {SAMPLE}_q15.bam > {SAMPLE}_dual.mpileup`

- `java -jar VarScan.v2.4.4.jar somatic {SAMPLE}_dual.mpileup {SAMPLE}_varscan2 --mpileup 1 --p-value 0.10 --strand-filter 0 --min-coverage-tumor 8 --min-coverage-normal 8 --min-var-freq 0.01 --somatic-p-value 1.0`
- `cat {SAMPLE}_varscan2.snp {SAMPLE}_varscan2.indel > {SAMPLE}_varscan2.txt`
- `awk '{print $1,$2-1,$2+1}' {SAMPLE}_varscan2.txt > {SAMPLE}_varscan2_pos.txt`
- `bam-readcount -w1 -f hg38.fasta -l {SAMPLE}_varscan2_pos.txt {SAMPLE}_varscan2.bam > {SAMPLE}_varscan2.txt.readcounts`
- `java -jar VarScan.v2.4.4.jar fpcfilter {SAMPLE}_varscan2.txt {SAMPLE}_varscan2.txt.readcounts --dream3-settings --keep-failures --output-file {SAMPLE}_varscan2_filtered.txt`

A description of the parameters and filters used for variant calling are provided in Appendix Table 1.

Appendix Table 1. Parameters and filters for running variant calling. Default values are stated where non-default values are used. Thresholds were set to a minimum stringency where appropriate, to output every considered position to allow creation of full ROC curves. All other parameters that are not indicated in the table were left as defaults. \*VarScan2’s P-value threshold to call a heterozygote was lowered from default, as recommended by the VarScan2 author’s as a result of fine-tuning for the TCGA-ICGC DREAM-3 SNV Challenge (Ewing *et al.*, 2015) (discussed in Chapter 2). Other additional fine tuning parameters for minimum VAF and coverage were not followed, so as to allow calling of lower VAF variants down to 0.05 (which were not included in the DREAM challenge training dataset) and to standardise minimum coverage between pipelines. \*\*VarScan2’s strand bias filter is indicated as being off by default, but it appears that the default setting is actually on. Therefore, as the authors recommend it being off as a result of their DREAM challenge fine-tuning, a parameter was included to specifically turn it off. \*\*\*Strelka2’s configuration file states higher aggregate variant score values for filtering, but it is apparent from the outputs that those values are ignored and the scores stated here are what is actually used.

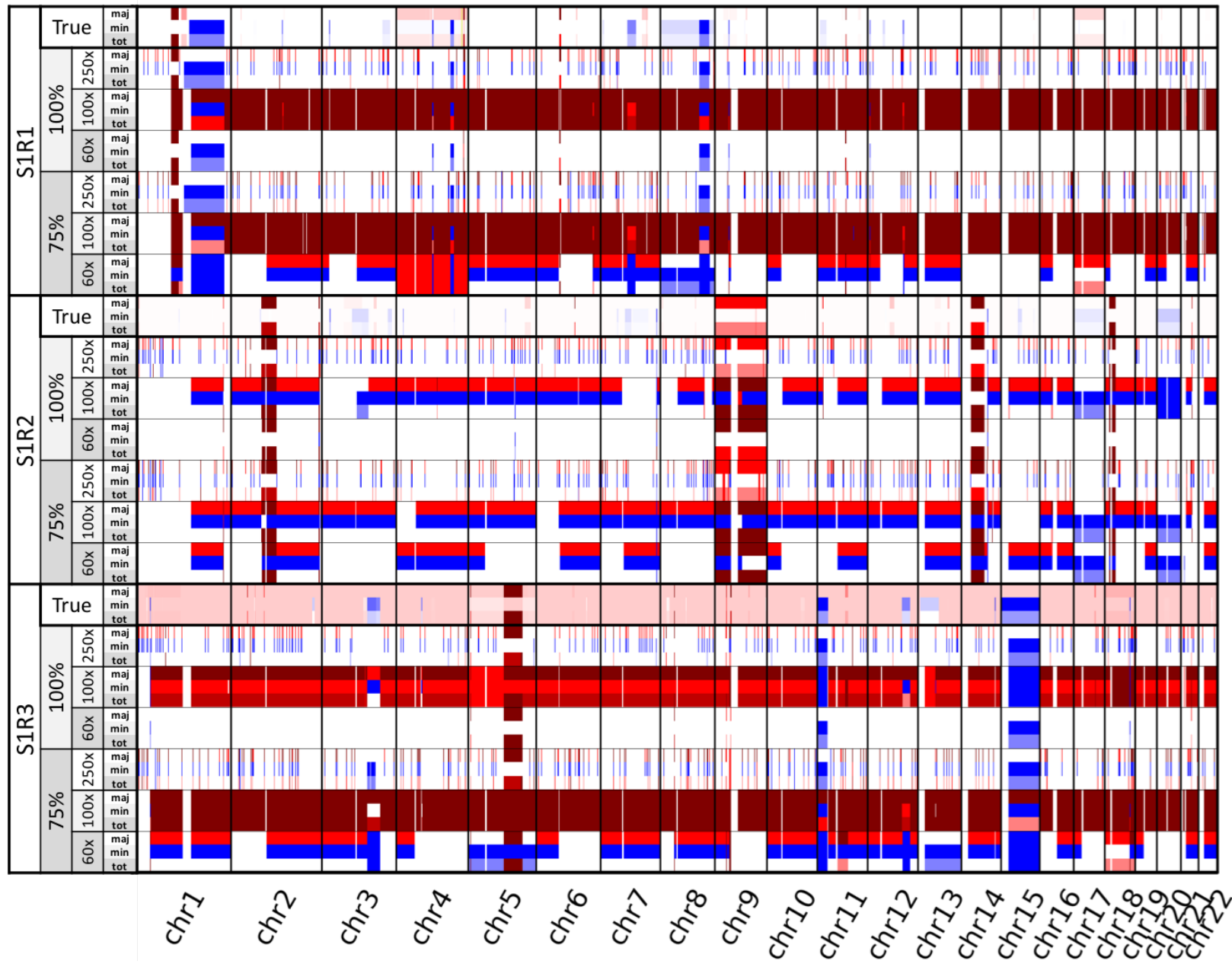
SNV calling pipeline	VarScan2	VarScan2 + ffilter	Mutect2	Mutect2 + FilterMutectCalls	Lancet_1	Lancet_2	Strelka2
Description	VarScan2 (+ samtools mpileup)	VarScan2 with ffilter. ffilter parameters were set based on the author’s fine-tuning for the TCGA-ICGC DREAM-3 SNV Challenge, using the --dream3-settings flag. (+ bam-readcount)	Mutect2 with default parameters	Mutect2 with FilterMutectCalls under default parameters.	Lancet with minor adjustments to parameters.	Lancet with less stringent parameters.	Strelka2 with default parameters.
Versions and references	samtools mpileup v1.9, htlib v1.9 (Li <i>et al.</i> , 2009), VarScan2 v2.4.4 (Koboldt <i>et al.</i> , 2012)	bam-readcount v0.8.0 ( <a href="https://github.com/genome/bam-readcount">https://github.com/genome/bam-readcount</a> )	GATK v4.1.2.0 (Benjamin <i>et al.</i> , 2019)	-	Lancet v1.1 (Narzisi <i>et al.</i> , 2018)	-	Strelka2 v2.9.10 (Kim <i>et al.</i> , 2018)
Threshold	Somatic P-value=1 (default=0.05) (dream3=0.05)	Somatic P-value=1 (default=0.05) (dream3=0.05)	LOD=3	LOD=3	Fisher’s exact test score=0 (default=5)	Fisher’s exact test score=0 (default=5)	Aggregate variant score***: SNVs=7 InDels=6

	Heterozygous P-value = 0.10 (default=0.99)* (dream3=0.10)	Heterozygous P-value = 0.10 (default=0.99)* (dream3=0.10)					
Minimum coverage in tumour/normal	8/8 (default=6/8) (dream=3/3)	8/8 (default=6/8) (dream=3/3)	NA - Manually filtered for 8/8 after calling	NA - Manually filtered for 8/8 after calling	8/8 (default=4/10)	8/8 (default=4/10)	NA - Manually filtered for 8/8 after calling
Minimum read mapping quality	15 (from previous filtering of reads with samtools, as recommended by the VarScan2 authors.)	30 alt/20 ref (default=15/15) (dream3=30/20)	Median≥50	Median alt≥30 (with an exception for longer indels)	15	15	20
Minimum base quality	13 (from mpileup default parameters)	30 alt/15 ref (default=15/15) (dream3=30/15)	10 alt/10 ref	10 alt/10 ref, Median alt≥20	17	17	NA
Minimum VAF	0.01 (default=0.20) (dream3=0.08)	0.05 (default=0.20) (dream3=0.08)	0	0	0.04	0.01 (default=0.04)	0
Minimum alt supporting reads	NA	3, 1 in low coverage regions) (default=4,2) (dream3=3, 1)	NA	0	3	2 (default=3)	NA
Strand bias filter	Off** (default=on - removes variants with >90% strand bias)	Off (default=minimum of 1% alt reads from each strand) (dream3=off)	NA	On – Performs strand artefact modelling	On – requires ≥1 alt reads on both strands	Off (default=on)	Off

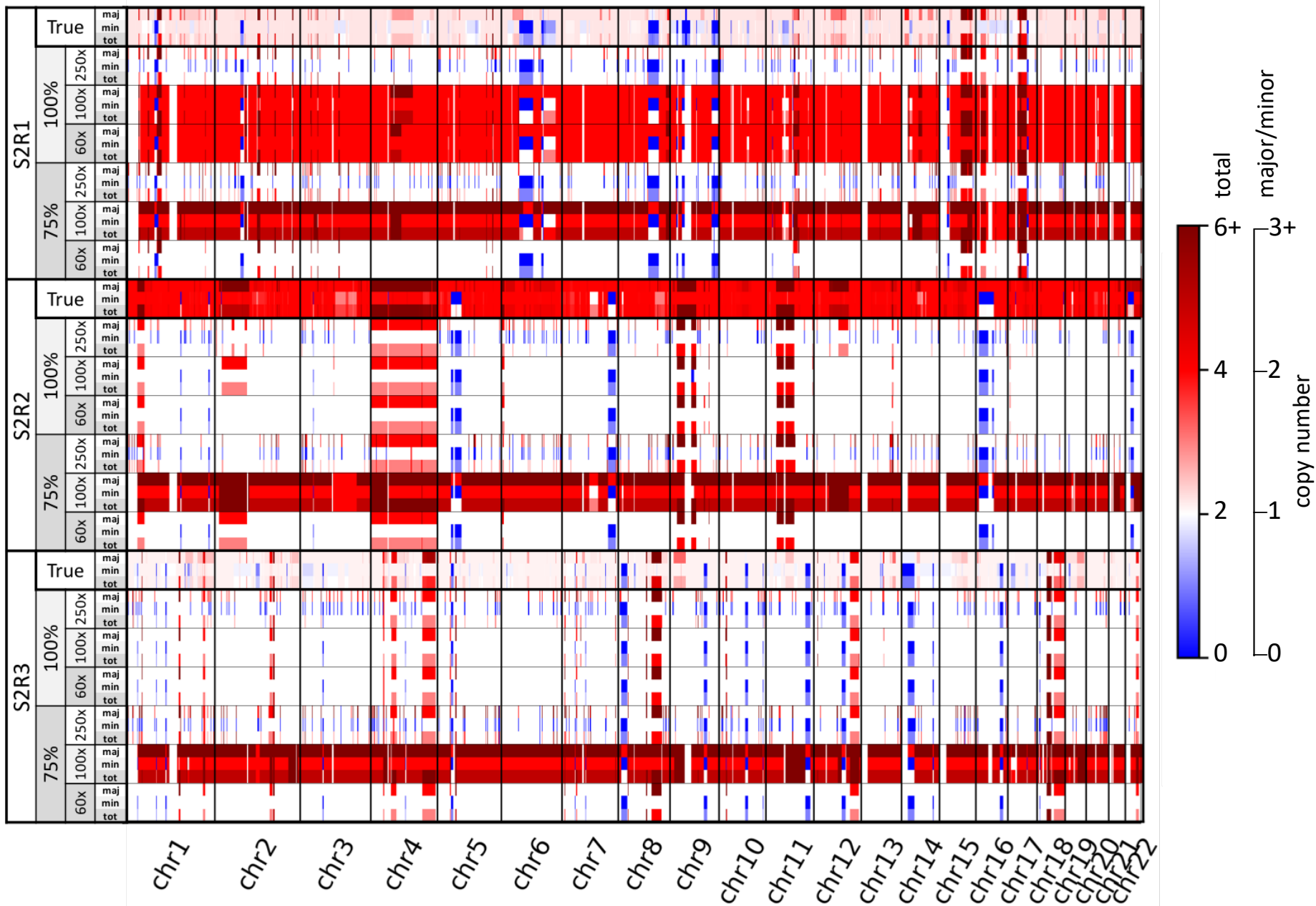
### 3.5.2 CNA heatmaps

Appendix Figure 1. (Below) Heatmaps of true and predicted overall copy numbers. FACETS, TITAN, and Sclust copy numbers are multiplied by the predicted cellular fractions containing them, thereby making them directly comparable to overall true copy number values.

Sequenza

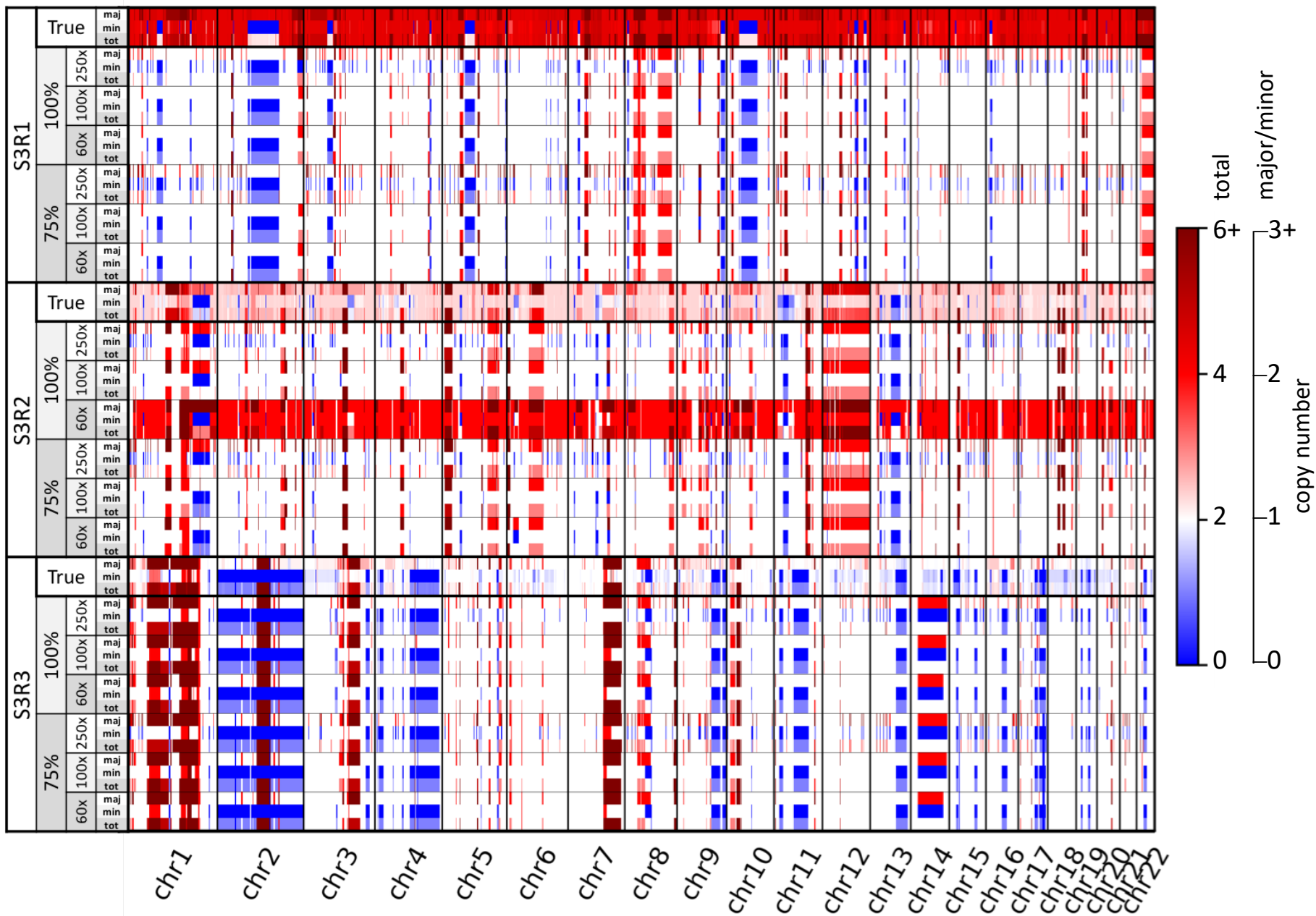


Sequenza

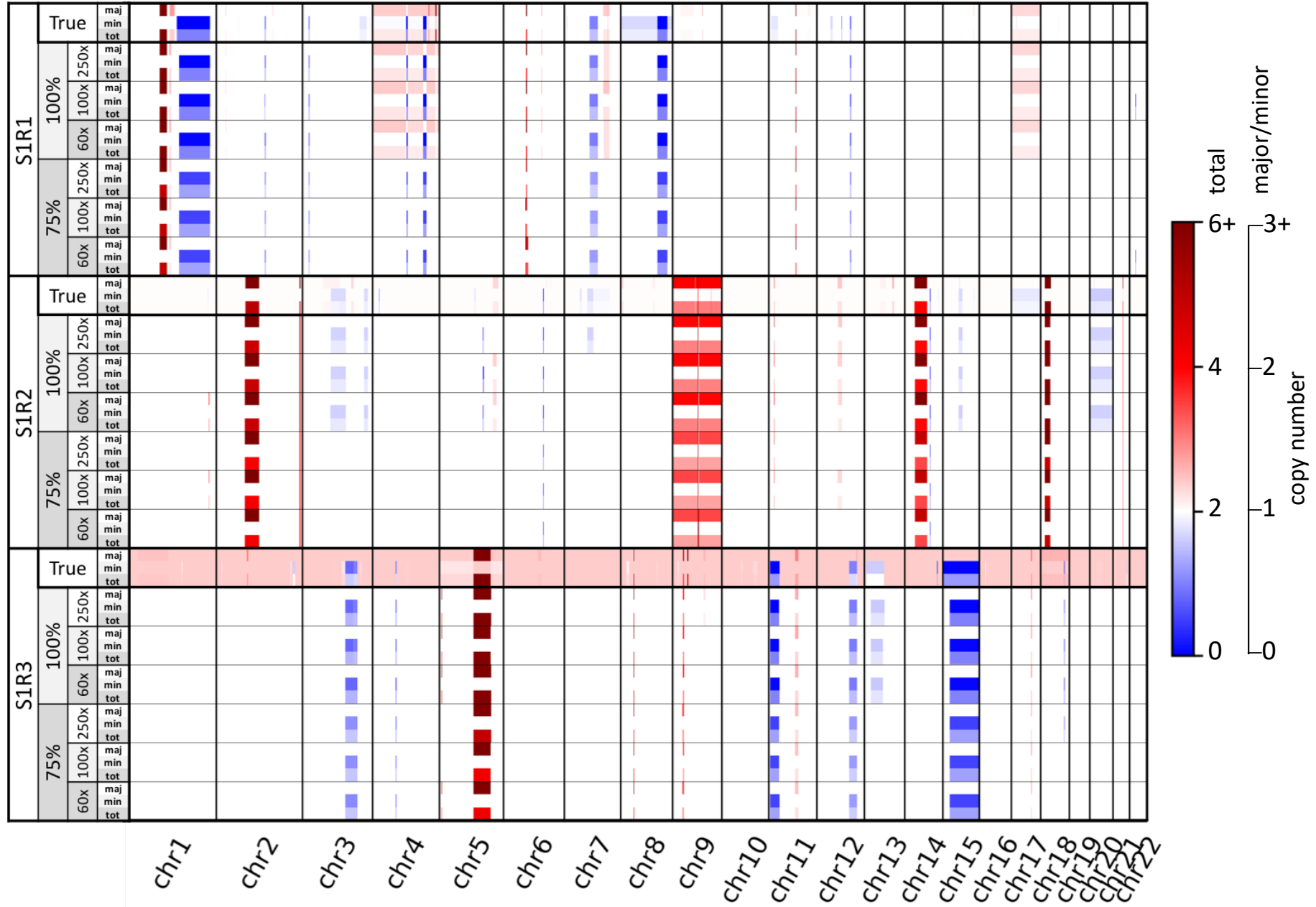




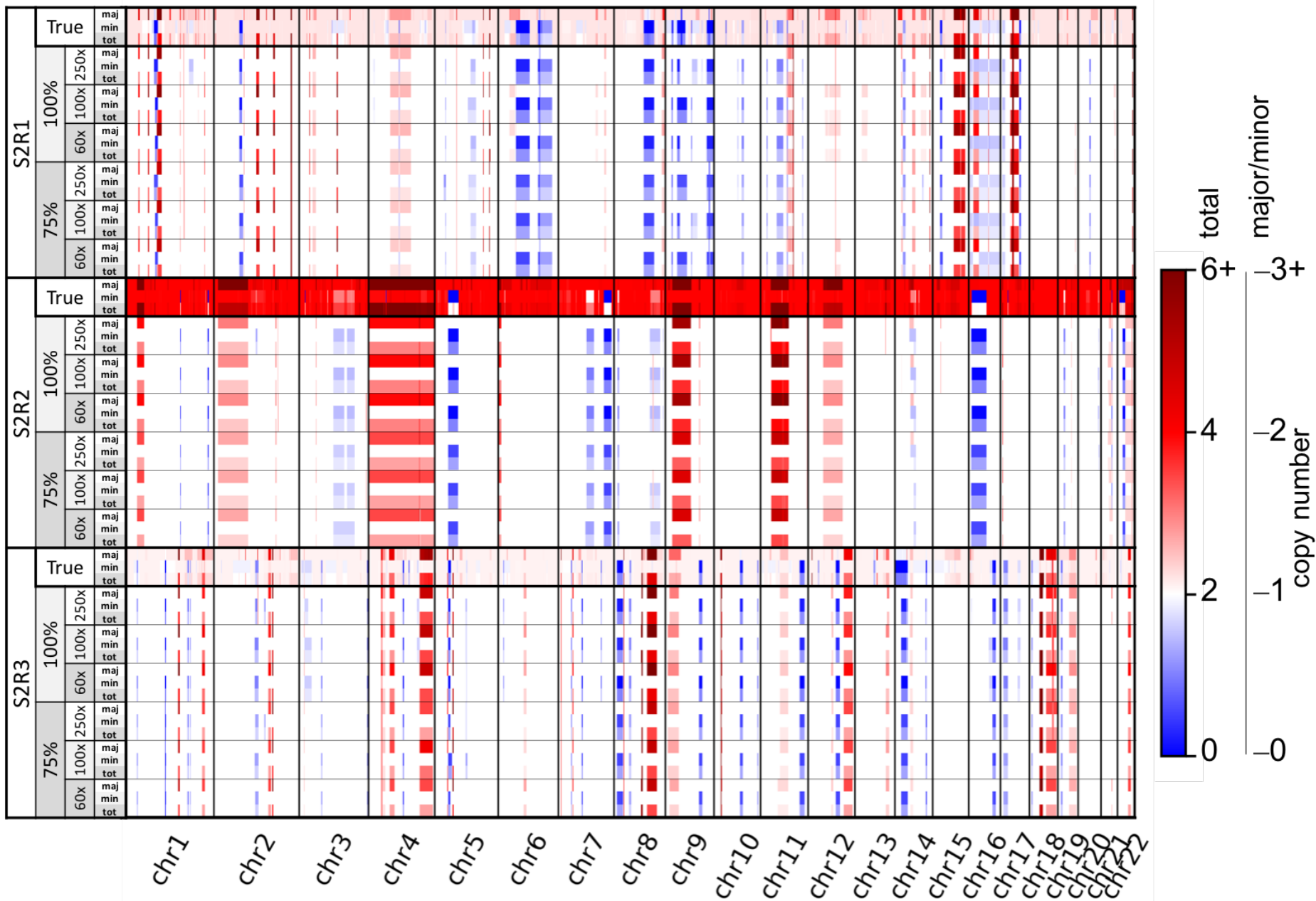
Sequenza



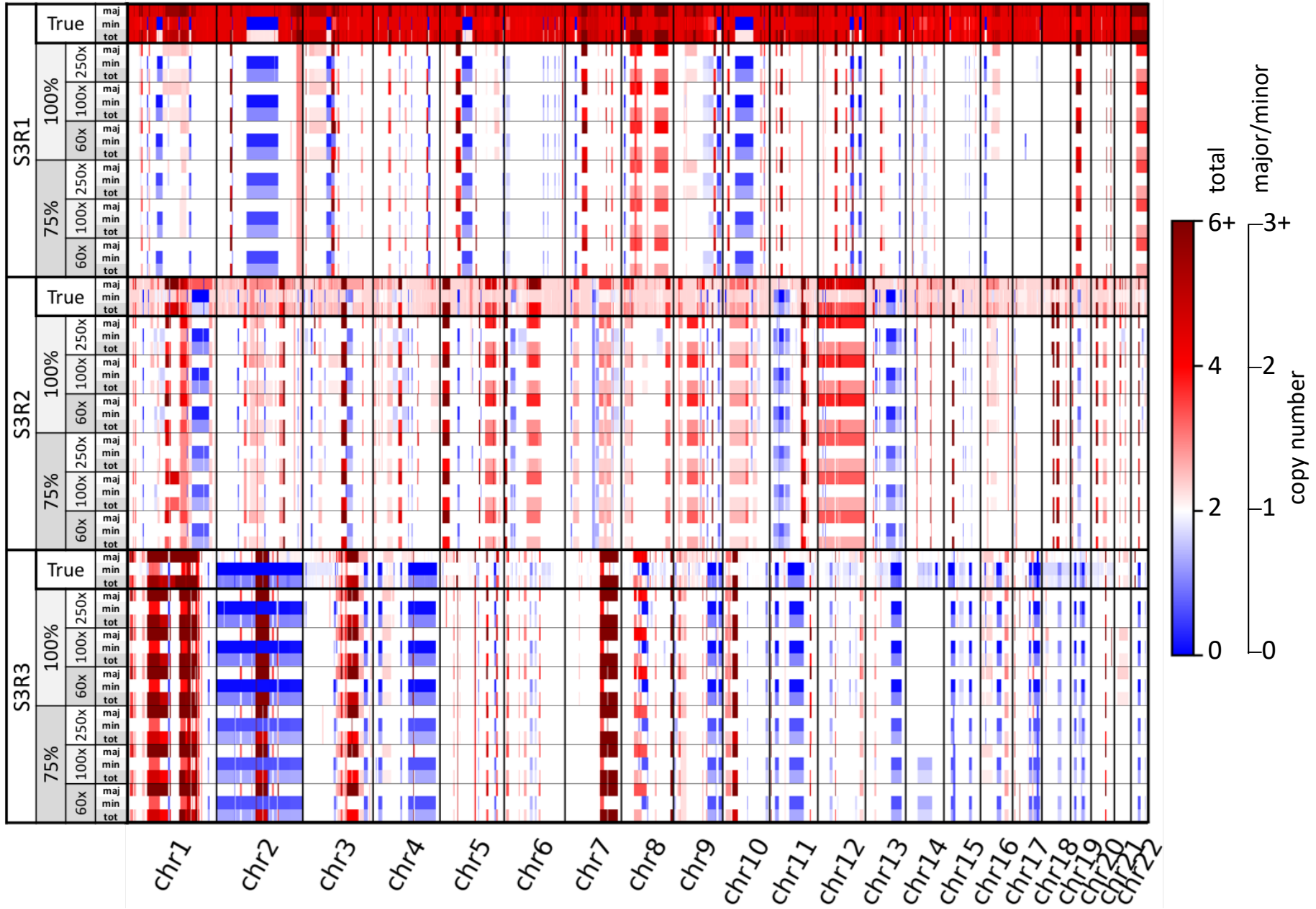
FACETS



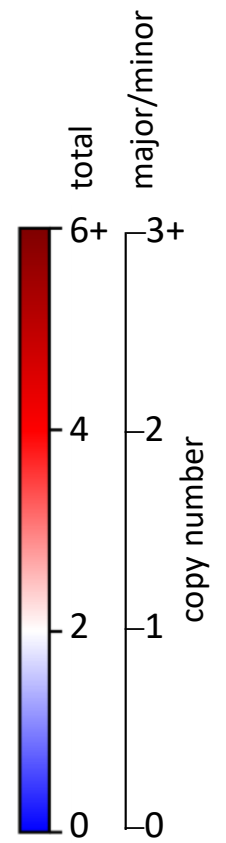
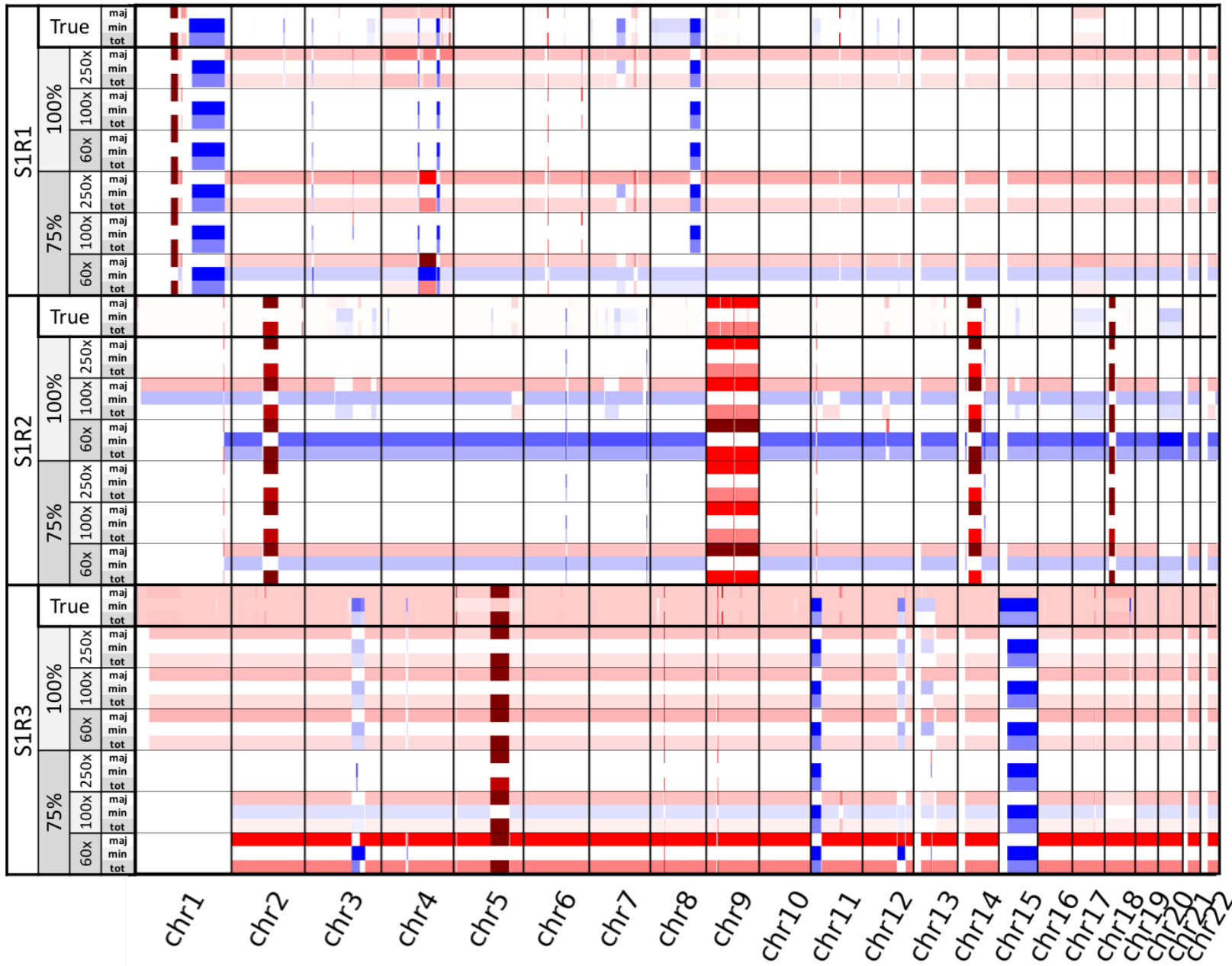
FACETS



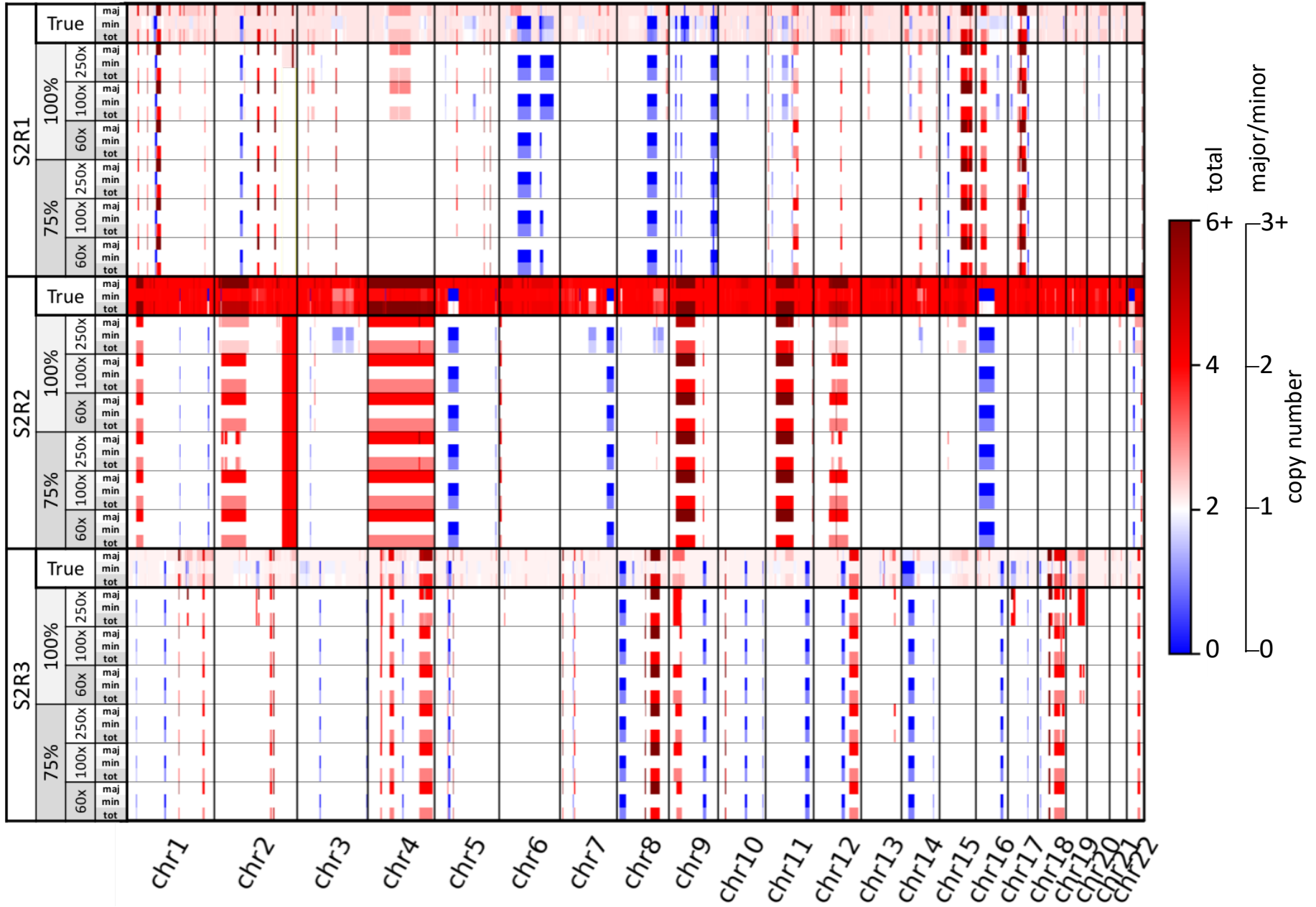
FACETS



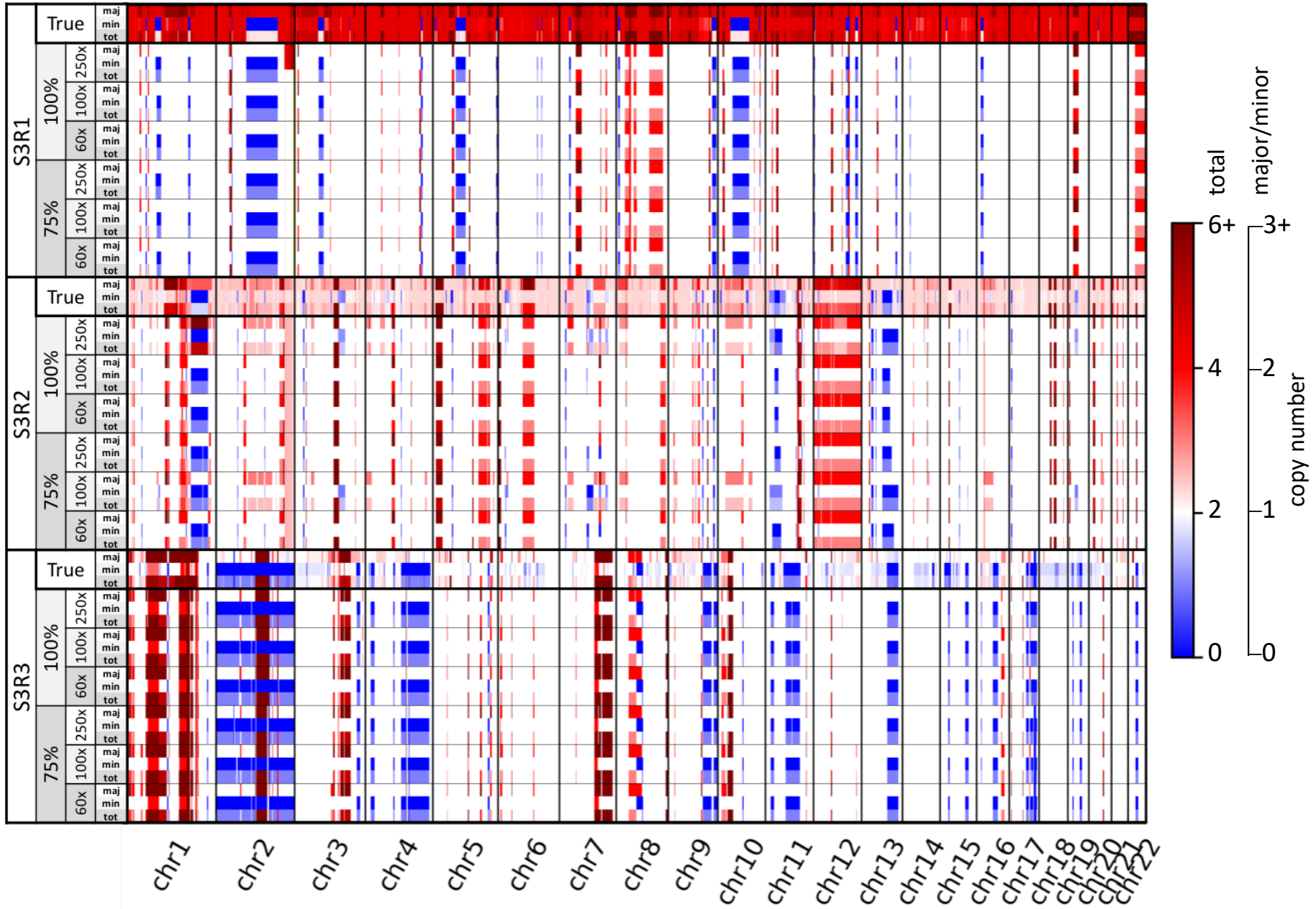
TITAN



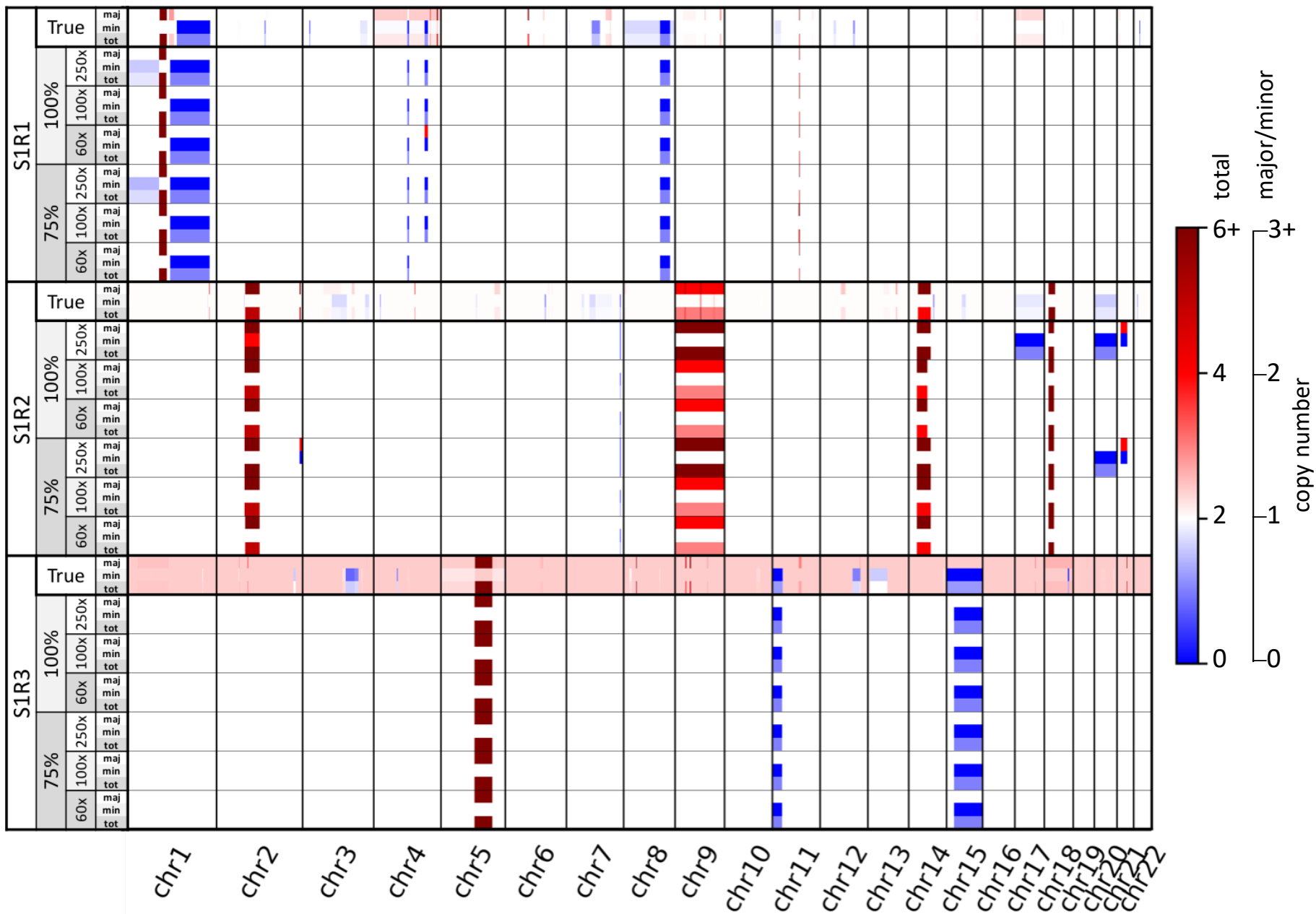
TITAN



TITAN

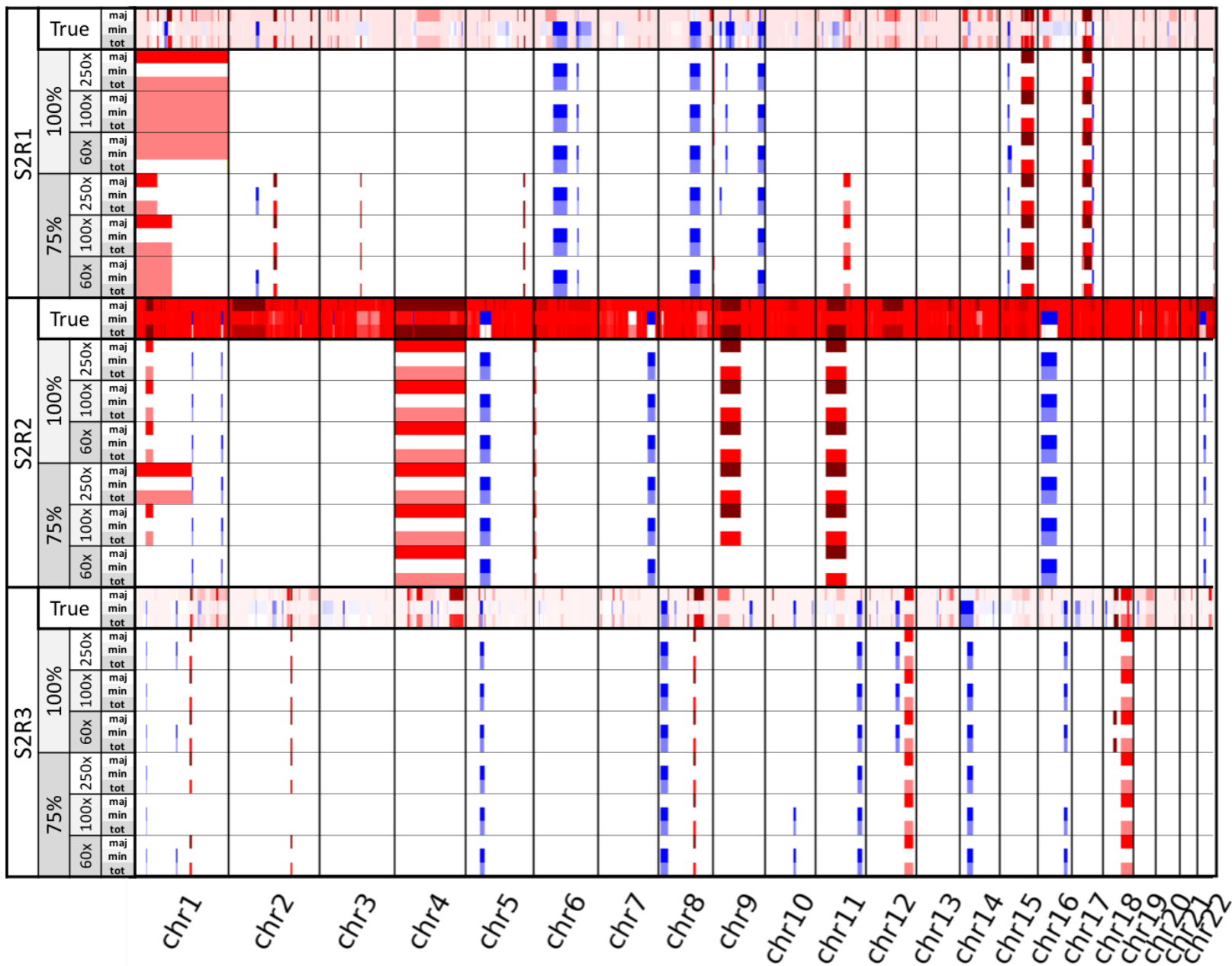


Sclust

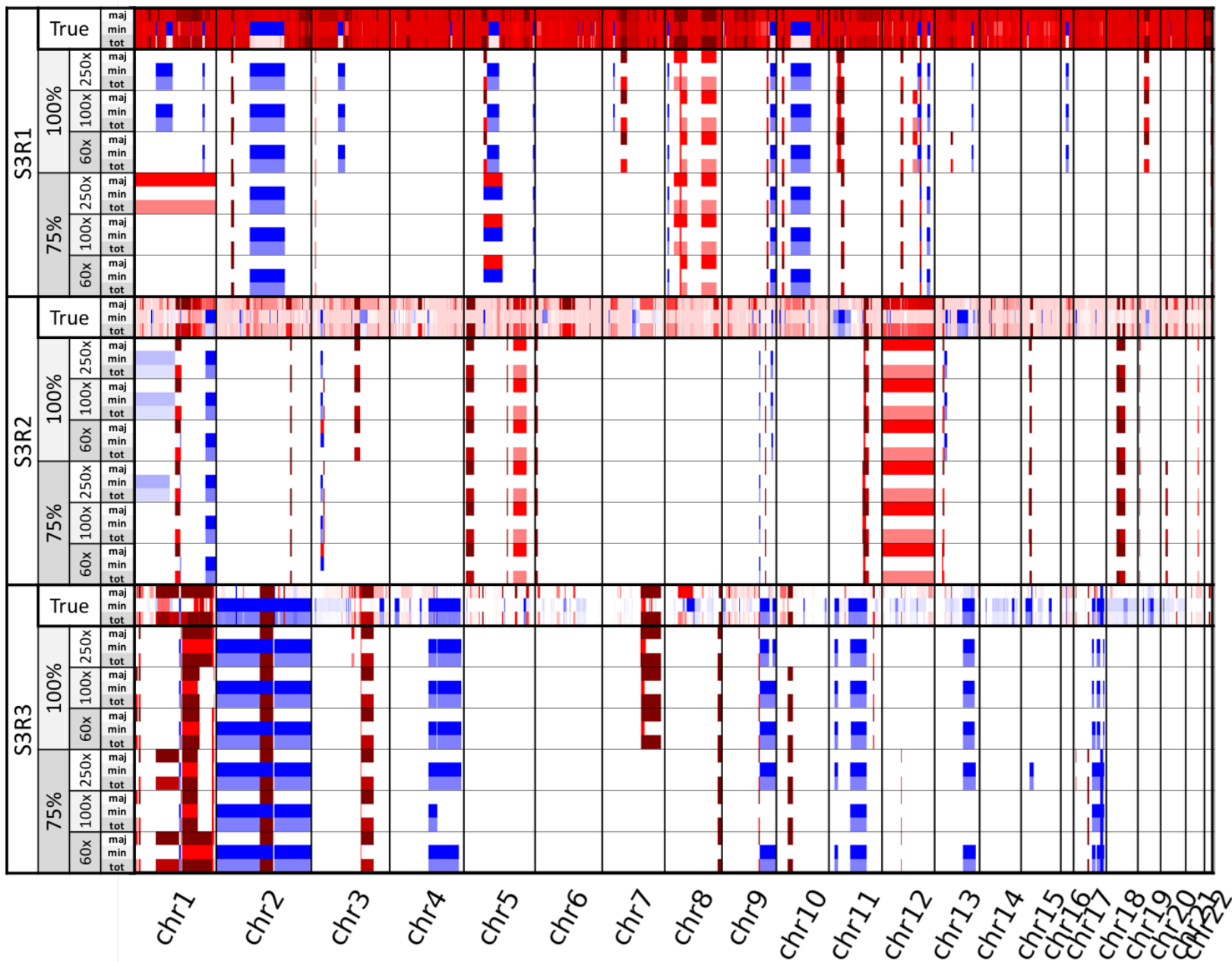




Sclust



Sclust



### 3.6 References

- 1000 Genomes Project Consortium, T. 1000 G.P. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
- Abécassis, J. *et al.* (2019) Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data. *PLoS One*, **14**.
- Andor, N. *et al.* (2016) Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.*, **22**, 105–13.
- Anzar, I. *et al.* (2019) NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC Med. Genomics*, **12**, 63.
- Aran, D. *et al.* (2015) Systematic pan-cancer analysis of tumour purity. *Nat. Commun.*, **6**, 8971.
- Van der Auwera, G.A. *et al.* (2013) From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.*
- Barthel, F.P. *et al.* (2019) Longitudinal molecular trajectories of diffuse glioma in adults. *Nature*, **576**, 112–120.
- Baysan, M. *et al.* (2017) Detailed longitudinal sampling of glioma stem cells *in situ* reveals Chr7 gain and Chr10 loss as repeated events in primary tumor formation and recurrence. *Int. J. Cancer*, **141**, 2002–2013.
- Benjamin, D. *et al.* (2019) Calling Somatic SNVs and Indels with Mutect2.
- Benjamini, Y. and Speed, T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, e72–e72.
- Beraldi, D. GitHub - dariober/cnv\_facets: Somatic copy variant caller (CNV) for next generation sequencing.
- Bhandari, V. *et al.* (2018) Quantifying the Influence of Mutation Detection on Tumour Subclonal Reconstruction. *bioRxiv*, 418780.
- Bohnert, R. *et al.* (2017) Comprehensive benchmarking of SNV callers for highly admixed tumor data. *PLoS One*, **12**, e0186175.
- Caravagna, G. *et al.* (2020) Subclonal reconstruction of tumors by using machine learning and population genetics. *Nat. Genet.*, **52**, 898–907.
- Chen, L. *et al.* (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science (80-. )*, **355**.
- Chen, Z. *et al.* (2020) Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci. Rep.*, **10**, 1–9.
- Chiang, D.Y. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.
- Chkhaidze, K. *et al.* (2019) Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *bioRxiv*, 544536.
- Cun, Y. *et al.* (2018) Copy-number analysis and inference of subclonal populations in cancer genomes using ScIust. *Nat. Protoc.*, **13**, 1488–1501.
- Dentro, S.C. *et al.* (2020) Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *bioRxiv*, 312041.
- DePristo, M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
- Detering, H. *et al.* (2019) Accuracy of somatic variant detection in multiregional tumor sequencing data. *bioRxiv*, 655605.
- Droop, A. *et al.* (2018) How to analyse the spatiotemporal tumour samples needed to investigate cancer evolution: A case study using paired primary and recurrent glioblastoma. *Int. J. Cancer*, **142**, 1620–1626.
- Durbin, R.M. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
- El-Kebir, M. *et al.* (2016) Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. *Cell Syst.*, **3**, 43–53.
- Ewing, A.D. *et al.* (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods*, **12**, 623–630.

- Favero, F. *et al.* (2015) Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, **26**, 64–70.
- Gawad, C. *et al.* (2016) Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.*, **17**, 175–188.
- Giroux Leprieur, E. *et al.* (2020) Sequential ctDNA whole-exome sequencing in advanced lung adenocarcinoma with initial durable tumor response on immune checkpoint inhibitor and late progression. *J. Immunother. Cancer*, **8**, e000527.
- Guillaumet-Adkins, A. *et al.* (2017) Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biol.*, **18**, 45.
- Guo, Y., Long, J., *et al.* (2012) Exome sequencing generates high quality data in non-target regions. *BMC Genomics*, **13**, 194.
- Guo, Y., Li, J., *et al.* (2012) The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics*, **13**, 666.
- Ha, G. *et al.* (2014) TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.*, **24**, 1881–93.
- Hu, Y. *et al.* (2013) Tumor-Specific Chromosome Mis-Segregation Controls Cancer Plasticity by Maintaining Tumor Heterogeneity. *PLoS One*, **8**, e80898.
- ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**, 82–93.
- Jiang, Y. *et al.* (2016) Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl. Acad. Sci.*, **113**, E5528–E5537.
- Johnson, B.E. *et al.* (2014) Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science (80-. )*, **343**, 189–193.
- Kandoth, C. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333.
- Kim, S. *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.
- Koboldt, D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–76.
- Körber, V. *et al.* (2019) Evolutionary Trajectories of IDH WT Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis. *Cancer Cell*, **35**, 692–704.e12.
- Krijgsman, O. *et al.* (2014) Focal chromosomal copy number aberrations in cancer—Needles in a genome haystack. *Biochim. Biophys. Acta - Mol. Cell Res.*, **1843**, 2698–2704.
- Kuipers, J. *et al.* (2017) Advances in understanding tumour evolution through single-cell sequencing. *Biochim. Biophys. Acta - Rev. Cancer*, **1867**, 127–138.
- Leshchiner, I. *et al.* (2018) Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. *bioRxiv*, 508127.
- Letouzé, E. *et al.* (2017) Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.*, **8**.
- Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*.
- Li, H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–9.
- Martelotto, L.G. *et al.* (2017) Whole-genome single-cell copy number profiling from formalin-fixed paraffin-embedded samples. *Nat. Med.*, **23**, 376–385.
- Mills, R.E. *et al.* (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.*, **16**, 1182–90.
- Miura, S. *et al.* (2018) Predicting clone genotypes from tumor bulk sequencing of multiple samples. *Bioinformatics*, **34**, 4017–4026.
- Mose, L.E. *et al.* (2019) Improved indel detection in DNA and RNA via realignment with ABRA2. *Bioinformatics*, **35**, 2966–2973.
- Mullaney, J.M. *et al.* (2010) Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.*, **19**, R131–6.
- Nam, J.-Y. *et al.* (2016) Evaluation of somatic copy number estimation tools for whole-exome sequencing data. *Brief. Bioinform.*, **17**, 185–192.
- Narzisi, G. *et al.* (2018) Genome-wide somatic variant calling using localized colored de Bruijn graphs.

- Commun. Biol.*, **1**, 1–9.
- Nilsen,G. *et al.* (2012) Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics*, **13**, 591.
- Noorbakhsh,J. *et al.* (2018) Distribution-based measures of tumor heterogeneity are sensitive to mutation calling and lack strong clinical predictive power. *Sci. Rep.*, **8**.
- Pedregosa,F. *et al.* (2011) Scikit-learn: Machine Learning in Python.
- Rashid,M. *et al.* (2013) Cake: A bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. *Bioinformatics*, **29**, 2208–2210.
- Rheinbay,E. *et al.* (2020) Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, **578**, 102–111.
- Rode,A. *et al.* (2016) Chromothripsis in cancer cells: An update. *Int. J. Cancer*, **138**, 2322–2333.
- Roth,A. *et al.* (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396–8.
- Rubanova,Y. *et al.* (2020) Reconstructing evolutionary trajectories of mutation signature activities in cancer using TrackSig. *Nat. Commun.*, **11**.
- Sakharkar,M.K. *et al.* (2004) Distributions of Exons and Introns in the Human Genome. *In Silico Biol.*, **4**, 387–393.
- Salcedo,A. *et al.* (2020) A community effort to create standards for evaluating tumor subclonal reconstruction. *Nat. Biotechnol.*, **38**, 97–107.
- Saunders,C.T. *et al.* (2012) Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, **28**, 1811–1817.
- Schwartz,R. and Schäffer,A.A. (2017) The evolution of tumour phylogenetics: Principles and practice. *Nat. Rev. Genet.*, **18**, 213–229.
- Shen,H. *et al.* (2013) Comprehensive Characterization of Human Genome Variation by High Coverage Whole-Genome Sequencing of Forty Four Caucasians. *PLoS One*, **8**, e59494.
- Shen,R. and Seshan,V.E. (2016) FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.*, **44**, e131–e131.
- Shi,W. *et al.* (2018) Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity. *Cell Rep.*, **25**, 1446–1457.
- Siegmund,K. and Shibata,D. (2016) At least two well-spaced samples are needed to genotype a solid tumor. *BMC Cancer*, **16**, 250.
- Sikorsky,J.A. *et al.* (2007) DNA damage reduces Taq DNA polymerase fidelity and PCR amplification efficiency. *Biochem. Biophys. Res. Commun.*, **355**, 431–7.
- Sun,R. *et al.* (2017) Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.*, **49**, 1015–1024.
- Talevich,E. *et al.* (2016) CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Comput. Biol.*, **12**, e1004873.
- Tauriello,D.V.F. *et al.* (2018) TGF $\beta$  drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature*, **554**, 538–543.
- Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.
- Wang,M. *et al.* (2015) Somatic Mutation Screening Using Archival Formalin-Fixed, Paraffin-Embedded Tissues by Fluidigm Multiplex PCR and Illumina Sequencing. *J. Mol. Diagn.*, **17**, 521–32.
- Wang,M. *et al.* (2020) SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. *Sci. Rep.*, **10**, 12898.
- Watkins,T.B.K. and Schwarz,R.F. (2018) Phylogenetic Quantification of Intratumor Heterogeneity. *Cold Spring Harb. Perspect. Med.*, **8**, a028316.
- Williams,M.J. *et al.* (2016) Identification of neutral tumor evolution across cancer types. *Nat. Genet.*, **48**, 238–244.
- Wong,S.Q. *et al.* (2014) Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. *BMC Med. Genomics*, **7**, 23.
- Xi,R. *et al.* (2011) Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, E1128-36.
- Xu,C. (2018) A review of somatic single nucleotide variant calling algorithms for next-generation sequencing

- data. *Comput. Struct. Biotechnol. J.*, **16**, 15–24.
- Yuan, K. *et al.* (2018) Ccube: A fast and robust method for estimating cancer cell fractions. *bioRxiv*, 484402.
- Zack, T.I. *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, **45**, 1134–1140.
- Zare, F. *et al.* (2017) An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*, **18**, 286.
- Zhao, M. *et al.* (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, **14 Suppl 11**, S1.

# Chapter 4 – Identification of pathways relevant to GBM progression through therapy

The results in section 4.2.1, involving the SubClonalSection model, were previously published in Nature (Barthel *et al.*, 2019). The material is reproduced here with the lead author's permission.

## 4.1 Introduction

### 4.1.1 Overview

The overall aim of this study is to identify cellular processes that influence glioblastoma (GBM) cells' ability to resist therapy. While the previous two chapters were focussed on identifying the most accurate pipelines for investigating such effects through subclonal deconvolution, using simulated datasets, this chapter reports on the analysis of real GBM datasets through additional and alternative approaches.

I had initially intended to extend the benchmarking of Chapter 3 to include multi-sample subclonal deconvolution methods, and then apply the most accurate pipeline to our group's in-house samples. However, whilst working on the benchmarking, my group's involvement in the Glioma Longitudinal AnalySiS (GLASS) consortium meant we gained access to a considerable dataset, containing genome sequencing data for 257 matched primary and recurrent glioma samples (GLASS Consortium, 2018). This dataset is an important and unprecedented resource for GBMs and other gliomas, so I therefore wanted to utilise it in my analyses. The raw sequencing reads were not accessible for the dataset, and instead I only had access to the mutation call sets. These had already been run through a subclonal deconvolution method in order to characterise the intratumour heterogeneity (ITH) in the samples, and so I decided to proceed with investigating the biology underlying the tumours' progression through therapy, using this available data.

### 4.1.2 The GLASS dataset

GLASS is a worldwide community-driven resource of pooled datasets containing longitudinal sequencing and clinical data from glioma patients. It currently contains 257 adult glioma patients, with somatic variant and copy number calls from a mix of whole-genome sequencing (WGS) and whole-exome sequencing (WES), as well as other omics datasets. Among the patients are 94 that had isocitrate dehydrogenase wild-type (IDHwt) GBMs at both primary and recurrent time points, had received standard therapy with both temozolomide (TMZ) and radiotherapy, and had high quality sequencing for both primary and recurrent tumours. These samples therefore allowed me to focus on processes that affect specifically GBMs (as opposed to other gliomas which may show different resistance mechanisms in response to therapy), in

response to standard treatment (as opposed to the minority of cases that may, for example, only receive radiotherapy or trial drugs) in part of the analyses.

Raw sequencing reads are not provided in the GLASS dataset, and I therefore use the provided mutation call sets. Specifically, somatic copy number alterations (CNAs), purity, and ploidy estimates are provided by TITAN (Ha *et al.*, 2014), and somatic point variants (single nucleotide variants (SNVs) and short insertions and deletions (InDels)) are provided by Mutect2 in multi-sample mode (Benjamin *et al.*, 2019). Both of these methods showed good results in the benchmarking in chapter 3. Cancer cell fractions (CCFs) are also provided in the dataset, estimated by PyClone (Roth *et al.*, 2014). While the benchmarking found that PyClone was less accurate than Ccube in estimating CCFs, I opted not to rerun the analysis with Ccube as the difference was relatively small. It's also likely that the CCF estimates in the GLASS dataset are more accurate than those in the benchmarking, as all samples for a patient (typically one primary and one recurrent) were jointly analysed in a multi-sample set-up. Furthermore, whereas the benchmarking only used clonal CNAs with PyClone, both clonal and subclonal CNAs were used in the GLASS analysis. While the latter may cause issues for PyClone's algorithm, which treats all CNAs as the overall copy number across all subclones, it might achieve better results.

As well as investigating variant changes in CCFs in response to therapy, this chapter also employs analyses using other information, such as VAFs, or the binary presence or absence of variants, which avoid any inevitable inaccuracies of CCF estimation.

### 4.1.3 Investigating the mode of tumour evolution

Looking for variants that confer therapy resistance assumes that there is a level of selection for subclones containing these alterations in response to treatment. This, however, is not guaranteed to be the case, and instead all subclones may survive equally well and the tumour continues to evolve neutrally. It's therefore of interest to determine how prevalent selection is in recurrent GBMs, and whether it is associated with therapy.

A novel method available for identifying selection in a tumour is the SubClonalSelection model (Williams *et al.*, 2018). Its underlying rationale is the observation that in neutrally evolving tumours, subclonal variant allele frequencies (VAFs) (in diploid regions, without distortions from copy number alterations) follow a power-law distribution (with a factor of -2), creating a neutral tail where the cumulative number of mutations has a linear relationship with the inverse of their allele frequency. This occurs as cells constantly gain new mutations and all continue to divide at a similar rate. Alternatively, any deviation from the linear relationship suggests variations in the growth rates between cells and, therefore, that there is selection of some subclones over others (Figure 30). However, a distortion will also be seen if the mutation rate of cells within a subclone is altered, even in the absence of selection (Williams *et al.*, 2016).



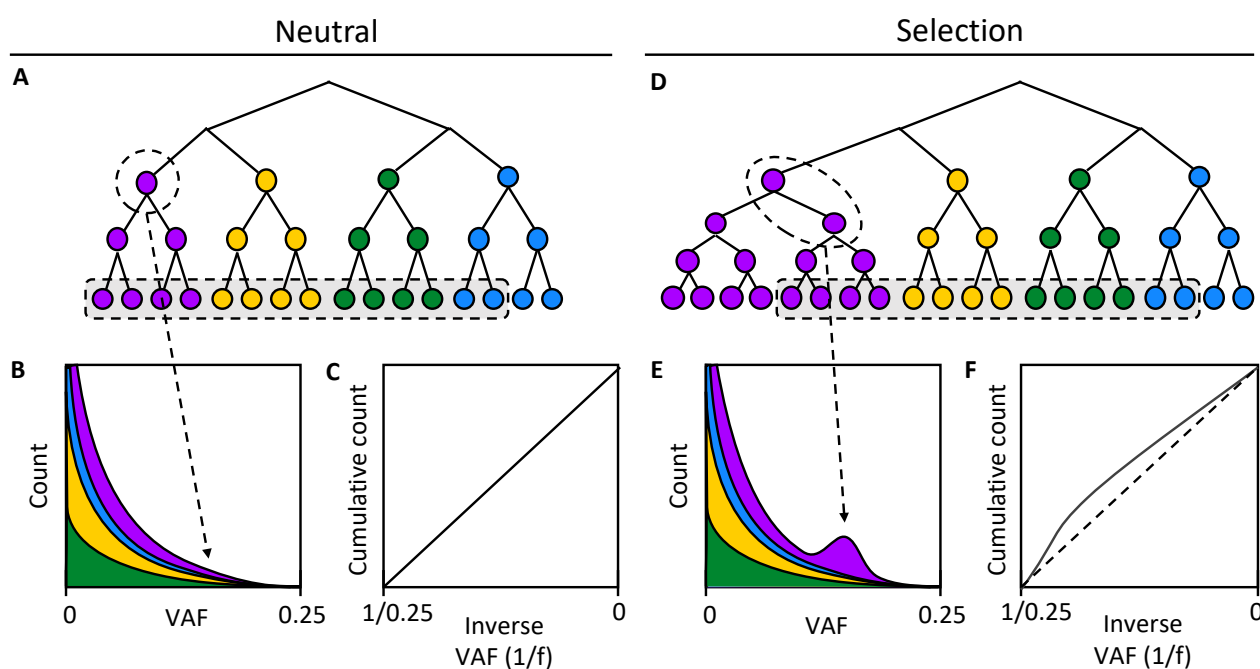


Figure 30. Representations of tumours evolving under either neutral or selection evolutionary modes. A+D) Phylogenetic trees representing cells coloured according to subclone. In A, all subclones grow and divide at the same rate, whereas in D, cells in the purple subclone have a selective advantage and divide more rapidly. The dashed boxes indicate sampled cells. B+E) Distributions of VAFs in the samples, coloured by subclone. In (E), the purple subclone distorts the overall distribution due to an increased number of variants that are ‘clonal’ to the sampled region of the purple subclone. C+F) Cumulative distributions of inverse VAFs. The neutrally evolving tumour (C) shows a linear relationship, whereas the tumour with selection of the purple clone (F) does not.

SubClonalSelection combines modelling of this with Bayesian model selection and parameter inference to determine the probability of either a neutrally evolving tumour or one with selection present. Furthermore, the model can predict the relative fitness advantage of up to 2 detected subclones compared to other cells in a tumour. Using this method with the GLASS dataset, I determine the prevalence of selection in GBM and other gliomas, and the affect that standard therapy has on tumour evolution.

#### 4.1.4 Identification of biological pathways relevant in GBM progression through therapy

This second part of this chapter is focused on identifying what specific cellular processes may be relevant to GBM progression through therapy. Although it has previously been shown that there is generally no genetic bottleneck for GBM through therapy (Körber *et al.*, 2019), even small selective advantages to cells can help us to understand what processes are behind therapy resistance, or may allow us to develop drugs that

prolong patient's lives by minimising the selective advantage of those cells. Furthermore, variants in cells that disappear from primary to recurrent, may inform us on cellular processes that confer sensitivity to therapy.

While the SubClonalSelection model is able to detect when selection is likely present in a tumour, allowing us to understand trends in selective pressures across a cohort, it is not able to predict which variants, and therefore which cellular processes, are driving selection, or even which variants are in the clonally expanding cells (though this can be estimated through the predicted cellular frequency of the expanding clone). It also cannot determine in which tumours the selection is in response to therapy. Furthermore, many tumours in the dataset have insufficient data for use with the model, and in others, selection may have been missed due to a number of reasons, which are discussed later. Therefore, I next aim to investigate which variants may confer therapy resistance or sensitivity, through patterns of their frequencies between primary and recurrent tumours.

Most variants in tumours are neutral passengers that have no significant effect on a subclone's progression (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). Differentiating these from driver variants, that increase a subclone's fitness and are under positive selection, is a large field of cancer bioinformatics (Bailey *et al.*, 2018; Martínez-Jiménez *et al.*, 2020). A common approach is to look for frequently mutated genes in cancers, or those that are mutated more often than expected by chance. Extensions of this include looking for variants that cluster to specific regions of a gene sequence, or that cause amino acid changes that cluster in three-dimensional space in a protein structure. Other approaches use machine learning with databases of cancer associated genes to predict the functional consequence of variants. Large pan-cancer studies have employed consensus approaches to investigate driver genes across cancers, by combining methods that use the above approaches (Bailey *et al.*, 2018; Martínez-Jiménez *et al.*, 2020). These studies identified 299-568 driver genes in total across cancers, most of which were cancer specific. Another commonly used approach to identifying driver genes is to compare the numbers of non-synonymous and synonymous variants within them, under the assumption that if alterations to a gene result in clonal sweeps of cells containing the variant, or even just expansions to detectable proportions, then a higher number of non-synonymous relative to synonymous variants is likely to be observed (Martincorena *et al.*, 2017).

The above approaches all focus on individual genes, many of which may lack significance when identifying individual drivers, particularly in smaller cohorts. To overcome this, researchers can instead look at the frequency of variants across multiple related genes in a pathway, thereby leveraging increased statistical power and gaining additional insights into the cellular mechanisms that result in the selection. In this study, I use such an approach with the GLASS dataset to identify variants that may drive therapy resistance or sensitivity in GBM. In doing so, altered driver genes that individually may not recur sufficiently to reach

significance on their own, have the potential to stand out. This is particularly important given that previous studies have found only scarce evidence of repeated genetic alterations in recurrent GBM tumours (Kraboth and Kalman, 2020; Barthel *et al.*, 2019; Körber *et al.*, 2019; Wang *et al.*, 2016; Cahill *et al.*, 2007).

Many methods are available for investigating the driver potential of genes based on their shared effect on pathways, and generally fall into three categories; pathway analysis through gene set enrichment analysis (GSEA), network analysis, and *de novo* methods. These can be applied to any dataset from which a list of altered genes can be obtained, such as, differentially expressed genes from RNAseq (although additional methods specific to this type of data that take into account the extent of differential expression in every gene without hard cut-offs, may be more suitable (García-Campos *et al.*, 2015)), genes with epigenetic modifications, or, in the case of this study, mutated genes from DNA sequencing.

GSEA pathway analysis methods look for over-representation of altered genes in predefined distinct biological pathways, each represented by a list of genes that all function to carry out a process (García-Campos *et al.*, 2015). Pathways are assigned enrichment scores based on the extent that the number of altered genes occurring in a set exceeds the number that would be expected by chance, often using a hypergeometric distribution to determine this. P-values for the enrichment scores for each pathway are then calculated. This is the underlying approach in all GSEA methods, though many more sophisticated adaptations have been developed. For example, PathScore calculates pathway enrichment separately per patient, allowing consideration of individual mutation rates, as well as accounting for gene lengths and gene specific background mutation rates (Gaffney and Townsend, 2016). Hundreds of databases of pathways are available for use in pathway analysis, with popular choices discussed in (García-Campos *et al.*, 2015) and a comprehensive list provided in the Pathguide website (Bader *et al.*, 2006).

The downside of GSEA pathway analysis is that information about cellular processes is split into distinct, and sometimes large and imprecise, individual pathways that discard crosstalk information between them. Network analysis methods avoid this by using whole networks such as protein-protein interaction networks to identify driver modules or subnetworks (Creixell *et al.*, 2015; Zhang and Zhang, 2018). For example, HotNet2 uses an approach of network propagation and heat diffusion from altered gene nodes along network edges to identify 'hot' recurrently mutated subnetworks (Leiserson *et al.*, 2015; Vandin *et al.*, 2011). Another method, MEMo (Ciriello *et al.*, 2012) combines prior network information with assessment of mutual exclusivity between genes (discussed below) to identify driver subnetworks. Whilst network methods are able to incorporate all the information in a predefined network, this is unlikely to fully or accurately represent all cellular processes, particularly in cells with altered physiology such as in tumours. To overcome this, *de novo* pathway methods, such as Dendrix (Vandin *et al.*, 2012), Multi-Dendrix (Leiserson *et al.*, 2013), and GAMToC (Melamed *et al.*, 2015), instead identify driver processes through patterns of altered genes (Creixell *et al.*, 2015; Zhang and Zhang, 2018; Vandin, 2017). Common approaches

include identifying co-occurring altered genes, which may indicate an additive or synergistic effect between two separate pathways. Examples of co-occurring altered genes have been found in pathways known to be altered in GBM, such as retinoblastoma protein (RB) and receptor tyrosine kinase (RTK) signalling, DNA damage, mitogenic, and cell cycle pathways (Gu *et al.*, 2013; Melamed *et al.*, 2015). Alternatively, patterns of mutual exclusivity between genes, where only one of the genes is typically altered per tumour, can instead be used to identify driver pathways containing both of these genes. In GBM, genes in known driver pathways, phosphoinositide 3-kinases (PI3K), p53, and RB, have been shown to have mutual exclusivity, as well as those in multiple novel gene sets involving several transcription factors (Ciriello *et al.*, 2012; McLendon *et al.*, 2008; Vandin *et al.*, 2012). It has been suggested, however, that many of these pairs of genes showing mutually exclusivity, do so due to alterations being subtype specific, and not because they are interchangeable in causing the same effect. A recent method, Subtype-specific Pathway Linear Progression Model (SPM), combines identification of driver pathways through mutual exclusivity, with predictions for relative timings of pathway alterations using CCFs, whilst aiming to classify patients into different subtypes. In GBM, this identified four sub-types, using inputs of 15 predefined known driver genes (Khakabimamaghani *et al.*, 2019).

My aim in this chapter is to use pathway analysis on the GLASS dataset, in a way that informs on processes that are specifically relevant to GBM progression through therapy. To achieve this, I subset variants based on changes in their cellular frequencies from primary to recurrent tumours, or whether variants are unique to primary tumours, unique to recurrent tumours, or shared between both. These distinct gene sets are then used to define the altered genes inputted into the pathway analysis. This may allow detection of pathways that are specifically relevant to tumour progression through therapy, as opposed to GBM tumours in general, as have been the focus of the above previous studies. For this purpose I chose to use a GSEA pathway analysis method, PathScore, as this allows easy direct comparison of identical pathways between results from different inputs. Furthermore, unlike *de novo* methods, GSEA is suitable for use with subsets of altered genes in a tumour, and able to detect driver pathways that may only be present in a small number of patients.

PathScore takes as input a list of patient-gene pairs, and for each pathway in a database, provides 1) the actual pathway size, based on the number of DNA bases across all genes in the set, 2) the effective pathway size, indicating the maximum likelihood estimate of pathway size given the per-patient alteration rates, pre-defined gene-specific background mutation rates, and gene transcript lengths, and 3) a P-value for the disparity between actual and effective pathway size from a likelihood ratio test. The effect size can then be determined from the ratios between the actual and effective pathway sizes, indicating the degree of overburden of pathways with alterations.

## 4.2 Results

### 4.2.1 Investigating the mode of tumour evolution

To identify the mode of evolution in GBM and other gliomas, I started with a set of 222 patients from the GLASS dataset that had been previously filtered for high quality sequencing and copy number data for matched primary and recurrent tumours (the gold set in (Barthel *et al.*, 2019)). I further filtered tumours for those with a minimum purity of 0.5, and that had  $\geq 25$  subclonal variants in copy neutral regions (ie. a copy number of 2 in diploid tumours, or 3 in triploid tumours), with a coverage of  $\geq 30x$ . This was based on the pre-processing thresholds recommended by the SubClonalSelection authors (Williams *et al.*, 2018), though they emphasise that the program's performance increases with increasing numbers of input subclonal variants. Excluding tumours for which the SubClonalSelection model was unable to complete successfully (those which could not converge upon a suitable result), the total number of included patients was 183, with 104 including estimates for both primary and recurrent tumours. Tumours were classified as evolving either neutrally or under selection based on the most probable scenario (i.e. whether the probability of neutral selection was greater or lower than 0.5). Similar proportions of neutral and selection classifications were seen when limiting results to only those with probabilities of greater than 0.7 for either neutral or selection (though this reduced our sample size by approximately a quarter, meaning less data for downstream analyses), thereby validating the 0.5 cut-off.

Neutral evolution was the most common evolutionary mode in both primary and recurrent tumours. In recurrent tumours across all subtypes (IDHmutant,1p/19q\_codeletion:n=19, IDHmutant\_non-codeletion:n=43, IDHwt:n=83), only 37% were classed as showing selection, although this increased to 47% in IDHwt recurrent tumours (of which the majority are GBMs) (Figure 31A). A Cox proportional hazards model, including age at first diagnosis, glioma subtype (IDHwt, IDHmut-noncodel, IDHmut-codel), and evolution mode in the recurrent tumour, demonstrated a significant association between shorter survival and selection at recurrence in all subtypes (Hazard ratio = 1.53, 95% confidence interval 1.00–2.41,  $P=0.048$ ). This survival difference is seen in a Kaplan–Meier curve for IDHwt gliomas ( $P=0.027$ , log-rank statistic) (Figure 31B). This suggests that recurrent tumours with subclonal selection are generally more aggressive. However, there was no significant associations between either chemotherapy ( $P=0.74$ , Fisher's exact test) or radiotherapy ( $P=0.68$ , Fisher's exact test) with selection in the recurrence. This may have been due to insufficient numbers of recurrent tumours from patients who did not receive therapy, to achieve a significant result. However, there was also no significant associations between chemotherapy ( $P=0.85$ , Fisher's exact test) or radiotherapy ( $P=0.72$ , Fisher's exact test) with selection in the recurrence when instead using primary tumours to infer the prevalence of selection in untreated tumours. These results suggest that therapy is not a prevelant factor in inducing strong selection of clones in recurrent tumours.

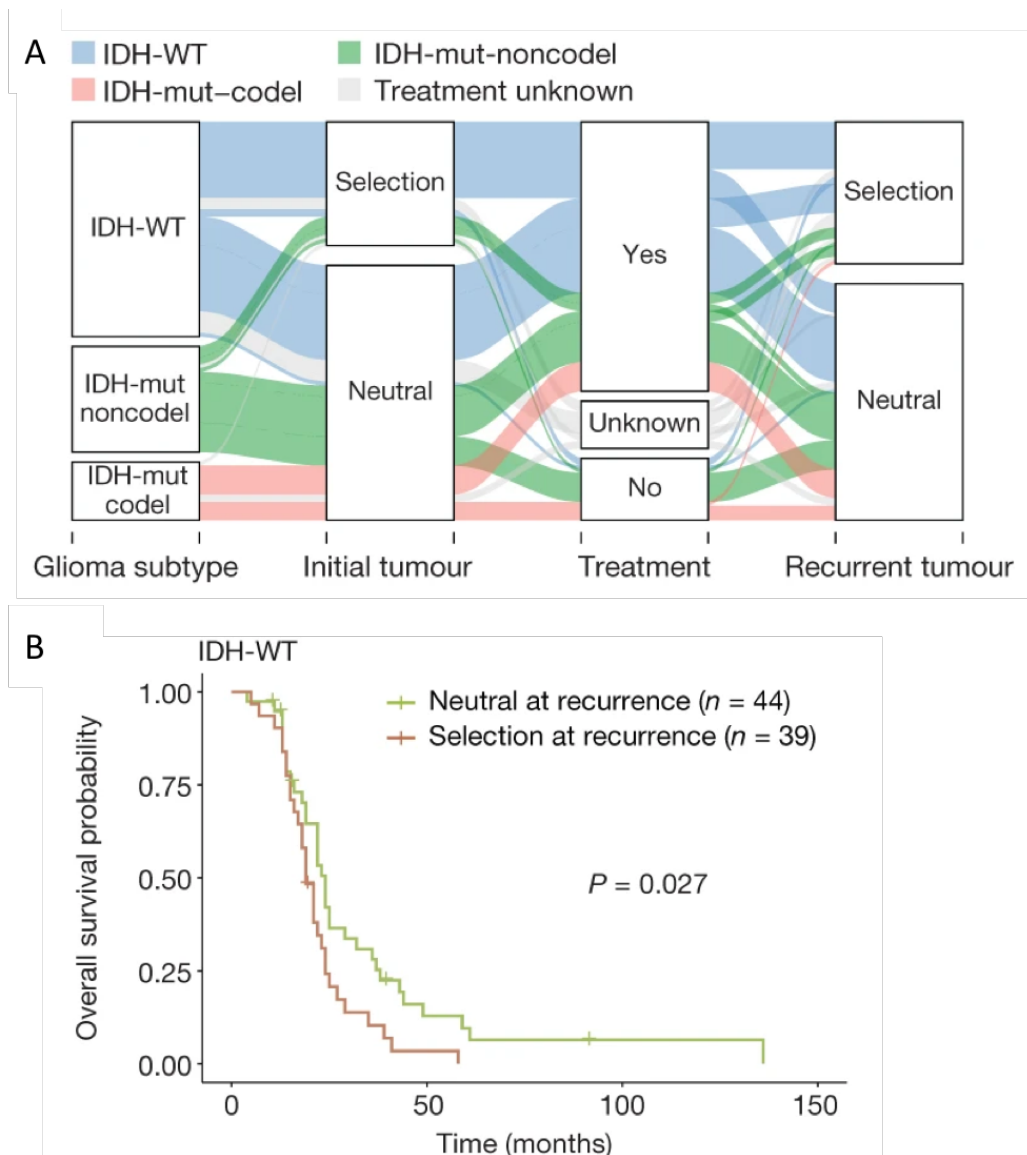


Figure 31. A) Sankey plot indicating the breakdown of SubClonalSelection evolutionary modes by subtype and therapy ( $n = 104$  for patients with model results at both primary and recurrent time points). The sizes of the bands reflect sample sizes and band colours highlight the glioma subtype. Grey colouring reflects instances when treatment information was not available. B) Kaplan–Meier curve showing survival differences between IDH-wild-type recurrent tumours demonstrating selection ( $n = 39$ ) compared with neutrally evolving tumours ( $n = 44$ ).  $P$  value determined by log-rank test.

### 4.2.3 Pathway analysis

Though the SubClonalSelection results suggest that selection in recurrent tumours is not significantly associated with therapy, it is still of interest to investigate how genetic alterations affect tumours' response to therapy. Identifying even small fitness advantages to therapy that may not have been detectable in the previous analysis, or which were present in only a small number of patients, could allow us to better understand the processes involved in GBM's ability to resist therapy. The same is true for alterations that

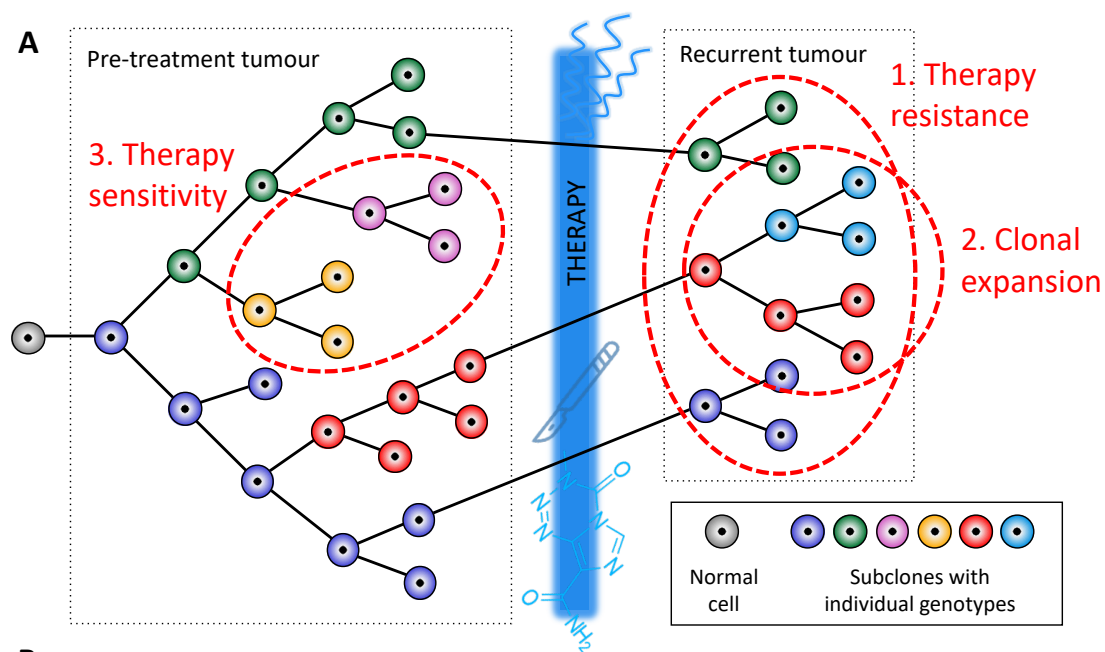
sensitise cells to therapy, which are not considered in the above method. I therefore looked for evidence of such scenarios by performing pathway analysis with PathScore, using subsets of variants that will inform on specific patterns of progression through therapy.

#### 4.2.4 Running PathScore

I started with a larger group of patients than in the SubClonalSelection analysis, to include samples previously filtered out due to poorer quality copy number data (the silver set in (Barthel *et al.*, 2019)), which is not so important in this analysis as variant copy numbers are ignored for most parts, and those parts that do take copy number into account are likely able to cope with a large amount of noise without negative effects on results. Increasing sample number, however, will increase the power of the analysis. This analysis included 117 patients with IDHwt GBM at both primary and recurrent.

To minimise noise from variants that have no effect on the function of the gene they're in, I include only those that are predicted to physiologically alter the protein, and filter out others where an effect is less likely. To achieve this, I annotated variants from the GLASS dataset with the Ensembl Variant Effect Predictor (McLaren *et al.*, 2016) (VEP) which characterises the effect that variants have on genomes. This includes providing scores from SIFT (Vaser *et al.*, 2015) and PolyPhen-2 (Adzhubei *et al.*, 2010) which predict the affect that amino acid modifying variants have on protein function based on sequence homology through evolution and the physical properties of amino acids. I then used this to filter the variants to create two sets of different stringencies; the high stringency set contained protein modifying variants predicted by either SIFT or PolyPhen to affect the function of a protein, and the low stringency set contains any protein modifying variant, as well as variants that may alter the regulation of the protein. Affected genes from both the high stringency and low stringency variant sets were run separately through PathScore. The results from each of the two sets generally agreed with each other, and therefore only those from the high stringency set are reported and discussed here.

Next, I assigned variants from each set into groups: 1) all primary tumour mutations, 2) all recurrent tumour variants, 3) variants private to the primary tumour, 4) variants private to the recurrent tumour, 5) shared variants, 6) shared variants that increase in VAF, and 7) shared variants that increase in CCF (Figure 32). Using both VAFs and CCFs to determine variants that increase through treatment has two benefits over using CCFs alone; i) Only around two thirds of variants in the GLASS dataset have a CCF estimate available, due to failed runs of PyClone, or coverage <30x, so using VAFs includes more variants, and ii) CCF estimates are not always accurate; While estimates from PyClone in the GLASS dataset are likely to be more accurate than those seen in the benchmarking from Chapter 3, as a result of differences between set-up, there is still a significant level of doubt in their accuracy.



B

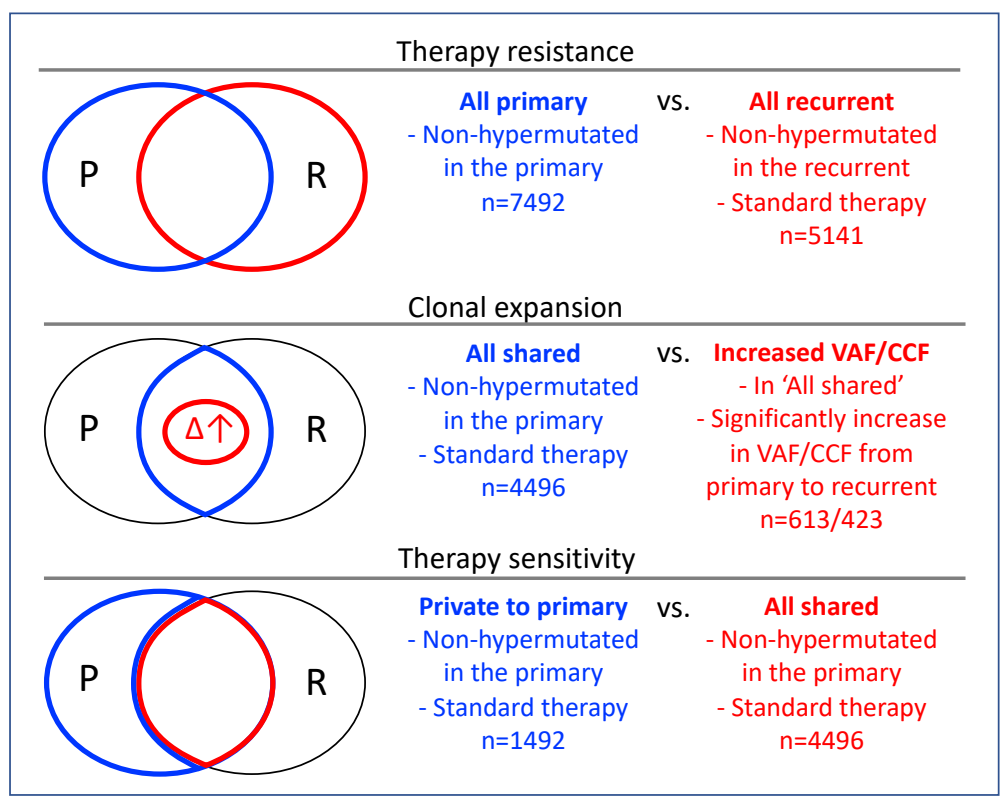


Figure 32. A) Evolutionary progression of genetic subclones in GBM through therapy, and the cell fractions this study aims to characterise in order to identify pathway candidates for conferring therapy resistance, clonal expansion, and therapy sensitivity. B) Variant groups used in the pathway analysis comparisons to characterise the relevant cell fractions. Venn diagrams indicate fractions of variants across primary (P) and recurrent (R) tumours.



Of the 117 patients with IDHwt GBM tumours, only those who received standard therapy with both TMZ and radiotherapy (n=94/117) were included in the analysis, with the exception of when looking at 'all primary' (ie. private to the primary or shared) variants, where patients with any therapy were included as it has no effect on this group. Patients with hypermutated primary tumours (n=4/117), defined as having greater than 10 variants per megabase, were excluded for primary and shared variants, and those with hypermutated recurrent tumours (n=16/117) were excluded for recurrent variants to reduce the level of noise (Figure 32B). The number of patients in each group were as follows: All primary:113, all recurrent:79, all shared:91, increased VAF:66, increased CCF:41, private to primary:91.

I ran PathScore using pathway genesets from the Molecular Signatures Database (*MSigDB*) (Liberzon *et al.*, 2011) which combines pathways from a number of sources, including KEGG, Biocarta, Reactome and Nature-NCI databases. Each group of variants is input separately, allowing a comparison of results between them. By ranking pathways on the fold change of number of patients altered in it between groups, this allows identification of pathways that are most differentially altered and, depending on which groups are being compared, are candidates for conferring therapy resistance, driving clonal expansion, or causing sensitivity to therapy (Figure 32).

Due to the small number of samples available in the GLASS dataset, the statistical significance able to be achieved in this analysis is limited. The goal is to only visualise the data in a meaningful biological context, with the aim of identifying pathways and hypotheses for further investigation, and without formally attributing statistical significance to changes in the pathways. Multiple-testing corrections are therefore not performed.

#### 4.2.5 Pathways differentially enriched between primary and recurrent tumours

To identify pathways that may confer therapy resistance, I looked for those that were altered in higher numbers of patients from all recurrent variants than from all primary variants, thereby identifying pathways that appear altered more frequently after therapy, and indicating an increased likelihood of the ability to confer resistance to therapy. Whilst many of the recurrent variants not seen in the primary will have developed post-therapy, be inconsequential, or were missed in the primary due to sampling bias, others may come from variants that were of too low frequency in the primary tumours to be detected, owing to their lack of a selective advantage in the absence of therapy, and which subsequently expanded to detectable frequencies in the recurrences. Such expansions, if small and with variants still at low frequency, may not have been identified by the SubClonalSelection model.

The most enriched pathways in both primary and recurrent tumours generally contained known driver genes (*MTOR*, *TP53*, *PTEN*, *EGFR*, *IDH1*, *IDH2*, *PIK3CA*, *PIK3R1*, *PIK3CG*, *RB1*, *NF1*) (Bailey *et al.*, 2018;

Krishnan *et al.*, 2020). In the recurrent variants, the most enriched pathway that didn't involve a commonly known driver gene was REACTOME\_SYNTHESIS\_OF\_BILE\_ACIDS\_AND\_BILE\_SALTS\_VIA\_7ALPHA\_HYDROXYCHOLESTEROL (155th out of 1329 pathways). More importantly, this was also the pathway altered in the highest number of recurrent tumours, out of the top 20 pathways that increased most in alterations from primary to recurrent (Table 6); 6/7492 variants across 5/113 patients (4.4%) occurred in the pathway in primary variants, and 9/5141 variants across 8/79 patients (10.1%) occurred in the pathway in all recurrent variants. (These numbers differ slightly to those in Table 6 as some genes were rejected by Pathscore in its calculations, due to a lack of valid gene identifiers. The values stated here reflect the true numbers of variants in the pathway.) This is a 2.19 fold enrichment of alterations to the pathway in the recurrent variants compared to the primary ( $P=0.06$ , 1-tailed Chi-squared). The PathScore enrichment score in all primary variants was 0.86, suggesting no driver effects (it being close to 1), whereas the enrichment score in all recurrent variants was 2.54, which does suggest a driver effect. When only looking at primary and recurrent variants from an identical set of patients, who had no hypermutation in either the primary or recurrent and received standard therapy, the numbers of variants in the pathway from primary and recurrent tumours was 2/4162 and 9/5056. This is an even larger enrichment of 3.70 in the recurrent tumours ( $P=0.06$ , 1-tailed Fisher's exact test (lower numbers required a different statistical test than above)). These results suggest that disrupted synthesis of bile acids and bile salts may confer a survival advantage to GBM cells undergoing standard treatment. I therefore looked into the variants in this pathway further. I could not find any evidence of gain of function variants although there was generally a lack of evidence to confirm either way (Table 7). A diagram of the REACTOME\_SYNTHESIS\_OF\_BILE\_ACIDS\_AND\_BILE\_SALTS pathway (which is an extension of REACTOME\_SYNTHESIS\_OF\_BILE\_ACIDS\_AND\_BILE\_SALTS\_VIA\_7ALPHA\_HYDROXYCHOLESTEROL) indicated that altered genes all lay along a linear pathway (Figure 33). 5/9 of the variants in the recurrent tumours are in genes in which a loss of function causes significant reductions in bile acid synthesis (*AKR1D1*, *HSD3B7* and *HSD17B4*) (Shea *et al.*, 2007; Lemonde *et al.*, 2003; Huyghe *et al.*, 2006; Fuchs *et al.*, 2001). A further 2 variants, in *ABCB11*, are likely to result in a decrease in export of bile acids and accumulation of intracellular bile acids.

Table 6. Top 20 pathways ranked by recurrent % patients / primary % patients, showing those potentially involved in causing therapy resistance. Pathways were filtered for those with  $\geq 3$  patients being altered by recurrent variants.

Pathway	All primary variants		All recurrent variants		% patients fold change
	Effect size	% patients n=113	Effect size	% patients n=79	
REACTOME_SIGNAL_REGULATORY_PROTEIN_SIRP_FAMILY_INTERACTIONS	0.244	0.9	1.441	5.1	5.67
REACTOME_POL_SWITCHING	0.201	0.9	1.203	5.1	5.67
REACTOME_THROMBOXANE_SIGNALLING_THROUGH_TP_RECEPTOR	0.33	0.9	1.425	3.8	4.22
BIOCARTA_TH1TH2_PATHWAY	0.244	0.9	1.059	3.8	4.22
KEGG_GALACTOSE_METABOLISM	0.272	2.7	0.931	8.9	3.30
REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS_VIA_7ALPHA_HYDROXYCHOLESTEROL	0.855	3.5	2.544	10.1	2.89
REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS	0.653	3.5	1.945	10.1	2.89
REACTOME_TRYPTOPHAN_CATABOLISM	0.557	1.8	1.626	5.1	2.83
REACTOME_IOTROPIC_ACTIVITY_OF_KAINATE_RECEPTORS	0.37	1.8	1.087	5.1	2.83
REACTOME_ACTIVATED_TAK1_MEDIATES_P38_MAPK_ACTIVATION	0.312	1.8	0.927	5.1	2.83
REACTOME_LAGGING_STRAND_SYNTHESIS	0.307	1.8	0.894	5.1	2.83
REACTOME_DIGESTION_OF_DIETARY_CARBOHYDRATE	0.196	1.8	0.563	5.1	2.83
PID_TCR_JNK_PATHWAY	0.616	2.7	1.862	7.6	2.81
KEGG_PRIMARY_BILE_ACID_BIOSYNTHESIS	0.559	2.7	1.636	7.6	2.81
REACTOME_DARPP_32_EVENTS	0.52	2.7	1.561	7.6	2.81
PID_SMAD2_3PATHWAY	0.448	2.7	1.344	7.6	2.81
REACTOME_N_GLYCAN_ANTENNAE_ELONGATION	0.915	3.5	2.421	8.9	2.54
REACTOME_N_GLYCAN_ANTENNAE_ELONGATION_IN_THE_MEDIAL_TRANS_GOLGI	0.644	3.5	1.703	8.9	2.54
BIOCARTA_PGC1A_PATHWAY	0.616	3.5	1.619	8.9	2.54
KEGG_MISMATCH_REPAIR	0.444	3.5	1.159	8.9	2.54

Table 7. Variants in the REACTOME\_SYNTHESIS\_OF\_BILE\_ACIDS\_AND\_BILE\_SALTS\_VIA\_7ALPHA\_HYDROXYCHOLESTEROL in patients with no hypermutation in either the primary or recurrent, and who received standard therapy.

Patient	Variant status	Gene	Chr	Position	Allele	Vaf	Additional information
GLSS-MD-0023	shared primary	ABCB11	2	169783809	C->A	0.504	Within a Q-loop which are involved in ATP-binding.
GLSS-MD-0023	shared recurrent	ABCB11	2	169783809	C->A	0.25	"
GLSS-CU-R006	private recurrent	ABCB11	2	169788951	C->T	0.46	Previously found in a large intestine tumour.
GLSS-MD-0026	private recurrent	AKR1C4	10	5242224	C->A	0.448	Stop gained in exon 4/11.
GLSS-SM-R060	private recurrent	AKR1D1	7	137798468	G->A	0.234	-
GLSS-MD-0025	private recurrent	AKR1D1	7	137790074	G->A	0.059	CTCF binding site - may disrupt gene regulation. Previously found in one prostate and multiple liver tumours.
GLSS-19-0271	private recurrent	HSD17B4	5	118810154	A->G	0.122	Previously found in an oligodendroglioma.
TCGA-06-0190	private recurrent	HSD17B4	5	118877604	->C	0.109	Frameshift - missing end of the protein, which facilitates the transfer of molecules between membranes, and contains the peroxisomal targeting signal.
GLSS-SM-R058	private recurrent	HSD3B7	16	30999380	C->G	0.182	Within a constrained element region.
GLSS-MD-0025	private recurrent	SLC27A5	19	59010520	G->T	0.166	Within a promoter and constrained elements region.
GLSS-LU-00C2	private primary	SCP2	1	53427247	C->T	0.149	-

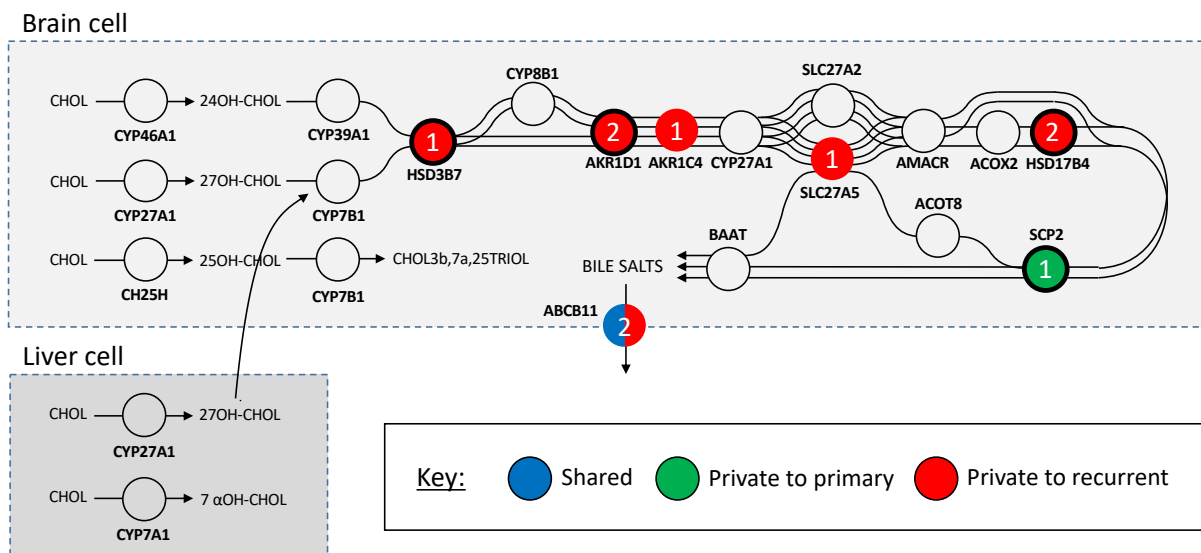


Figure 33. A diagram of the variants in the REACTOME\_SYNTHESIS\_OF\_BILE\_ACIDS\_AND\_BILE\_SALTS pathway in patients with no hypermutation in either the primary or recurrent, and who received standard therapy. Circles indicate all proteins in the gene set, with filled in circles indicating genes that contain variants private to the primary (green), private to the recurrent (red), or in both primary and recurrent (blue), and the number of variants in each gene written on each. Additional pathways involving these genes occur in the liver and other tissues, but here pathways are only shown when relevant to the brain or if a gene does not function in the brain. Altered genes with bold circles indicates that loss of function in these genes has previously been found to significantly reduce bile acid levels.

#### 4.2.6 Pathways enriched in clonally expanding cells

I next looked at whether there was any evidence of pathways driving clonal expansion. For this, I provided PathScore with only variants that were shared between both the primary and recurrent tumours, which provides a background pathway alteration rate for all variants that survived through therapy (Figure 34). I then provided PathScore with only the shared variants that significantly increased in VAF from the primary to the recurrent, thus representing variants in subclones that clonally expanded through therapy (Figure 34A). I repeated this with the shared variants that increased in CCF from the primary to the recurrent (Figure 34B). Both of these, when compared to the results from using all shared variants, allow identification of pathways that are more enriched in clonally expanding subclones and therefore likely to be driving the increased fitness.

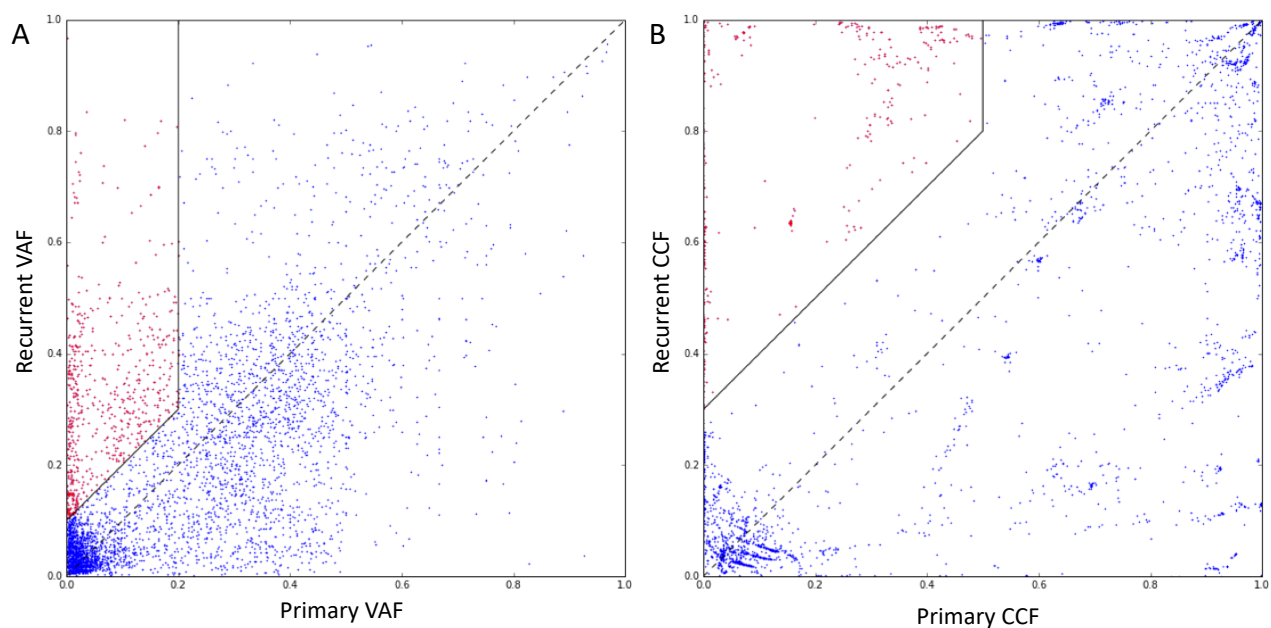


Figure 34. Variants shared between both the primary and recurrent tumours, plotted with A) VAF in the primary against VAF in the recurrent, and B) CCF in the primary against CCF in the recurrent. Red markers indicate variants that were considered to be significantly increased from primary to recurrent.

Pathways were generally altered in fewer patients in the increased CCF variants than the increased VAF variants as around a third of shared variants (1490/4496) did not have a CCF available from the GLASS dataset. Many pathways altered by shared variants were found to be enriched in variants that increase in either VAF (Table 8) or CCF (Table 9).

Table 8. Top 20 pathways ranked by increasing VAF number of patients / shared number of patients, showing those potentially causing clonal expansion. Pathways were filtered for those with  $\geq 3$  patients being altered by increasing VAF variants.

Pathway	Shared variants		Increasing VAF variants		% patients fold change
	Effect size	% patients n=91	Effect size	% patients n=91	
KEGG_VALINE_LEUCINE_AND_Isoleucine_DEGRADATION	0.47	4.4	2.85	3.3	0.75
KEGG_N_GLYCAN_BIOSYNTHESIS	0.5	5.5	2.26	3.3	0.60
REACTOME_ABORTIVE_ELONGATION_OF_HIV1_TRANSCRIPT_IN_THE_ABSENCE_OF_TAT	1.27	5.5	5.54	3.3	0.60
PID_HIF2PATHWAY	0.7	5.5	3.09	3.3	0.60

REACTOME_FORMATION_OF_THE_HIV1_EARLY_ELONGATION_COMPLEX	0.94	5.5	4.1	3.3	0.60
PID_ALK2_PATHWAY	2.66	5.5	12.31	3.3	0.60
KEGG_BUTANOATE_METABOLISM	1.19	8.8	4.78	4.4	0.50
REACTOME_HS_GAG_BIOSYNTHESIS	0.73	6.6	2.59	3.3	0.50
REACTOME_FORMATION_OF_RNA_POL_II_ELONGATION_COMPLEX	0.83	6.6	3	3.3	0.50
KEGG_ALANINE_ASPARTATE_AND_GLUTAMATE_METABOLISM	0.82	6.6	3.01	3.3	0.50
PID_ADISS_2PATHWAY	2.08	26.4	6.02	12.1	0.46
PID_BMP_PATHWAY	0.87	7.7	2.86	3.3	0.43
REACTOME_MRNA_CAPPING	1.56	7.7	4.8	3.3	0.43
REACTOME_TRANSCRIPTIONAL_ACTIVITY_OF_SMAD2_SMAD3_SMAD4_HETEROTRIMER	0.88	7.7	2.72	3.3	0.43
PID_WNT_SIGNALING_PATHWAY	1.18	7.7	3.66	3.3	0.43
REACTOME_GLUCOSE_METABOLISM	0.57	7.7	1.79	3.3	0.43
PID_FOXO_PATHWAY	0.85	7.7	2.67	3.3	0.43
REACTOME_SIGNALING_BY_TGF_BETA_RECEPTOR_COMPLEX	1.01	13.2	3.15	5.5	0.42
BIOCARTA_ERK_PATHWAY	5.95	29.7	14.96	12.1	0.41
KEGG_ADHERENS_JUNCTION	1.97	35.2	4.88	14.3	0.41

Table 9. Top 20 pathways ranked by increasing CCF number of patients / shared number of patients, showing those potentially causing clonal expansion. Pathways were filtered for those with  $\geq 3$  patients being altered by increasing CCF variants. CCF numbers of patients are not directly comparable to those of shared variants due to approximately a third of variants not having an available CCF estimate.

Pathway	Shared variants		Increasing CCF variants		% patients fold change
	Effect size	% patients n=91	Effect size	% patients n=91	
PID_ALK2_PATHWAY	2.66	5.5	18.03	3.3	0.60

REACTOME_SMOOTH_MUSCLE_CONTRACTIO N	1.83	9.9	9.8	4.4	0.44
PID_BMP_PATHWAY	0.87	7.7	4.18	3.3	0.43
REACTOME_MRNA_CAPPING	1.56	7.7	7.05	3.3	0.43
REACTOME_LATE_PHASE_OF_HIV_LIFE_CYCL E	0.59	14.3	2.52	5.5	0.38
KEGG_BUTANOATE_METABOLISM	1.19	8.8	5.32	3.3	0.38
SIG_REGULATION_OF_THE_ACTIN_CYTOSKEL ETON_BY_RHO_GTPASES	0.91	8.8	3.94	3.3	0.38
REACTOME_TRIGLYCERIDE_BIOSYNTHESIS	1.08	8.8	4.12	3.3	0.38
REACTOME_HIV_LIFE_CYCLE	0.61	15.4	2.29	5.5	0.36
REACTOME_TRANSPORT_TO_THE_GOLGI_AN D_SUBSEQUENT_MODIFICATION	1.31	9.9	4.39	3.3	0.33
PID_AURORA_B_PATHWAY	0.96	9.9	3.65	3.3	0.33
KEGG_MELANOGENESIS	1.05	20.9	3.27	6.6	0.32
KEGG_ADHERENS_JUNCTION	1.97	35.2	5.29	11.0	0.31
PID_CMYB_PATHWAY	0.96	17.6	3.18	5.5	0.31
BIOCARTA_CHREBP2_PATHWAY	1.31	11.0	3.92	3.3	0.30
BIOCARTA_AMI_PATHWAY	1.42	11.0	4.59	3.3	0.30
REACTOME_RNA_POL_II_TRANSCRIPTION	0.89	15.4	2.65	4.4	0.29
KEGG_TGF_BETA_SIGNALING_PATHWAY	0.87	15.4	2.64	4.4	0.29
REACTOME_PKB_MEDIATED_EVENTS	1.73	12.1	5.12	3.3	0.27
REACTOME_ASPARAGINE_N_LINKED_GLYCOS YLATION	0.99	16.5	2.64	4.4	0.27

The top result when ranking pathways by number of patients altered by increased VAF variants divided by number of patients altered by all shared variants, was KEGG\_VALINE\_LEUCINE\_AND\_Isoleucine\_DEGRADATION. This was altered by 6/4496 shared variants across 4/91 patients (4.4%), of which 4/613 variants across 3/91 patients (3.3%) had a significantly increased VAF, and 3/423 variants across 2/91 patients (2.2%) had a significantly increased CCF. (These numbers differ slightly to those in Table 8 and 9 as some genes were rejected by Pathscore in its calculations, due to a lack of valid gene identifiers. The values stated here reflect the true numbers of variants in the pathway.) This is a 4.9 fold enrichment in variants



that increase in VAF (P=0.0238, 1-tailed Fisher's exact test), and a 3.6 fold enrichment in variants that increase in CCF (P=0.0888, 1-tailed Fisher's exact test). Shared altered genes in this pathway are listed in Table 10.

Table 10. Genes altered in the KEGG\_VALINE\_LEUCINE\_AND\_Isoleucine\_DEGRADATION pathway in variants shared between primary and recurrent tumours. Those that significantly increased in VAF are shaded.

Gene	VAF change	CCF change
<i>BCAT1</i>	0.096->0.281	0.154->0.634
<i>OXCT1</i>	0.081->0.267	0.154->0.634
<i>HMGCS2</i>	0.015->0.347	0.0002->1.00
<i>OXCT2</i>	0.086->0.218	0.310->0.422
<i>ACAT2</i>	0.654->0.431	0.966->0.587
<i>ECHS1</i>	0.061->0.057	0.040->0.069

Another pathway with enrichment in increasing VAF variants is KEGG\_BUTANOATE\_METABOLISM. This pathway overlaps with KEGG\_VALINE\_LEUCINE\_AND\_Isoleucine\_DEGRADATION and contains the same set of variants as that pathway, other than *BCAT1*, plus additional variants. This consists of 9/4496 shared variants across 9/91 patients (9.9%), of which 4/613 variants across 4/91 patients (4.4%) had a significantly increased VAF and 3/423 variants across 3/91 patients (3.3%) had a significantly increased CCF. Due to the far higher numbers of background variants that did not significantly increase in frequency, this is a 3.3 fold enrichment in variants that increase in VAF (P=0.0606, 1-tailed Fisher's exact test), and a 2.7 fold enrichment in variants that increase in CCF (P=0.1451, 1-tailed Fisher's exact test). Shared altered genes in this pathway are listed in Table 11.

Table 11. Genes altered in the KEGG\_BUTANOATE\_METABOLISM pathway in variants shared between primary and recurrent tumours. Those that significantly increased in VAF are shaded.

Gene	VAF change	CCF change
<i>GAD2</i>	0.164->0.701	0.387->0.984
<i>OXCT1</i>	0.081->0.267	0.154->0.634
<i>HMGCS2</i>	0.015->0.347	0.0002->1.00
<i>OXCT2</i>	0.086->0.218	0.310->0.422
<i>ACAT2</i>	0.654->0.431	0.966->0.587
<i>ECHS1</i>	0.061->0.057	0.040->0.069
<i>PDHA1</i>	0.814->0.280	0.862->0.197
<i>ACSM3</i>	0.407->0.267	0.702->0.882
<i>ACSM5</i>	0.454->0.234	NA

The top result when ranking pathways by number of patients altered by increased CCF variants divided by number of patients altered by all shared variants was PID\_ALK2\_PATHWAY. This pathway was altered by

5/4496 shared variants across 5/91 patients (5.5%), and by 3/613 and 3/423 significantly increasing VAF and CCF variants across 3/91 patients (3.3%). This is a 4.4 fold enrichment in variants that increase in VAF (P=0.0606, 1-tailed Fisher's exact test), and a 5.3 fold enrichment in variants that increase in CCF (P=0.0446, 1-tailed Fisher's exact test). Shared altered genes in this pathway are listed in Table 12.

Table 12. Genes altered in the PID\_ALK2\_PATHWAY pathway in variants shared between primary and recurrent tumours. Those that significantly increased in CCF are shaded.

Gene	VAF change	CCF change
SMAD5	0.012->0.285	0.005->0.439
BMP2R	0.010->0.408	0.0002->1.00
AMHR2	0.118->0.321	0.259->0.657
AMHR2	0.333->0.330	0.619->0.992
AMHR2	0.716->0.584	NA

#### 4.2.7 Pathways under-enriched in cells surviving through therapy

I next wanted to see if there was evidence of pathways potentially causing therapy sensitivity. We expect to see a subclone wiped out through therapy if it contained an alteration to a pathway that, when altered, confers sensitivity. I therefore looked for pathways that decreased significantly from 'private primary' variants to those shared between primary and recurrent. Many pathways did show this pattern (Table 14), with the majority being altered by *IDH1* or *IDH2* variants. Whilst all included tumours in the analysis were classed as *IDHwt* in the GLASS dataset, despite this, many of them have low frequency *IDH1* and *IDH2* variants; 33/1492 (14 *IDH1*, 19 *IDH2*) private to primary variants across 28/91 patients (30.8%), and 2/4496 shared variants across 2/91 patients (2.2%). Interestingly, there were 50 *IDH1* or *IDH2* variants private to the recurrent across 28/79 patients (35.4%), with 7/79 (8.9%) having at least 2-4 separate variants in the same tumour, showing that new variants in these genes commonly develop once therapy has stopped. All but 1 of these *IDH* variants, across primary and recurrences, were at the canonical R132 (*IDH1*) and R172 (*IDH2*) positions causing gain of function (Dang *et al.*, 2009). The majority of these were only supported by 1 read and all had VAFs under 0.1. It's therefore possible that some of them are sequencing errors, although the fact that they're mostly at canonical positions and no other genes were altered to such an extent, would suggest otherwise. Therefore, these results show that GBM tumours classed as *IDHwt* regularly have subclonal canonical *IDH* variants and that those clones rarely survive through therapy.

Table 14. Top 20 pathways ranked by private to primary % patients / shared % patients, showing those potentially involved in causing therapy sensitivity. Pathways were filtered for those with  $\geq 3$  patients being altered by private to primary variants. Pathways involving *IDH1* or *IDH2* are shaded.

Pathway	Private to primary variants		Shared variants		% patients fold change
	Effect size	% patients n=91	Effect size	% patients n=91	
BIOCARTA_TOB1_PATHWAY	3.86	3.3	0	0	0.00
REACTOME_TIGHT_JUNCTION_INTERACTIONS	2.27	3.3	0	0	0.00
REACTOME_PEROXISOMAL_LIPID_METABOLISM	9	13.2	0.24	1.1	0.08
KEGG_CITRATE_CYCLE_TCA_CYCLE	15.24	31.9	0.93	6.6	0.21
KEGG_GLUTATHIONE_METABOLISM	15.8	31.9	0.97	6.6	0.21
BIOCARTA_KREB_PATHWAY	27.59	19.8	1.89	4.4	0.22
REACTOME_PYRUVATE_METABOLISM_AND_CITRIC_ACID_TCA_CYCLE	12.77	31.9	0.92	7.7	0.24
KEGG_PEROXISOME	7.8	33	0.63	8.8	0.27
REACTOME_CITRIC_ACID_CYCLE_TCA_CYCLE	13.85	19.8	1.19	5.5	0.28
REACTOME_INHIBITION_OF_INSULIN_SECRETION_BY_ADRENALINE_NORADRENALINE	2.99	3.3	0.34	1.1	0.33
REACTOME_ACYL_CHAIN_REMODELLING_OF_PG	3.97	3.3	0.45	1.1	0.33
REACTOME_OXYGEN_DEPENDENT_PROLINE_HYDROXYLATION_OF_HYPOXIA_INDUCIBLE_FACTOR_ALPHA	5.03	3.3	0.58	1.1	0.33
REACTOME_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_TRANSPORT	6.47	31.9	0.93	14.3	0.45
BIOCARTA_FMLP_PATHWAY	1.7	4.4	0.28	2.2	0.50
KEGG_SELENOAMINO_ACID_METABOLISM	2.69	4.4	0.45	2.2	0.50
REACTOME_BILE_SALT_AND_ORGANIC_ANION_SLC_TRANSPORTERS	5.91	4.4	0.99	2.2	0.50
REACTOME_PYRIMIDINE_METABOLISM	3.56	5.5	0.69	3.3	0.60
REACTOME_COMPLEMENT_CASCADE	1.69	5.5	0.34	3.3	0.60
REACTOME_DOPAMINE_NEUROTRANSMITTER_RELEASE_CYCLE	3.4	3.3	0.74	2.2	0.67
REACTOME_NOREPINEPHRINE_NEUROTRANSMITTER_RELEASE_CYCLE	3.76	3.3	0.82	2.2	0.67

Interestingly, patients with *IDH* alterations in recurrent tumours ( $P=0.054$ , log-rank statistic), but not those with *IDH* alterations in primary tumours ( $P=0.56$ , log-rank statistic), had a shorter survival. The difference in survival was also seen when increasing the sample size by 2 ( $P=0.037$ , log-rank statistic), by including patients who had an IDHwt oligodendroglioma or astrocytoma as a primary that changed to a GBM in the recurrent after receiving standard therapy (Figure 35).

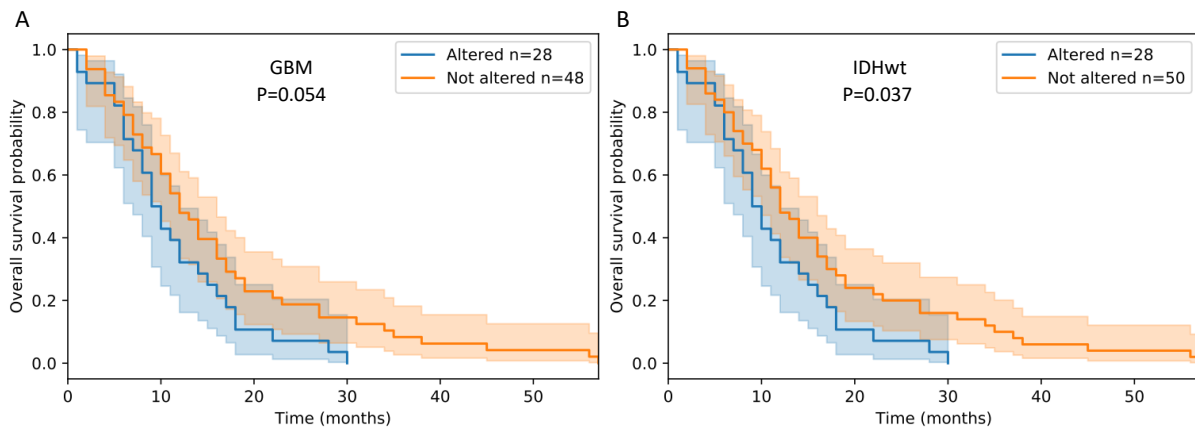


Figure 35. Kaplan-Meier plot showing the difference in survival between patients with and without subclonal *IDH1* or *IDH2* alterations (though those ‘with’ alterations had them at insufficient levels to be classified as IDHwt), in the recurrent GBM tumour for patients whose primary tumour was A) a GBM ( $n=76$ ), or B) any IDHwt high grade glioma ( $n=78$ ).

*IDH* mutant tumours are shown to respond better to TMZ and radiotherapy (Christians *et al.*, 2019).

Therefore, a possible explanation for the shorter survival in patients with recurrent subclonal *IDH* alterations is that they represent those who had less therapy due to being too unwell and therefore able to accumulate the alterations. To test this, I performed a Cox proportional hazards model, taking into account whether a patient had an *IDH* variant in their recurrent tumour, and also number of TMZ chemotherapy cycles they underwent between the primary and recurrent surgeries (for 64 patients that had this information available). This revealed that the presence of an *IDH* mutation was not associated with a shorter survival when, and only when taking into account number of TMZ cycles (Hazard ratio = 0.94, 95% confidence interval 0.53-1.66,  $P=0.82$ ). Additionally, the presence of *IDH* variants in the recurrent tumour was negatively correlated with the number of TMZ cycles (ordinary least squares statistic, coefficient = 4.56, 95% confidence interval 2.236-6.884,  $P = 2.2 \times 10^{-4}$ ).

Two other pathways not involving *IDH* stood out. Both BIOCARTA\_TOB1\_PATHWAY and REACTOME\_TIGHT\_JUNCTION\_INTERACTIONS were altered in 3/1492 genes across 3/91 primary tumours (3.3%), but none were shared with the recurrent. These differences in the numbers of altered genes between primary and shared variants had an uncorrected P-value of 0.0154 (1-tailed Fisher’s exact test) for both pathways. BIOCARTA\_TOB1\_PATHWAY was altered in genes involved specifically in transforming

growth factor beta (TGF $\beta$ ) signalling; *TGF $\beta$ 1* (VAF=0.234), *TGF $\beta$ 2* (VAF=0.218), *TGF $\beta$ 3* (VAF=0.288). REACTOME\_TIGHT\_JUNCTION\_INTERACTIONS was altered in 3 variants private to the primary. This increased to 4/14752 genes across 4/94 patients (4.3%) when also looking at hypermutated primary tumours, with no genes altered in the shared variants. The primary altered genes included: *CLDN8* (VAF=0.143), *PARD3* (VAF=0.089) and 2 in *MPP5* (VAF=0.156, 0.02). In patients that did not receive both radiotherapy and TMZ, 1/23 (4.3%) patient had a private to primary variant in this pathway, and 2/23 (8.7%) had a shared variant. This is a significant difference in number of patients with shared variants in the pathway, between patients who had and hadn't received standard treatment (P=0.0373, 1-tailed Fisher's exact test), and which does not require multiple testing correction.

## 4.3 Discussion

### 4.3.1 Summary

In this chapter, I used longitudinal GBM samples to show that strong selection is present in a minority of recurrent GBMs and other gliomas, but that it is not significantly associated with therapy. Despite this, by carrying out pathway analysis using subsets of variants that are informative of GBM progression, I found evidence of many pathways that may confer therapy resistance or sensitivity, and which are candidates for further investigation.

### 4.3.2 Investigating the mode of tumour evolution

The SubClonalSelection model predicted that 37% of all recurrent gliomas, and 47% of IDHwt recurrent gliomas, are evolving with strong selection of some subclones over others. This is a higher proportion of selection than was found in colon cancers from The Cancer Genome Atlas (21%) and gastric cancers (29%), lower than in non-small-cell lung cancer from the TRACERx cohort (97%), and comparable to mixed metastatic tumours in the MET500 cohort (51%) (Williams *et al.*, 2018).

This selection detected in this analysis was not associated with patients receiving therapy and is therefore likely to be due to other factors in the majority of cases. However, it is possible that scenarios where genetic mutations do confer increased therapy resistance, were not detected by the SubClonalSelection model for several reasons:

- i) When there is insufficient evidence to reliably predict the mode of evolution due to low sequencing depth, insufficient numbers of subclonal variants, or when the relative fitness of a clone is not large, the model defaults towards neutral.
- ii) Evidence of selection could become lost over time between cessation of therapy and biopsy of the recurrence, which is plausible given that resistance mechanisms will likely hinder subclone

growth in the absence of therapy compared to other clones (Yamamoto, Tomiyama, *et al.*, 2018; Enriquez-Navas *et al.*, 2015).

iii) Genetic alterations can confer resistance to surrounding clones through non-autonomous effects (Inda *et al.*, 2010), thereby maintaining the current subclonal architecture without selection of specific clones.

iv) Tumours undergoing a clonal sweep whereby a clone expanded to the extent that it appears clonal, and then proceeded to evolve neutrally, would not be detected as selection by the model (Bozic *et al.*, 2019). Evidence for this latter scenario is apparent in multiple patients when comparing VAFs and CCFs between primary and recurrent tumours. The recurrent samples from those patients with the clearest evidence of this occurring were either classed as showing selection anyway (likely from subsequent additional expansions) or had insufficient data for use with the model.

v) Sampling bias can have a large effect on the detection of subclones, resulting in reduced power to detect selection in single samples (Sun *et al.*, 2017; Siegmund and Shibata, 2016; Chkhaidze *et al.*).

Nonetheless, the results do support previous observations that factors other than genetic mechanisms are involved in driving the therapy resistance seen in gliomas (Eyler *et al.*, 2020; Körber *et al.*, 2019). This is also supported by the observation that the majority of shared variants are subclonal in both the primary and recurrent tumours, a scenario that is only possible in oligoclonal progression where multiple subclones survive through therapy, with a tendency to remain at a similar frequency. Evidence of oligoclonal progression in most GBM patients has been previously reported (Körber *et al.*, 2019). Further support for a general lack of variants driving selection through therapy in glioma has been shown via the dN/dS method, where the global ratio of non-synonymous to synonymous variants, does not show an increase for those private to the recurrence, particularly in IDHwt gliomas (Barthel *et al.*, 2019). Although, given that this ratio is averaged across all variants private to the recurrent tumours, it's still possible that a small minority of these variants are driving selection, but that the signal is diluted by many more that are not.

### 4.3.3 Pathway analysis

The emerging evidence supporting adaptive epigenetic reprogramming of cell states as driving GBM therapy resistance, is also accompanied by evidence that highlights the influence of genetic factors. Recent studies using single-cell RNAseq in GBM show that some somatic profiles predispose cells to certain cellular states, of which some survive therapy better than others (Neftel *et al.*, 2019). Others show that epigenetic changes through therapy can be accompanied by genetic factors that further increase the cell's resistance, either before or after the initial onset of therapy (Eyler *et al.*, 2020). It is therefore important to investigate

which specific gene or pathway alterations may be influencing cell survival through therapy, by investigating patterns of genetic alterations across longitudinal GBM samples. Even if genetic factors influence resistance in only a small minority of patients, or just cause pockets of treatment resistance or sensitivity that are missed when looking for evidence of strong selection across a full tumour profile, they may allow us to learn about processes that affect therapeutic response.

By using subsets of variants and comparing results between them, I was able to use pathway analysis to identify genes and pathways that may influence therapy resistance in GBM. This builds on previous studies that found a limited number of recurrence specific single gene (Kraboth and Kalman, 2020; Barthel *et al.*, 2019; Körber *et al.*, 2019; Wang *et al.*, 2016; Cahill *et al.*, 2007), and highlights the benefit of using a systems biology approach to pool information across multiple genes a pathway.

#### 4.3.4 Pathways differentially enriched between primary and recurrent tumours

Pathways that are significantly more enriched in the recurrent than primary tumours are candidates for conferring therapy resistance. The pathway that stood out most in this study as showing this pattern was REACTOME\_SYNTHESIS\_OF\_BILE\_ACIDS\_AND\_BILE\_SALTS\_VIA\_7ALPHA\_HYDROXYCHOLESTEROL. Bile synthesis occurs through several overlapping pathways, each via differing oxysterols (McMillin and DeMorrow, 2016). In the brain, these include 24S-hydroxycholesterol (24OH-CHOL), which is synthesised *de novo* from cholesterol, and 27-hydroxycholesterol (27OH-CHOL), which is mostly imported from outside the central nervous system (Heverin *et al.*, 2005). Synthesis via 7alpha-hydroxycholesterol (7 $\alpha$ OH-CHOL) is not thought to occur in the brain, however the altered genes found in this pathway in the recurrent tumours are also involved in synthesis of bile acids and salts from 24OH-CHOL and 27OH-CHOL. Previous studies provide multiple lines of evidence as to how this alters GBM biology; 1) Bile acids affect neurotransmission and regulate neurological function (McMillin and DeMorrow, 2016; Kiriya and Nochi, 2019). GBM cells have been found to form synapses and integrate into neuronal circuits to promote proliferation, and will therefore likely be affected by changes in neuron functioning (Venkatesh *et al.*, 2019; Venkataramani *et al.*, 2019). 2) Bile acids activate numerous nuclear receptors, including farnesoid X, pregnane X, vitamin D, constitutive androstane, glucocorticoid, M2 and M3 muscarinic, formyl-peptide, sphingosine-1-phosphate receptor 2, Takeda G-protein 5 receptors, and also activate several ion channels (McMillin and DeMorrow, 2016; Kiriya and Nochi, 2019), which could bring about changes in gene expression or reprogramming of cell states, to allow GBM cells to better resist therapy. 3) Bile acids increase cellular cholesterol levels in the brain (McMillin *et al.*, 2018) :

Cholesterol is highly implicated in many types of cancer, including in GBM where feedback mechanisms are altered to increase intracellular cholesterol levels to fuel the cells' growth (Ahmad *et al.*, 2019). Bile is synthesised from cholesterol via the oxysterols, 24OH-CHOL and 27OH-CHOL. In healthy brain cells,

conversion of cholesterol to oxysterols, primarily to 24OH-CHOL by cytochrome P450 family 46 subfamily A member 1 (Cyp46A1) (Russell *et al.*, 2009), is the main route of clearing excess cholesterol from the brain (Figure 33), and is regulated by a negative feedback loop (Figure 36); the produced oxysterols i) inhibit sterol regulatory element-binding protein 1 (SREBP1) activity, causing a reduction in expression of low density lipoprotein receptor (*LDLR*) which imports exogenous cholesterol into cells, and ii) activate liver X receptor (LXR) to increase expression of genes responsible for cholesterol efflux and degradation of *LDLR* (Han *et al.*, 2020). In GBM cells, Cyp46A1 is downregulated leading to a ten-fold reduction in 24OH-CHOL (Villa *et al.*, 2016) compared to normal brain. This results in increased cholesterol import, reduced efflux from cells, and a reduced clearance through conversion to 24OH-CHOL (Villa *et al.*, 2016; Han *et al.*, 2020). In addition, GBM cells increase cholesterol intake via epidermal growth factor receptor (EGFR) activated, PI3K/SREBP-1-dependent upregulation of *LDLR* (Guo *et al.*, 2011). There are conflicting reports on whether GBM cells upregulate or downregulate *de novo* cholesterol synthesis through the mevalonate pathway (Villa *et al.*, 2016; Patel *et al.*, 2019; Kambach *et al.*, 2017). The 24OH-CHOL that does get produced in GBM cells can be converted into bile acids. Bile acids have a positive feedback mechanism to increase intracellular cholesterol further by downregulating Cyp46A1 via farnesoid X receptor (FXR) activation (McMillin *et al.*, 2018) and upregulating *LDLR* via activation of mitogen-activated protein kinase cascades (Nakahara *et al.*, 2002) (Figure 36).

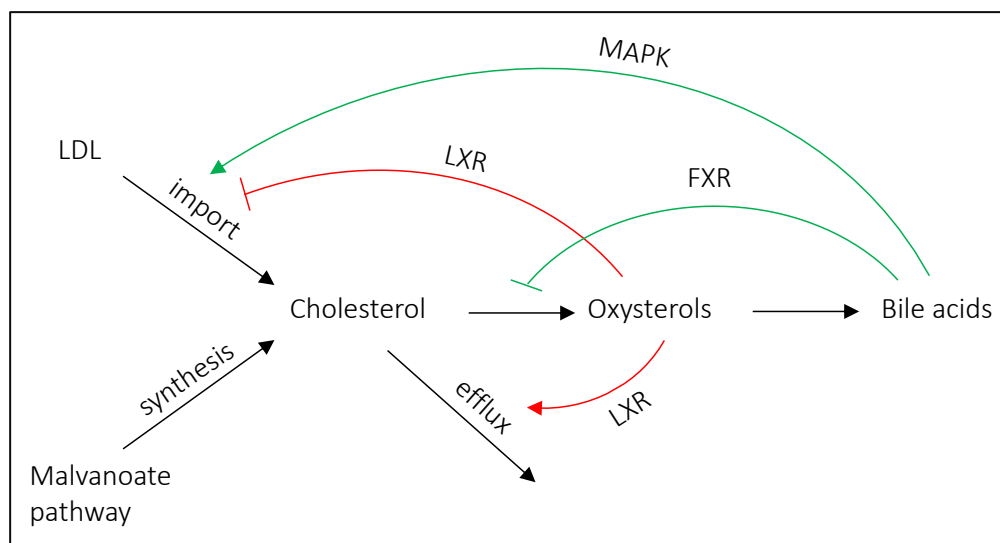


Figure 36. Key processes that regulate intracellular cholesterol in healthy cells. Green lines indicate processes that positively regulate cholesterol levels, and red lines indicate processes that negatively regulate cholesterol levels.

Targeting cholesterol metabolism in GBM has proven to be a promising area for potential new treatments. An LXR agonist resulted in decreased cholesterol and death in GBM cells, while leaving healthy brain cells unharmed (Villa *et al.*, 2016). An anti-HIV drug, Efavirenz, that activates CYP46A1 leading to an increase in 24OH-CHOL and decrease in cholesterol, inhibited GBM growth *in vivo* (Han *et al.*, 2020). However, whilst



lowering cholesterol promotes GBM cell death, a study found that intracellular cholesterol levels are lower in TMZ resistant GBM cells and higher in sensitive cells, and that treating either sensitive or resistant cells to modulate cellular cholesterol levels resulted in changes in TMZ induced apoptosis in a similar manner. This was shown to be due to both cholesterol and TMZ upregulating death receptor 5 (*DR5*), potentially from endoplasmic reticulum stress. Furthermore, a single treatment with TMZ increased intracellular cholesterol, thereby increasing pressure from cholesterol on cells that have reduced ability to clear the excess (Yamamoto, Tomiyama, *et al.*, 2018). Interestingly, many studies have shown that statins, which block *de novo* cholesterol synthesis along with many other anti-tumour effects (Pisanti *et al.*, 2014), applied at very high doses increase GBM cell death in combination with TMZ (Yamamoto, Sasaki, *et al.*, 2018), however when used at only physiological doses, statins decreased the effect of TMZ (Yamamoto, Tomiyama, *et al.*, 2018). Together, this suggests that, while important to GBM survival, cholesterol positively regulates TMZ response, and decreasing cholesterol levels through a reduction of bile acids may enable the cells to better resist therapy. Therefore, further investigations into the effects of bile acids on intracellular cholesterol and survival in GBM cells is warranted.

#### 4.3.5 Pathways enriched in clonally expanding cells

Many pathways are implicated when focussing only on those shared mutations that increase in frequency after treatment and which are likely in clonally expanding cells, though these pathways were altered in only a small fraction of patients. In general the numbers of increased VAF variants in pathways agreed with the numbers of increased CCF variants, therefore suggesting that using VAFs to indicate clonally expanding subclones is appropriate, whilst including more variants than using CCFs (which are absent for around a third of variants).

One pathway that showed evidence of causing clonal expansion was KEGG\_VALINE\_LEUCINE\_AND\_ ISOLEUCINE\_DEGRADATION, which is involved in the breakdown of branched chain amino acids. Altered metabolism of branched chain amino acids has been found in many cancers and plays a key role in proliferation and aggressiveness, including in GBM (Prabhu *et al.*, 2019; Suh *et al.*, 2019). One altered gene in this pathway, 3-oxoacid CoA-transferase 1 (*OXCT1*), has previously been found to be downregulated in GBM (Chang *et al.*, 2013; Vallejo *et al.*, 2019), thereby supporting the possible significance of this pathway. However, upregulation of another altered gene in the pathway, branched chain amino acid transaminase 1 (*BCAT1*), promotes GBM proliferation, and suppression of the gene blocked secretion of toxic glutamate from cells leading to inhibited GBM growth (Tönjes *et al.*, 2013). It's possible that the variant in this gene is a gain of function, though I could not find any evidence to determine this. Alternatively, the reason for variants in this pathway being enriched amongst clonally expanded cell populations may be due to its overlap with KEGG\_BUTANOATE\_METABOLISM, another significantly altered pathway in cell populations expanded after treatment, which contains all the altered variants of KEGG\_VALINE\_LEUCINE\_AND\_

ISOLEUCINE\_DEGRADATION, other than *BCAT1*, plus additional ones. Sodium butanoate has previously been shown to induce senescence and inhibit invasion of GBM cells (Nakagawa *et al.*, 2018), thereby supporting the hypothesis that altering butanoate metabolism can lead to subclonal expansion, and further indicates that the methods in this analysis are identifying pathways truly relevant to GBM progression and which may be worth further investigating.

Another pathway found to be enriched in clonally expanded clones is PID\_ALK2\_PATHWAY. This is involved in regulating transforming growth factor beta (TGF $\beta$ )/bone morphogenetic protein (BMP) signalling, and is made up of a subset of genes, specifically involving BMP, in a larger interconnected network of TGF $\beta$  signalling. These processes are tightly balanced and highly tumour specific, with numerous and opposing roles in cell cycle regulation, differentiation, motility, autophagy, T cell activation, therapy resistance, and activation of the apoptotic pathway (Ikushima and Miyazono, 2010; Massagué *et al.*, 2000; Caja *et al.*, 2015). BMP signalling in particular induces epithelial-mesenchymal transition (EMT) in tumours (Tan *et al.*, 2015; Beck *et al.*, 2016), which leads to more stem like cell types, with increased therapy resistance and suppression of cancer cell proliferation (Zheng *et al.*, 2015; Fischer *et al.*, 2015; Shibue and Weinberg, 2017), and is associated with a worsened prognosis in GBM patients (Phillips *et al.*, 2006). In glioma stem cells, treatment with BMP, or activation of the BMP signalling pathway, inhibits cell proliferation and invasiveness, induces differentiation and apoptosis, increases DNA repair processes, and confers resistance to radiotherapy and TMZ chemotherapy (Sachdeva *et al.*, 2019; Xi *et al.*, 2017; Raja *et al.*, 2017; Nayak *et al.*, 2020). Of the three variants that increased in frequency, 2 of them (in *SMAD5* and bone morphogenetic protein receptor type 2 (*BMPR2*)) are likely to reduce EMT, reduce therapy resistance, and increase proliferation, whereas the other (in anti-Mullerian hormone receptor Type 2 (*AMHR2*), which was also the gene affected by the 2 variants in the pathway that did not increase in frequency) is likely to increase EMT, increase therapy resistance, and decrease proliferation (Beck *et al.*, 2016), and therefore may just reflect noise in the analysis.

I choose to investigate which pathways are enriched specifically in clonally expanding cells from primary to recurrent tumours, under the hypothesis that these represent cells that were able to better resist therapy than other subclones and therefore a larger proportion survived through to the recurrent. However, the results suggest that this may not be the case. Alterations to the above two pathways are most likely to result in increased proliferation, explaining their enrichment in clonally expanding cells, but this reduced senescence is also associated with reduced therapy resistance (Tomicic and Christmann, 2018). Therefore, in contrary to the previous assumption that variants in clonally expanding cells indicate candidates for those driving therapy resistance, these variants may actually be increasing in frequency solely because they increase proliferation, perhaps more so in the recurrent tumour than the primary as there is less competition from surrounding clones after surgical resection. Whilst this reduces therapy resistance, and therefore these clones would otherwise be expected to reduce in frequency, the results suggest that the

selective effect of increased proliferation overpowers the negative selection from therapy, once therapy is stopped and the tumour regrows. It may also, therefore, be that variants that confer therapy resistance are in clonally reduced cells, as a result of reduced proliferation, (Suvà and Tirosh, 2020; Liao *et al.*, 2017), though it will likely be challenging to distinguish those from variants that decrease in frequency simply due to deleterious effects on cells.

Though unlikely to be conferring therapy resistance, the pathways identified as enriched in clonally expanding cells may still be of interest in developing novel treatments for GBM patients. While TMZ and radiotherapy aim to kill cancer cells, drugs that instead prevent the tumour from growing back so quickly may provide a way to extend survival in patients, particularly if it could offer a less unpleasant treatment for patients to switch to when TMZ or radiotherapy can no longer be tolerated.

#### 4.3.6 Pathways under-enriched in cells surviving through therapy

When looking for pathways potentially causing therapy sensitivity, evidenced by an under enrichment of genes with shared variants compared to private to primary variants, many pathways initially stood out. However, after investigating these further I found that they were mostly due to just two genes; *IDH1* and *IDH2*. Around a third of both primary and recurrent tumours contained subclonal canonical gain of function alterations in these genes. Such variants are commonly found in several types of cancers, including in 70% of World Health Organisation (WHO) grade II and III astrocytomas and oligodendrogliomas, and in GBMs that developed from these (Yan *et al.*, 2009; Hartmann *et al.*, 2009; Watanabe *et al.*, 2009; Parsons *et al.*, 2008). However, they are generally thought to be near absent in *de novo* GBMs (that did not originate from lower grade gliomas). This may be the case for clonal *IDH* variants, however this study shows that this is not true for low frequency subclonal *IDH* variants. These results also contrast previous studies that found no evidence of repeatedly altered genes in recurrent GBMs (Körber *et al.*, 2019), possibly because others did not include such low frequency variants.

*IDH1* and *IDH2* genes code for isocitrate dehydrogenase 1 and 2 enzymes, present in the cytoplasm and mitochondria respectively, that function in the citric acid cycle as well as maintaining the cellular redox state. Their primary function in healthy cells is the oxidative decarboxylation of isocitrate (ICT) with nicotinamide adenine dinucleotide phosphate (NADP<sup>+</sup>), into  $\alpha$ -ketoglutarate ( $\alpha$ -KG), CO<sub>2</sub> and the reduced form of NADP<sup>+</sup>, NADPH. In hypoxic conditions this process is reversed due to a lack of inhibition by ICT (Wise *et al.*, 2011). The lack of ICT also allows an additional reductive reaction by *IDH1* and *IDH2* to convert  $\alpha$ -KG and NADPH into 2-hydroxyglutarate (2HG) and NADP<sup>+</sup> (Wise *et al.*, 2011). Alteration of certain positions in the proteins cause a loss of inhibition by ICT due to a change in their binding sites (Bascur *et al.*, 2019). These positions typically include R132 and R140 in *IDH1*, and R172 in *IDH1*, although other positions are also capable of causing the same gain of function affect (Ward *et al.*, 2012). The loss of inhibition by ICT causes an increase in the reductive reactions of IDH, resulting in a 10-100 fold increase in cellular 2HG

(Ward *et al.*, 2012, 2010; Dang *et al.*, 2009). 2HG inhibits the activity of  $\alpha$ -KG-dependent dioxygenases, resulting in multiple cellular effects (Ye *et al.*, 2018), most noticeably, the introduction of genome-wide histone alterations and DNA hypermethylation (Xu *et al.*, 2011) thereby preventing cellular differentiation (Lu *et al.*, 2012). While the *IDH* alterations were of low frequency, there is some suggestion that *IDH* variants may enable nonautonomous increases of 2HG in surrounding cells (Wouters, 2017; Amary *et al.*, 2011) and therefore have a stronger influence on the overall tumour growth.

*IDH* variants have also been found to increase sensitivity to both TMZ and radiotherapy, through NADPH depletion, and NAD<sup>+</sup> depletion from 2HG independent downregulation of nicotinate phosphoribosyltransferase (*NAPRT1*) expression and inhibition of nicotinamide phosphoribosyltransferase (*NAMPT*) (Tateishi *et al.*, 2015, 2017; Gujar *et al.*, 2016; Molenaar *et al.*, 2018; Bleeker *et al.*, 2010; Wahl *et al.*, 2017). '*IDH* mutant' gliomas are also known to result in significantly better patient prognosis, especially in patients receiving radiotherapy or TMZ (Christians *et al.*, 2019; SongTao *et al.*, 2012; Houillier *et al.*, 2010; Taal *et al.*, 2011; Okita *et al.*, 2012). This explains the repeated loss of *IDH* variant containing subclones through therapy seen in this study, as well as the association between *IDH* variants and fewer TMZ cycles.

The BIOCARTA\_TOB1\_PATHWAY is a pathway not involving *IDH1* or *IDH2*, that showed potential sensitising effects when altered. The pathway covers the role of transducer of ERBB2 (Tob) in T-cell activation. However, the affected genes in the primary (TGF $\beta$ 1, TGF $\beta$ 2, TGF $\beta$ 3) were specifically involved in TGF $\beta$  signalling. As mentioned above, TGF $\beta$  signalling has numerous roles in cancer, covering a large and complex network of processes. In the part of the network relevant to the altered genes identified here, TGF $\beta$  ligands, TGF $\beta$ 1, TGF $\beta$ 2 and TGF $\beta$ 3, bind to the receptor TGF $\beta$ R1, causing it to heterodimerise with TGF $\beta$ R2 (Cheifetz *et al.*, 1987; Massagué *et al.*, 2000). The formed complex then initiates cellular changes through phosphorylation of SMAD2 and SMAD3 transcription factors (Massagué, 1998; Massagué *et al.*, 2000). TGF $\beta$ R3 augments binding of TGF $\beta$  ligands to TGF $\beta$ R1 and TGF $\beta$ R2, but is also involved in an additional non-canonical and non-redundant signalling pathway through interaction with the GIPC1 scaffolding protein (Sánchez *et al.*, 2011). TGF- $\beta$  is known to induce EMT (Hao *et al.*, 2019), and disruption of signalling through TGF $\beta$  receptor inhibitors has been shown to increase GBM sensitisation to radiotherapy (Liu *et al.*, 2016) and improve survival *in vivo* (Zhang *et al.*, 2011). This supports the observation in this study that inhibition of the TGF $\beta$  signaling pathway prevents cells from surviving through therapy, and again provides confidence in the method's ability to yield findings worth further investigation. However, a phase 2 clinical trial of the TGF $\beta$ R1 inhibitor galunisertib in combination with TMZ and radiotherapy, did not show any increase in survival compared to standard therapy alone (Wick *et al.*, 2020). Additionally, in glioma stem cells, pre-treatment with TGF- $\beta$  significantly increased the amount of cell death that occurred with TMZ and radiotherapy, as well as an increase in cell proliferation (Sachdeva *et al.*, 2019).

Another pathway that stood out was REACTOME\_TIGHT\_JUNCTION\_INTERACTIONS. Tight junctions are protein complexes involved in cell–cell adhesion and cell polarity, two features that are lost during EMT (Shibue and Weinberg, 2017; Wodarz and Näthke, 2007). Two of the altered genes in the primary variants, par-3 family cell polarity regulator (*PARD3*) and membrane palmitoylated protein 5 (MPP5), in particular are highly conserved proteins in two complexes that interact in tight junction formation and required for normal cell polarity (Straight *et al.*, 2004; Hurd *et al.*, 2003; Chen *et al.*, 2017). Specifically, knockdown of *PARD3* in glioma cells has been found to promote proliferation and migration by regulating ras homolog family member A (RhoA) through atypical protein kinase C/nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) signalling (Li *et al.*, 2019). These effects provide a potential explanation for the increased effect size of REACTOME\_TIGHT\_JUNCTION\_INTERACTIONS in primary tumours. However, another consequence of altering tight junctions formation is that it impairs the blood brain barrier (BBB). In GBM and other gliomas, the BBB is already impaired through VEGF mediated downregulation of the tight junction protein ceroid-lipofuscinosis neuronal protein 5 (CLN-5) (Argaw *et al.*, 2009). It seems possible that the addition of genetic alterations to other tight junction proteins further increases the permeability of the BBB, thereby allowing increased TMZ delivery to tumour cells and preventing those subclones to survive through therapy. Furthermore, *PARD3* also functions as a subunit of the DNA-dependent protein kinase complex and therefore plays an essential role in repairing double-strand DNA breaks (Fang *et al.*, 2007), providing an additional reason why that variant did not survive through therapy which functions through damaging DNA. Whilst breakdown of the BBB leading to cerebral oedema is the main cause of mortality in GBM patients, targeting a protein such as *PARD3* to both increase TMZ delivery whilst simultaneously impairing DNA repair, before the recurrent tumour grows back, may be worth further investigation.

#### 4.3.7 Conclusions and future directions

In this chapter, I provide evidence that genetic ITH is not the primary driving force behind therapy resistance in GBM. However, by performing pathway analysis, using genesets defined by patterns of variants across primary and recurrent tumours, I identified pathways that are candidates for increasing or decreasing therapy resistance in GBM. Many of these have already been investigated in the lab and in the clinic, thereby supporting that the approach used is identifying truly relevant pathways, and that others not yet investigated, with regards to GBM and therapy resistance, may be worth looking into further.

The findings of the current study could be validated in future work using larger non-longitudinal publicly available GBM datasets, such as those in The Cancer Genome Atlas, by confirming that proportions of non-matched primary or recurrent tumours altered in the candidate pathways are consistent with those in this study. It is also of interest to investigate whether these, or other pathways, stand out in expression data, under the hypothesis that cells will have altered expression of genes in pathways relevant to surviving through therapy.

There are several ways in which the pathway analysis could be furthered to better assess the effect of genetic alterations on GBM progression through therapy. I focussed on using only genes affected by point mutations and not CNAs. Whilst this leaves out a lot of relevant information about how GBM cells are affected, the fact that CNAs can each affect thousands of genes, many of which wouldn't have a significant effect on the cell's function, means that including these would likely introduce a lot of noise into the analysis. This may however be worth exploring in the future. Also, although I filtered out tumours that were hypermutated to reduce noise in the analysis, as PathScore accounts for differing mutation rates per sample, it may instead be better to include these in further analyses.

While the use of a GSEA pathway analysis method allowed for direct comparison between the effect of mutations from different groups of mutations (eg. primary vs. recurrent), they have a disadvantage to network based methods which aren't constrained to rigid pathways and can consider effects that span across multiple pathways or only affect part of a pathway. Therefore, future work may benefit from taking variants from the most relevant PathScore pathways (in terms of increased enrichment from primary to recurrent etc.) and overlaying these onto protein-protein interaction networks (Coker *et al.*, 2019; Kanehisa *et al.*, 2012; Croft *et al.*, 2011). Alternatively, using a method that creates networks of individual pathways from GSEA results, such as Enrichment Map (Merico *et al.*, 2010), would also overcome this issue.

Future work in our group will be to apply the pathway analysis approaches in this chapter to our new in-house samples. Having access to the raw data means that we can perform our own mutation calling and subclonal deconvolution, through an optimised pipeline, which may result in reduced noise in the dataset when looking at changes in variant frequencies. This might provide further support for the relevance of pathways identified in this study, as would rerunning the analysis on future, larger releases of the GLASS dataset, to better separate random noise from pathways having a real effect on GBM progression through therapy. Additionally, it would be interesting to apply this pathway analysis approach to methylation datasets available within our group, with the aim of identifying if changes in CCFs characterised by changes in methylation signal intensities, and not genetic VAFs, are able to highlight pathways relevant to GBM progression through therapy.

## 4.4 Methods

### 4.4.1 SubClonalSelection methods

Variants from the GLASS dataset were included only if they had  $\geq 2$  supporting reads,  $\geq 30x$  coverage, and were in copy neutral regions, identified where  $\text{round}(\text{region copy number} - \text{tumour ploidy}) = 0$ . Those that passed these filters were then randomly downsampled to a maximum of 5000 in order to limit run time.

The SubClonalSelection model (<https://github.com/marcjwilliams1/SubClonalSelection.jl>) (Williams *et al.*, 2018) was run with the recommended  $10^6$  iterations, and 1,000 particles, double the recommended minimum in order to reduce failed runs. ‘ploidy’ values were set to rounded ploidy estimates from TITAN in the GLASS dataset. ‘min\_cellularity’ and ‘max\_cellularity’ values were set to TITAN purity estimates,  $\pm 10\%$ . ‘min\_vaf’ and ‘f\_min’ were set to the maximum point of the leftmost peak in a histogram of VAFs with bin sizes of 0.01, for each sample. ‘read\_depth’ was set to the average coverage across included variants.

Clinical data for the samples used in the analyses are accessible via Synapse (<http://synapse.org/glass>).

#### 4.4.2 Pathway analysis methods

Variants from the GLASS dataset were included with as low as 1 supporting read in a tumour. This is lower than in the SubClonalSelection analysis as the accuracy of VAFs is not as important for the pathway analysis, and added noise from false positive is unlikely to be an issue, especially as most would be filtered out in the next step. To determine those that had a physiological effect on cells, variants were ran through Ensembl VEP (McLaren *et al.*, 2016) and then filtered into high and low stringency sets (Table 15).

Table 15. Annotations filtered for from VEP to create the high and low stringency sets of variants.

Set	Included annotations
High stringency	IMPACT=HIGH SIFT=deleterious SIFT=tolerated_low_confidence SIFT=deleterious_low_confidence PolyPhen=possibly_damaging PolyPhen=probably_damaging PolyPhen=unknown
Low stringency	splice_donor_variant splice_acceptor_variant stop_gained frameshift_variant stop_lost start_lost inframe_insertion inframe_deletion protein_altering_variant missense_variant splice_region_variant incomplete_terminal_codon_variant stop_retained_variant coding_sequence_variant mature_miRNA_variant 5_prime_UTR_variant 3_prime_UTR_variant non_coding_transcript_exon_variant

PathScore was run through the web-app implementation of the model at <http://pathscore.publichealth.yale.edu>, using the 'Gene length, BMR-scaled' method with default background mutation rates. Only variants that could be annotated with a valid Hugo/HGNC Symbol and Entrez ID, as required by PathScore, were included in the analysis. Enrichment scores for pathways were calculated by dividing the 'll\_effective' by 'll\_actual' pathway sizes.

## 4.5 References

- Adzhubei, I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
- Ahmad, F. *et al.* (2019) Cholesterol metabolism: A potential therapeutic target in glioblastoma. *Cancers (Basel)*, **11**.
- Amary, M.F. *et al.* (2011) Ollier disease and Maffucci syndrome are caused by somatic mosaic mutations of IDH1 and IDH2. *Nat. Genet.*, **43**, 1262–1265.
- Argaw, A.T. *et al.* (2009) VEGF-mediated disruption of endothelial CLN-5 promotes blood-brain barrier breakdown. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 1977–1982.
- Bader, G.D. *et al.* (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.*, **34**, D504-6.
- Bailey, M.H. *et al.* (2018) Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, **173**, 371-385.e18.
- Barthel, F.P. *et al.* (2019) Longitudinal molecular trajectories of diffuse glioma in adults. *Nature*, **576**, 112–120.
- Bascur, J.P. *et al.* (2019) IDH1 and IDH2 mutants identified in cancer lose inhibition by isocitrate because of a change in their binding sites. *bioRxiv*, 425025.
- Beck, T.N. *et al.* (2016) Anti-Müllerian Hormone Signaling Regulates Epithelial Plasticity and Chemoresistance in Lung Cancer. *Cell Rep.*, **16**, 657–671.
- Benjamin, D. *et al.* (2019) Calling Somatic SNVs and Indels with Mutect2.
- Bleeker, F.E. *et al.* (2010) The prognostic IDH1R132 mutation is associated with reduced NADP+-dependent IDH activity in glioblastoma. *Acta Neuropathol.*, **119**, 487–494.
- Bozic, I. *et al.* (2019) On measuring selection in cancer from subclonal mutation frequencies. *PLOS Comput. Biol.*, **15**, e1007368.
- Cahill, D.P. *et al.* (2007) Loss of the mismatch repair protein MSH6 in human glioblastomas is associated with tumor progression during temozolomide treatment. *Clin. Cancer Res.*, **13**, 2038–2045.
- Caja, L. *et al.* (2015) Transforming growth factor  $\beta$  and bone morphogenetic protein actions in brain tumors. *FEBS Lett.*, **589**, 1588–1597.
- Chang, H.T. *et al.* (2013) Ketolytic and glycolytic enzymatic expression profiles in malignant gliomas: Implication for ketogenic diet therapy. *Nutr. Metab.*, **10**, 47.
- Cheifetz, S. *et al.* (1987) The transforming growth factor- $\beta$  system, a complex pattern of cross-reactive ligands and receptors. *Cell*, **48**, 409–415.
- Chen, X. *et al.* (2017) Rare Deleterious PARD3 Variants in the aPKC-Binding Region are Implicated in the Pathogenesis of Human Cranial Neural Tube Defects Via Disrupting Apical Tight Junction Formation. *Hum. Mutat.*, **38**, 378–389.
- Chkhaidze, K. *et al.* Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data.
- Christians, A. *et al.* (2019) The prognostic role of IDH mutations in homogeneously treated patients with anaplastic astrocytomas and glioblastomas. *Acta Neuropathol. Commun.*, **7**, 156.
- Ciriello, G. *et al.* (2012) Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, **22**, 398–406.
- Coker, E.A. *et al.* (2019) canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic Acids Res.*, **47**, D917–D922.
- Creixell, P. *et al.* (2015) Pathway and network analysis of cancer genomes. *Nat. Methods*, **12**, 615–621.
- Croft, D. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids*



- Res.*, **39**, D691-7.
- Dang,L. *et al.* (2009) Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature*, **462**, 739–744.
- Enriquez-Navas,P.M. *et al.* (2015) Application of evolutionary principles to cancer therapy. *Cancer Res.*, **75**, 4675–4680.
- Eyler,C.E. *et al.* (2020) Single-cell lineage analysis reveals genetic and epigenetic interplay in glioblastoma drug resistance. *Genome Biol.*, **21**.
- Fang,L. *et al.* (2007) Cell polarity protein Par3 complexes with DNA-PK via Ku70 and regulates DNA double-strand break repair. *Cell Res.*, **17**, 100–116.
- Fischer,K.R. *et al.* (2015) Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. *Nature*, **527**, 472–476.
- Fuchs,M. *et al.* (2001) Disruption of the Sterol Carrier Protein 2 Gene in Mice Impairs Biliary Lipid and Hepatic Cholesterol Metabolism. *J. Biol. Chem.*, **276**, 48058–48065.
- Gaffney,S.G. and Townsend,J.P. (2016) PathScore: a web tool for identifying altered pathways in cancer data. *Bioinformatics*, **32**, btw512.
- García-Campos,M.A. *et al.* (2015) Pathway analysis: State of the art. *Front. Physiol.*, **6**.
- GLASS Consortium (2018) Glioma through the looking GLASS: Molecular evolution of diffuse gliomas and the Glioma Longitudinal Analysis Consortium. *Neuro. Oncol.*, **20**, 873–884.
- Gu,Y. *et al.* (2013) Network analysis of genomic alteration profiles reveals co-altered functional modules and driver genes for glioblastoma. *Mol. Biosyst.*, **9**, 467–477.
- Gujar,A.D. *et al.* (2016) An NAD<sup>+</sup>-dependent transcriptional program governs self-renewal and radiation resistance in glioblastoma. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, E8247–E8256.
- Guo,D. *et al.* (2011) An LXR agonist promotes glioblastoma cell death through inhibition of an EGFR/AKT/SREBP-1/LDLR-dependent pathway. *Cancer Discov.*, **1**, 442–456.
- Ha,G. *et al.* (2014) TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.*, **24**, 1881–93.
- Han,M. *et al.* (2020) Therapeutic implications of altered cholesterol homeostasis mediated by loss of CYP46A1 in human glioblastoma. *EMBO Mol. Med.*, **12**.
- Hao,Y. *et al.* (2019) TGF- $\beta$ -mediated epithelial-mesenchymal transition and cancer metastasis. *Int. J. Mol. Sci.*, **20**.
- Hartmann,C. *et al.* (2009) Type and frequency of IDH1 and IDH2 mutations are related to astrocytic and oligodendroglial differentiation and age: A study of 1,010 diffuse gliomas. *Acta Neuropathol.*, **118**, 469–474.
- Heverin,M. *et al.* (2005) Crossing the barrier: Net flux of 27-hydroxycholesterol into the human brain. *J. Lipid Res.*, **46**, 1047–1052.
- Houillier,C. *et al.* (2010) IDH1 or IDH2 mutations predict longer survival and response to temozolomide in low-grade gliomas. *Neurology*, **75**, 1560–1566.
- Hurd,T.W. *et al.* (2003) Direct interaction of two polarity complexes implicated in epithelial tight junction assembly. *Nat. Cell Biol.*, **5**, 137–142.
- Huyghe,S. *et al.* (2006) Peroxisomal multifunctional protein-2: The enzyme, the patients and the knockout mouse model. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids*, **1761**, 973–994.
- ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**, 82–93.
- Ikushima,H. and Miyazono,K. (2010) TGF $\beta$  2 signalling: A complex web in cancer progression. *Nat. Rev. Cancer*, **10**, 415–424.
- Inda,M.D.M. *et al.* (2010) Tumor heterogeneity is an active process maintained by a mutant EGFR-induced cytokine circuit in glioblastoma. *Genes Dev.*, **24**, 1731–1745.
- Kambach,D.M. *et al.* (2017) Disabled cell density sensing leads to dysregulated cholesterol synthesis in glioblastoma. *Oncotarget*, **8**, 14860–14875.
- Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109-14.
- Khakabimamaghani,S. *et al.* (2019) Uncovering the subtype-specific temporal order of cancer pathway dysregulation. *PLOS Comput. Biol.*, **15**, e1007451.
- Kiriyama,Y. and Nochi,H. (2019) The biosynthesis, signaling, and neurological functions of bile acids. *Biomolecules*, **9**.

- Körber, V. *et al.* (2019) Evolutionary Trajectories of IDH WT Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis. *Cancer Cell*, **35**, 692–704.e12.
- Krabbath, Z. and Kalman, B. (2020) Longitudinal Characteristics of Glioblastoma in Genome-Wide Studies. *Pathol. Oncol. Res.*, **26**, 2035–2047.
- Krishnan, P.D.G. *et al.* (2020) Rab gtpases: Emerging oncogenes and tumor suppressive regulators for the editing of survival pathways in cancer. *Cancers (Basel)*, **12**.
- Leiserson, M.D.M. *et al.* (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, **47**, 106–114.
- Leiserson, M.D.M. *et al.* (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Comput. Biol.*, **9**, e1003054.
- Lemond, H.A. *et al.* (2003) Mutations in SRD5B1 (AKR1D1), the gene encoding  $\delta$  4-3-oxosteroid 5 $\beta$ -reductase, in hepatitis and liver failure in infancy. *Gut*, **52**, 1494–1499.
- Li, J. *et al.* (2019) Pard3 suppresses glioma invasion by regulating RhoA through atypical protein kinase C/NF- $\kappa$ B signaling. *Cancer Med.*, **8**, 2288–2302.
- Liau, B.B. *et al.* (2017) Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance. *Cell Stem Cell*, **20**, 233–246.e7.
- Liberzon, A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
- Liu, C.-C. *et al.* (2016) Transforming Growth Factor-beta Promotes Glioblastoma Multiforme to Establish Radiation Resistance by Mesenchymal Differentiation. *FASEB J.*, **30**, 1227.1–1227.1.
- Lu, C. *et al.* (2012) IDH mutation impairs histone demethylation and results in a block to cell differentiation. *Nature*, **483**, 474–478.
- Martincorena, I. *et al.* (2017) Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, **171**, 1029–1041.e21.
- Martínez-Jiménez, F. *et al.* (2020) A compendium of mutational cancer driver genes. *Nat. Rev. Cancer*, 1–18.
- Massagué, J. (1998) TGF- $\beta$  SIGNAL TRANSDUCTION. *Annu. Rev. Biochem.*, **67**, 753–791.
- Massagué, J. *et al.* (2000) TGF $\beta$  signaling in growth control, cancer, and heritable disorders. *Cell*, **103**, 295–309.
- McLaren, W. *et al.* (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.
- McLendon, R. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
- McMillin, M. *et al.* (2018) FXR-Mediated Cortical Cholesterol Accumulation Contributes to the Pathogenesis of Type A Hepatic Encephalopathy. *CMGH*, **6**, 47–63.
- McMillin, M. and DeMorrow, S. (2016) Effects of bile acids on neurological function and disease. *FASEB J.*, **30**, 3658–3668.
- Melamed, R.D. *et al.* (2015) An information theoretic method to identify combinations of genomic alterations that promote glioblastoma. *J. Mol. Cell Biol.*, **7**, 203.
- Merico, D. *et al.* (2010) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One*, **5**, e13984.
- Molenaar, R.J. *et al.* (2018) Wild-type and mutated IDH1/2 enzymes and therapy responses. *Oncogene*, **37**, 1949–1960.
- Nakagawa, H. *et al.* (2018) Sodium butyrate induces senescence and inhibits the invasiveness of glioblastoma cells. *Oncol. Lett.*, **15**, 1495.
- Nakahara, M. *et al.* (2002) Bile acids enhance low density lipoprotein receptor gene expression via a MAPK cascade-mediated stabilization of mRNA. *J. Biol. Chem.*, **277**, 37229–37234.
- Nayak, S. *et al.* (2020) Bone Morphogenetic Protein 4 Targeting Glioma Stem-Like Cells for Malignant Glioma Treatment: Latest Advances and Implications for Clinical Application. *Cancers (Basel)*, **12**, 516.
- Neftel, C. *et al.* (2019) An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell*, **178**, 835–849.e21.
- Okita, Y. *et al.* (2012) IDH1/2 mutation is a prognostic marker for survival and predicts response to chemotherapy for grade II gliomas concomitantly treated with radiation therapy. *Int. J. Oncol.*, **41**, 1325–1336.
- Parsons, D.W. *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science (80-. )*, **321**, 1807–1812.
- Patel, D. *et al.* (2019) LXR $\beta$  controls glioblastoma cell growth, lipid balance, and immune modulation

- independently of ABCA1. *Sci. Rep.*, **9**, 1–15.
- Phillips, H.S. *et al.* (2006) Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell*, **9**, 157–173.
- Pisanti, S. *et al.* (2014) Novel prospects of statins as therapeutic agents in cancer. *Pharmacol. Res.*, **88**, 84–98.
- Prabhu, A.H. *et al.* (2019) Abstract 5280: Branched chain amino acids represent indispensable metabolic substrates in glioblastoma. In, *Cancer Research*. American Association for Cancer Research (AACR), pp. 5280–5280.
- Raja, E. *et al.* (2017) Bone morphogenetic protein signaling mediated by ALK-2 and DLX2 regulates apoptosis in glioma-initiating cells. *Oncogene*, **36**, 4963–4974.
- Roth, A. *et al.* (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396–8.
- Russell, D.W. *et al.* (2009) Cholesterol 24-Hydroxylase: An Enzyme of Cholesterol Turnover in the Brain. *Annu. Rev. Biochem.*, **78**, 1017–1040.
- Sachdeva, R. *et al.* (2019) BMP signaling mediates glioma stem cell quiescence and confers treatment resistance in glioblastoma. *Sci. Rep.*, **9**, 1–14.
- Sánchez, N.S. *et al.* (2011) The cytoplasmic domain of TGF $\beta$ R3 through its interaction with the scaffolding protein, GIPC, directs epicardial cell behavior. *Dev. Biol.*, **358**, 331–343.
- Shea, H.C. *et al.* (2007) Analysis of HSD3B7 knockout mice reveals that a 3 $\alpha$ -hydroxyl stereochemistry is required for bile acid function. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 11526–11533.
- Shibue, T. and Weinberg, R.A. (2017) EMT, CSCs, and drug resistance: The mechanistic link and clinical implications. *Nat. Rev. Clin. Oncol.*, **14**, 611–629.
- Siegmund, K. and Shibata, D. (2016) At least two well-spaced samples are needed to genotype a solid tumor. *BMC Cancer*, **16**, 250.
- SongTao, Q. *et al.* (2012) IDH mutations predict longer survival and response to temozolomide in secondary glioblastoma. *Cancer Sci.*, **103**, 269–273.
- Straight, S.W. *et al.* (2004) Loss of PALS1 Expression Leads to Tight Junction and Polarity Defects. *Mol. Biol. Cell*, **15**, 1981–1990.
- Suh, E.H. *et al.* (2019) In vivo assessment of increased oxidation of branched-chain amino acids in glioblastoma. *Sci. Rep.*, **9**.
- Sun, R. *et al.* (2017) Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.*, **49**, 1015–1024.
- Suvà, M.L. and Tirosh, I. (2020) The Glioma Stem Cell Model in the Era of Single-Cell Genomics. *Cancer Cell*, **37**, 630–636.
- Taal, W. *et al.* (2011) First-line temozolomide chemotherapy in progressive low-grade astrocytomas after radiotherapy: molecular characteristics in relation to response. *Neuro. Oncol.*, **13**, 235–41.
- Tan, E.J. *et al.* (2015) Reprogramming during epithelial to mesenchymal transition under the control of TGF $\beta$ . *Cell Adhes. Migr.*, **9**, 233–246.
- Tateishi, K. *et al.* (2015) Extreme Vulnerability of IDH1 Mutant Cancers to NAD<sup>+</sup> Depletion. *Cancer Cell*, **28**, 773–784.
- Tateishi, K. *et al.* (2017) The alkylating chemotherapeutic temozolomide induces metabolic stress in IDH1-mutant cancers and potentiates NAD<sup>+</sup> depletion-mediated cytotoxicity. *Cancer Res.*, **77**, 4102–4115.
- Tomicic, M.T. and Christmann, M. (2018) Targeting anticancer drug-induced senescence in glioblastoma therapy. *Oncotarget*, **9**, 37466–37467.
- Tönjes, M. *et al.* (2013) BCAT1 promotes cell proliferation through amino acid catabolism in gliomas carrying wild-type IDH1. *Nat. Med.*, **19**, 901–908.
- Vallejo, F.A. *et al.* (2019) Exploiting metabolic susceptibilities in glioblastoma via glycolytic inhibition and ketogenic therapy. *J. Clin. Oncol.*, **37**, e13558–e13558.
- Vandin, F. *et al.* (2011) Algorithms for detecting significantly mutated pathways in cancer. In, *Journal of Computational Biology.*, pp. 507–522.
- Vandin, F. (2017) Computational Methods for Characterizing Cancer Mutational Heterogeneity. *Front. Genet.*, **8**, 83.
- Vandin, F. *et al.* (2012) De novo discovery of mutated driver pathways in cancer. *Genome Res.*, **22**, 375–385.
- Vaser, R. *et al.* (2015) SIFT missense predictions for genomes.

- Venkataramani, V. *et al.* (2019) Glutamatergic synaptic input to glioma cells drives brain tumour progression. *Nature*, **573**, 532–538.
- Venkatesh, H.S. *et al.* (2019) Electrical and synaptic integration of glioma into neural circuits. *Nature*, **573**, 539–545.
- Villa, G.R. *et al.* (2016) An LXR-Cholesterol Axis Creates a Metabolic Co-Dependency for Brain Cancers. *Cancer Cell*, **30**, 683–693.
- Wahl, D.R. *et al.* (2017) Glioblastoma therapy can be augmented by targeting IDH1-mediated NADPH biosynthesis. *Cancer Res.*, **77**, 960–970.
- Wang, J. *et al.* (2016) Clonal evolution of glioblastoma under therapy. *Nat. Genet.*, **48**, 768–776.
- Ward, P.S. *et al.* (2012) Identification of additional IDH mutations associated with oncometabolite R(-)-2-hydroxyglutarate production. *Oncogene*, **31**, 2491–2498.
- Ward, P.S. *et al.* (2010) The Common Feature of Leukemia-Associated IDH1 and IDH2 Mutations Is a Neomorphic Enzyme Activity Converting  $\alpha$ -Ketoglutarate to 2-Hydroxyglutarate. *Cancer Cell*, **17**, 225–234.
- Watanabe, T. *et al.* (2009) IDH1 mutations are early events in the development of astrocytomas and oligodendrogliomas. *Am. J. Pathol.*, **174**, 1149–1153.
- Wick, A. *et al.* (2020) Phase 1b/2a study of galunisertib, a small molecule inhibitor of transforming growth factor-beta receptor I, in combination with standard temozolomide-based radiochemotherapy in patients with newly diagnosed malignant glioma. *Invest. New Drugs*, 1–10.
- Williams, M.J. *et al.* (2016) Identification of neutral tumor evolution across cancer types. *Nat. Genet.*, **48**, 238–244.
- Williams, M.J. *et al.* (2018) Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.*, **50**, 895–903.
- Wise, D.R. *et al.* (2011) Hypoxia promotes isocitrate dehydrogenase-dependent carboxylation of  $\alpha$ -ketoglutarate to citrate to support cell growth and viability. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 19611–19616.
- Wodarz, A. and Näthke, I. (2007) Cell polarity in development and cancer. *Nat. Cell Biol.*, **9**, 1016–1024.
- Wouters, B.J. (2017) Hitting the target in IDH2 mutant AML. *Blood*, **130**, 693–694.
- Xi, G. *et al.* (2017) Therapeutic Potential for Bone Morphogenetic Protein 4 in Human Malignant Glioma. *Neoplasia (United States)*, **19**, 261–270.
- Xu, W. *et al.* (2011) Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of  $\alpha$ -ketoglutarate-dependent dioxygenases. *Cancer Cell*, **19**, 17–30.
- Yamamoto, Y., Tomiyama, A., *et al.* (2018) Intracellular cholesterol level regulates sensitivity of glioblastoma cells against temozolomide-induced cell death by modulation of caspase-8 activation via death receptor 5-accumulation and activation in the plasma membrane lipid raft. *Biochem. Biophys. Res. Commun.*, **495**, 1292–1299.
- Yamamoto, Y., Sasaki, N., *et al.* (2018) Involvement of intracellular cholesterol in temozolomide-induced glioblastoma cell death. *Neurol. Med. Chir. (Tokyo)*, **58**, 296–302.
- Yan, H. *et al.* (2009) IDH1 and IDH2 mutations in gliomas. *N. Engl. J. Med.*, **360**, 765–773.
- Ye, D. *et al.* (2018) Metabolism, Activity, and Targeting of D- and L-2-Hydroxyglutarates. *Trends in Cancer*, **4**, 151–165.
- Zhang, J. and Zhang, S. (2018) The discovery of mutated driver pathways in cancer: Models and algorithms. *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, **15**, 988–998.
- Zhang, M. *et al.* (2011) Blockade of TGF- $\beta$  signaling by the TGF $\beta$ R-I kinase inhibitor LY2109761 enhances radiation response and prolongs survival in glioblastoma. *Cancer Res.*, **71**, 7155–7167.
- Zheng, X. *et al.* (2015) Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. *Nature*, **527**, 525–530.

# Chapter 5 - Comparison of intratumour heterogeneity between GBM patient biopsies and patient-derived orthotopic xenografts

The majority of the work presented in this chapter was originally published by *Acta Neuropathologica* in (Golebiewska *et al.*, 2020). The material is reproduced here under the Creative Commons Attribution 4.0 International License.

## 5.1 Introduction

*In vivo* cancer models are a frequently used resource for investigating cancer biology and assessing therapeutic effects. It is important that they adequately recapitulate relevant aspects of patient tumours so that they can reliably inform us on how real tumours react to interventions. The previous chapters have focussed on understanding how genetic intratumour heterogeneity and the clonal architectures of glioblastoma (GBM) tumours evolves over time, but the same techniques can also be used to address the question of whether these clonal architectures are conserved between patient biopsies and the models derived from them. Such an approach is carried out in this chapter to investigate the conservation of clonal architecture in a novel model system.

Patient-derived xenografts (PDX) are commonly used models, where tumour cells are implanted into an immunodeficient animal, typically mice. Traditionally, this is achieved by injecting patient-derived tumour cell lines subcutaneously into a mouse. After multiple rounds of passages in culture, these cells undergo genetic drift and no longer recapitulate the full biology of the original biopsy. Alternatively, tumour fragments taken directly from a patient biopsy can be implanted subcutaneously, which better maintain the phenotype of the original parent tumour and correlate more closely in therapeutic responses (Yoshida, 2020). Despite this, studies have shown that mouse PDXs undergo mouse-specific tumour evolution, where rapid gains and losses of copy number alterations (CNAs) from selection of pre-existing minor clones occurs during PDX passaging in a manner that differs to those acquired during progression in patients (Ben-David *et al.*, 2017). A possible reason for this is due to differences in the tumour microenvironment. For brain tumours specifically, this differing selection may result from PDXs lacking the effects imposed by the blood-brain-barrier and cerebrospinal fluid.

To mitigate the above issues with PDXs, studies have alternatively used patient-derived orthotopic xenograft (PDOX) models, where the patient-derived tumour cells are inserted into the same corresponding location the tumour was derived from. Direct transplantation of bulk tissue into mouse brains is challenging, and therefore GBM PDOXs typically rely on injection of enzyme dissociated cells (Lai *et al.*, 2017), often cultured for unspecified times and passages. This results in the loss of the tumour tissue

architecture, leading to chromosomal instability and potential loss of tumour heterogeneity (Bolhaqueiro *et al.*, 2019; Knouse *et al.*, 2018). Researches have addressed some of these issues by creating GBM PDOXs from dissociated cells that were cultured through only a low passage number. This resulted in PDOX tumours that accurately recapitulated genetic features, histopathological properties, and treatment response of the original patient tumour (Joo *et al.*, 2013). However, this method still relied on passaging *in vitro*, and loses the tissue architecture of the original biopsy.

To overcome these limitations, our collaborators at the NorLux Neuro-oncology laboratory based at the Luxembourg Institute of health, have developed glioma PDOX models using 3D organoids derived from mechanically minced tumour tissue, without enzyme dissociation, that was only briefly maintained in culture without any *in vitro* passaging (Golebiewska *et al.*, 2020). To investigate whether these models more closely recapitulated human GBM tumour biology, they assessed several aspects and compared them to both patient biopsies and existing models. Common driver variants were maintained from biopsy to PDOXs, and copy numbers (CNs) were mostly very similar, with only minor and glioma specific differences that may result from sampling bias. Unsupervised hierarchical clustering of transcriptomic profiles showed that normal human brain, biopsies, PDOXs, and cell lines all clustered in groups depending on the sample type, however PDOXs clustered more closely to biopsies and normal brain, than to cell lines. Inter-patient differences and heterogeneous expression of stem cell markers were retained in PDOXs, and similar cellular subpopulations were found in mouse-derived tumour microenvironments as in patients. Some changes in methylation profiles were seen between biopsies and PDOXs, though on the whole they correlated well, clustering primarily by isocitrate dehydrogenase (*IDH*) status. Furthermore, O6-methylguanine-DNA methyl-transferase (*MGMT*) promoter methylation status was conserved in all but 2 out of 28 PDOXs.

It was not known, however, to what extent the genetic subclonal architecture was maintained between biopsies and PDOXs. I therefore collaborated with the group to estimate cancer cell fractions (CCFs) of variants in the tumours, and assess whether these changed between biopsies and resulting PDOX models, from three patients with GBMs. This involved first point variant calling and determination of variant allele frequencies (VAFs), and estimation of absolute CNs at each variant position. From these, CCFs for each variant can then be estimated and clustered to determine changes in subclone frequencies.

## 5.2 Results

### 5.2.1 Sample details

Three patients were included in the analysis, with all samples from them indicating *IDHwt* GBMs (Table 16). LIH0192 and LIH0347 had longitudinal samples for one and two recurrences, respectively, having received radio and chemotherapy since the primary surgery.

Table 16. Samples from the three patients included in the analysis.

Patient	Sample type	Primary	1st Recurrence	2 <sup>nd</sup> Recurrence
LIH0192	Biopsy	T192_Biopsy	T233_Biopsy	T251_Biopsy
	PDOX	T192_PDX	T233_PDX	T251_PDX
LIH0347	Biopsy	T347_Biopsy	T470_Biopsy	-
	PDOX	T347_PDX	T470_PDX	-
LIH0158	Biopsy	T158_Biopsy	-	-
	PDOX	T158_PDX	-	-

### 5.2.2 Variant calling

Sequencing data for targeted panels of either 150 or 234 genes were available for biopsy and PDOX samples. No matched normal samples were available so standard somatic variant calling, where variants are checked against a normal to determine either germline or somatic status, was not an option. Instead, I performed variant calling using VarScan2's germline method and then filtered out variants at common polymorphic sites (Sherry *et al.*, 1999) which is likely to remove over 90% of germline variants (Shen *et al.*, 2013; Auton *et al.*, 2015). I then further filtered the variants for those in regions targeted by the 150 gene panel, which is almost entirely a subset of the larger 234 gene panel regions, in order to make matched primary and recurrent samples comparable.

### 5.2.3 Copy number calling

Subclonal deconvolution requires estimation of CNs at each variant position. Whilst allele-specific CNs are preferable for higher accuracy, these are not possible to calculate without germline variant B-allele frequencies, for which the sequencing data available was not sufficient due to small target size and lack of a normal reference. I therefore instead estimated absolute total CNs from array Comparative Genomic Hybridisation (aCGH), using the DNACopy R package (Seshan E. and Olshen, 2019). This found chr7 amplification and chr10 loss in all samples, both of which are present in the majority of IDHwt GBMs (Barthel *et al.*, 2019; McNulty *et al.*, 2019; Körber *et al.*, 2019; Gerstung *et al.*, 2020). Coincidentally, all analysed samples across the three patients also all showed a gain of chr19 and chr20. While these are not as common as alterations to chr7 and chr10, they are associated with GBMs containing focal epidermal growth factor receptor (EGFR) amplifications (McNulty *et al.*, 2019), which is present in all of the tumours in this study, other than T158\_PDOX. I developed a custom script to calculate purities and absolute CNs from segmented normalised log<sub>2</sub> ratios, based on the assumption that chr7 likely had a clonal single copy gain in all the samples included. This was due to the observation that all samples had a similar increased log<sub>2</sub> ratio

for chr7, chr19 and chr20, meaning that all three chromosomes have the same total number of copies in each sample. If these weren't split evenly between cells as clonal events, then some fluctuation in the relative log2 ratios between these chromosomes in different samples would be expected, due to variations in subclonal frequencies. The exception to this is if the gains were all in the same subclone, though this is less likely, particularly as chr7 gain is thought to be a tumour initiating event (Körber *et al.*, 2019; Gerstung *et al.*, 2020). Assuming there are an equal number of clonal gains in chr7, chr19 and chr20, the most likely scenario is that they all have a gain of one copy, as has been seen in 90% of GBMs for at least one of these chromosomes (Gerstung *et al.*, 2020). Furthermore, the gains are equal in magnitude to the loss of chr10 in most of the samples, which is limited to the loss of just 1 copy, and also thought to be a clonal tumour initiating event in GBMs (Körber *et al.*, 2019; Gerstung *et al.*, 2020), providing additional support for the gains being of just 1 copy. Therefore, taking into account the known log2 value for a ratio of 3 copies to 2 reference copies, the formula I used to calculate purity is:

$$purity = \frac{\overline{chr7}}{\log_2\left(\frac{3}{2}\right)}$$

where  $\overline{chr7}$  is the mean log2 ratio value across chr7 (excluding values  $\geq 1$  or  $\leq 0.2$ , in order to avoid the effects of focal amplifications on the calculation). These estimates are likely to be accurate owing to the fact that they take into account known prior information about GBMs, as well as observations spanning multiple samples. CNs (cn) were calculated for each segmented region using the formula:

$$cn = 2\left(2^{\left(\frac{lr}{purity}\right)}\right)$$

where *lr* is the log2 ratio for a segment. I then rounded CNs to absolute values (abs\_cn) using the code:

$$\text{if } cn \geq (\overline{chr10} + 0.2), \text{ then } abs\_cn = \text{round}(cn)$$

$$\text{if } cn < (\overline{chr10} + 0.2), \text{ then } abs\_cn = \max(\text{round}(cn - (\overline{chr10} - 1)), 0)$$

where  $\overline{chr10}$  is the mean log2 value across chr10 (excluding values  $\geq -0.02$  or  $\leq -1$ ). This complex form of rounding was required to prevent chr10 and other regions showing clear losses, being rounded up to 2 in some samples.

The normalised log2 ratios I was provided with had many probes missing (ranging from ~5-80% of total probes), which had been filtered out during the normalisation step due to poor quality. This meant CNs could not be estimated for large portions of the genome, and other regions, with low number of probes, having biased and very inaccurate segmented log2 values (Figure 37A). However, I also had access to the non-normalised log2 ratios for the full set of probes. I therefore overcame the issue by first removing probes for any chromosome that had less than 600 included, and then assuming these chromosomes and



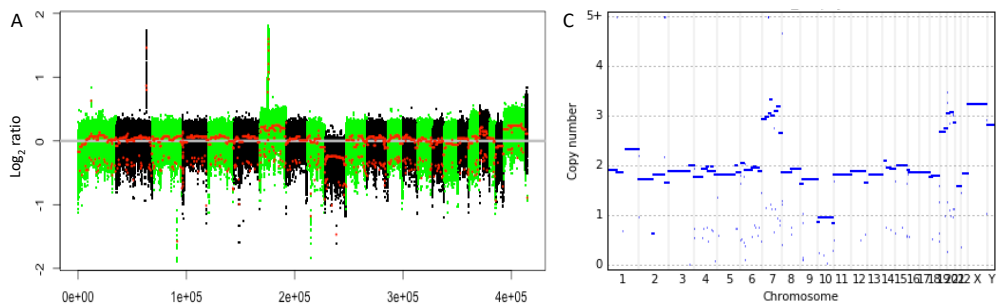
all other regions not covered by a segmented log<sub>2</sub> ratio, had a log<sub>2</sub> ratio of 0. By inspecting the non-normalised log<sub>2</sub> ratios, by eye, I concluded that this approach was largely accurate and did not remove any obvious CNAs seen in the non-normalised dataset (Figure 37). Some inaccurate CN estimates still remained, where the CN was estimated at 0 despite many variants found in those regions, and with read depths that were representative of the whole sample and therefore unlikely to be present only subclonally. This mostly affected just T192-PDX, T233-PDX, and T251-Biopsy.

Overall, CN estimates were very similar between biopsy and PDOX samples, with generally only small focal regions showing differences, many of which are likely due to noise.

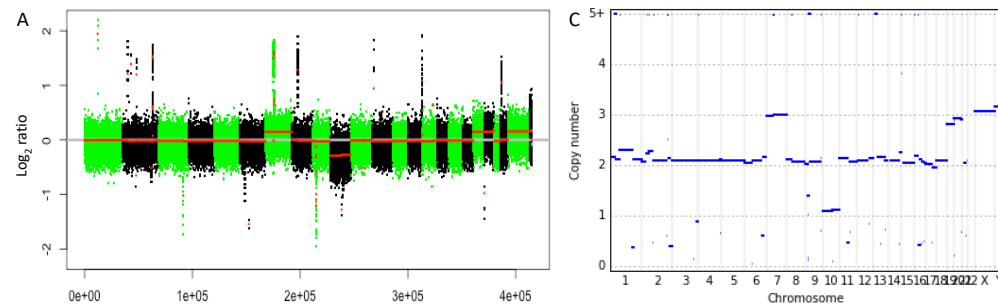
Figure 37. (Below) Processing of array-CGH data. A) Segmented non-normalised log<sub>2</sub> ratios. B) Segmented and normalised log<sub>2</sub> ratios. C) Unrounded CN estimates, calculated from log<sub>2</sub> ratios using purity estimates. D) Rounded absolute CN estimates, calculated using a custom algorithm.

# Patient LIH0192

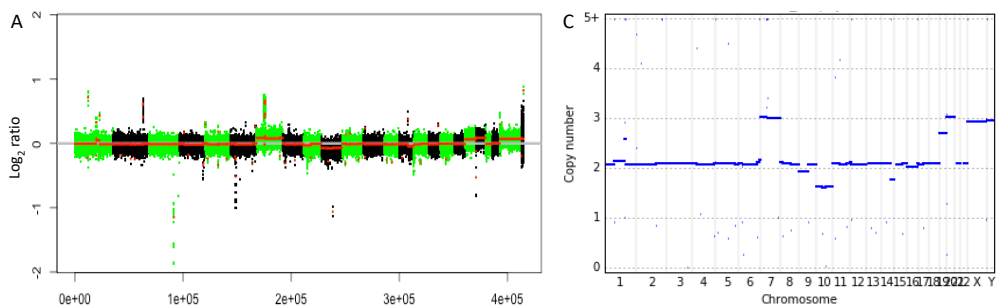
T192\_Biopsy – purity=0.853



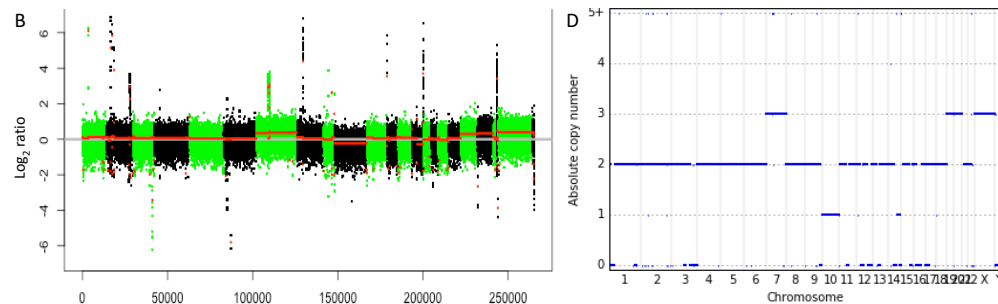
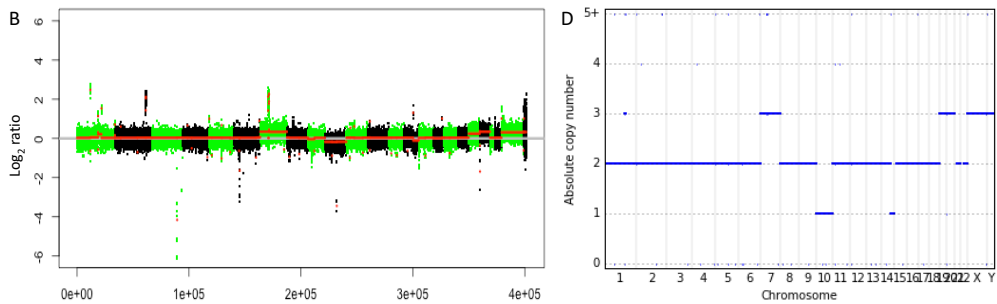
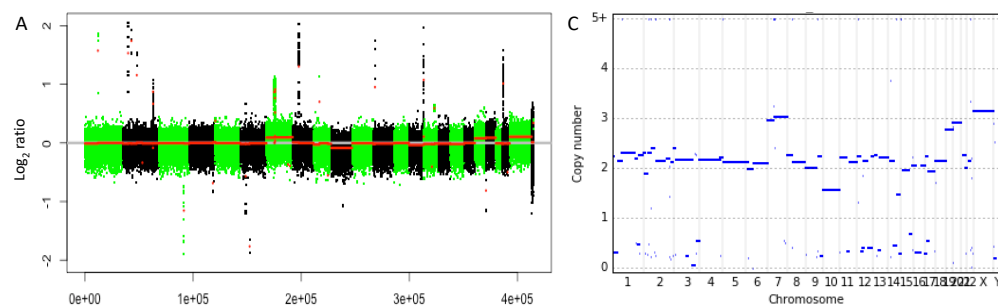
T192\_PDX – purity=1.001



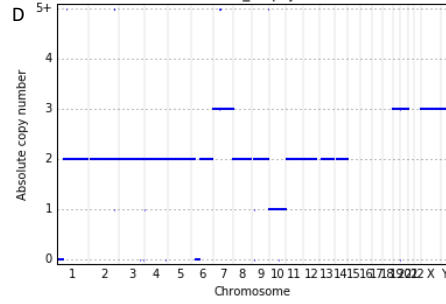
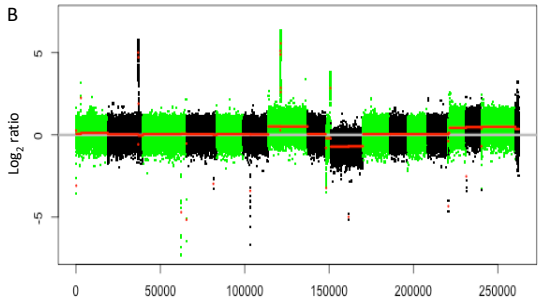
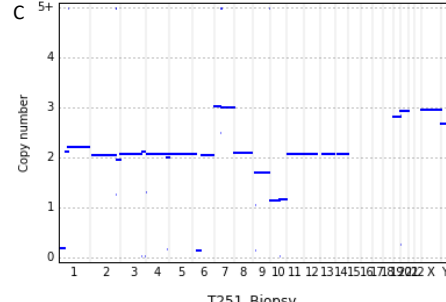
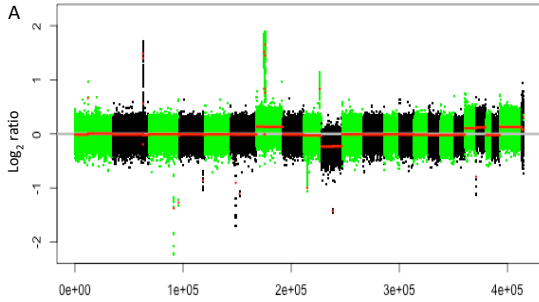
T233\_Biopsy – purity=0.577



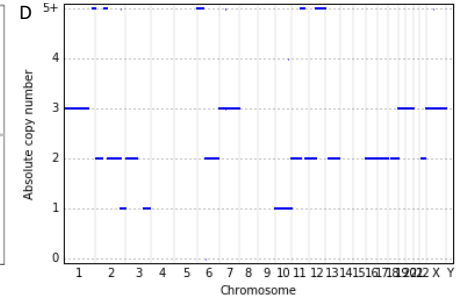
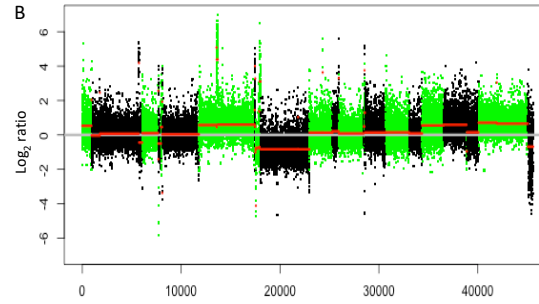
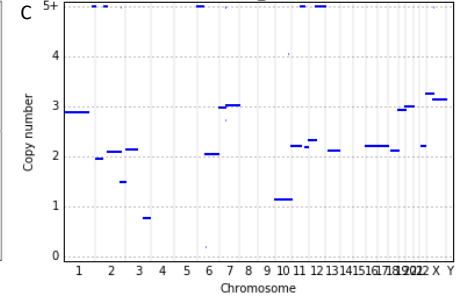
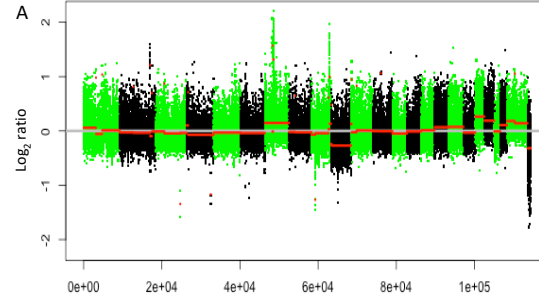
T233\_PDX – purity=0.628



T251\_Biopsy – purity=0.877

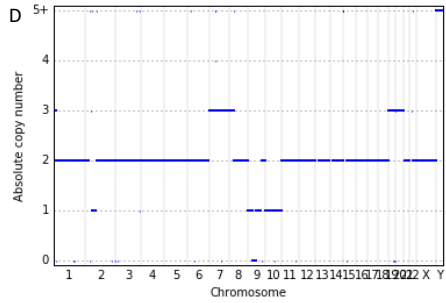
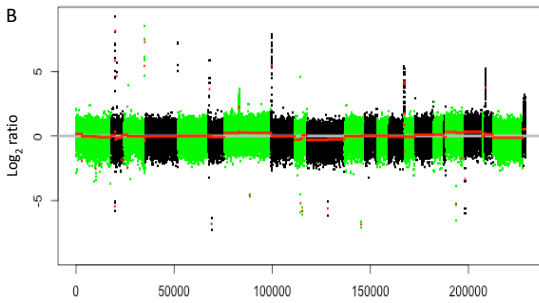
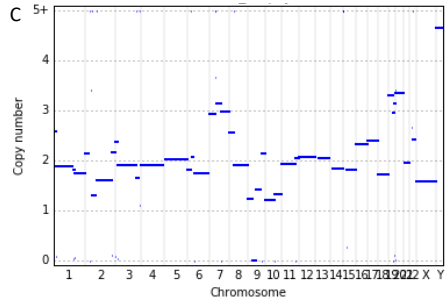
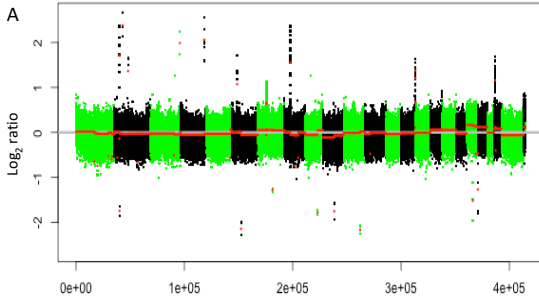


T251\_PDX – purity=1.007

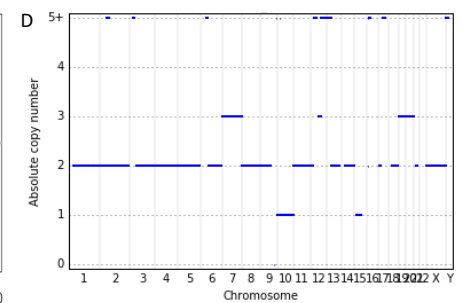
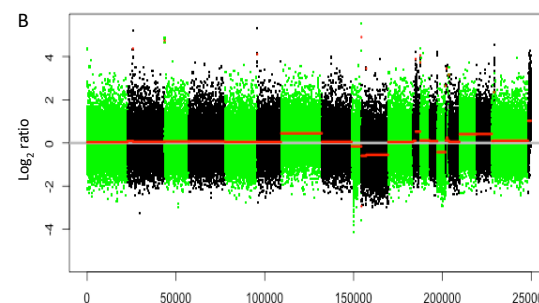
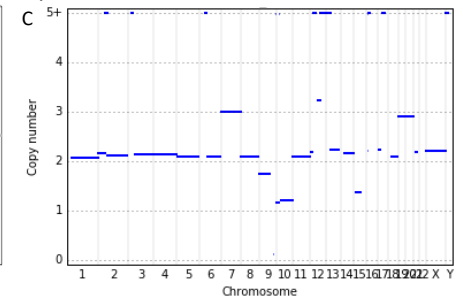
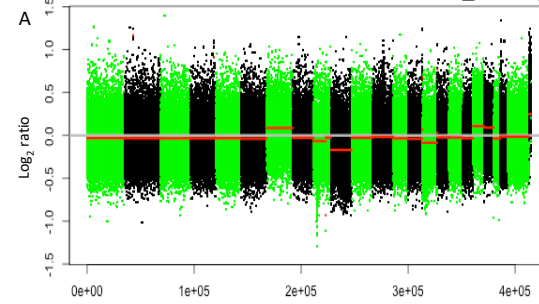


### Patient LIH0158

T158\_Biopsy – purity=0.425

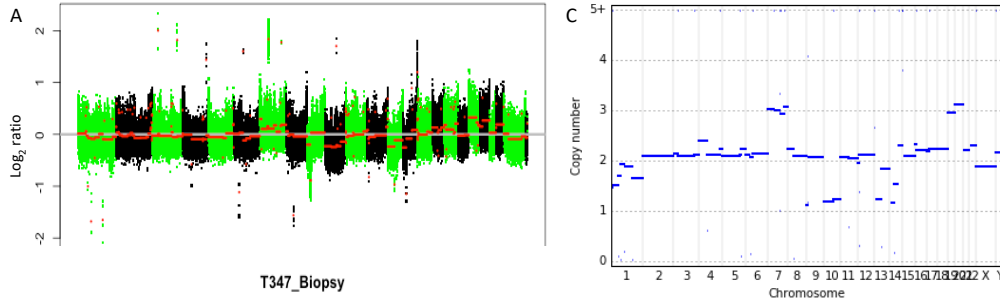


T158\_PDX – purity=0.763

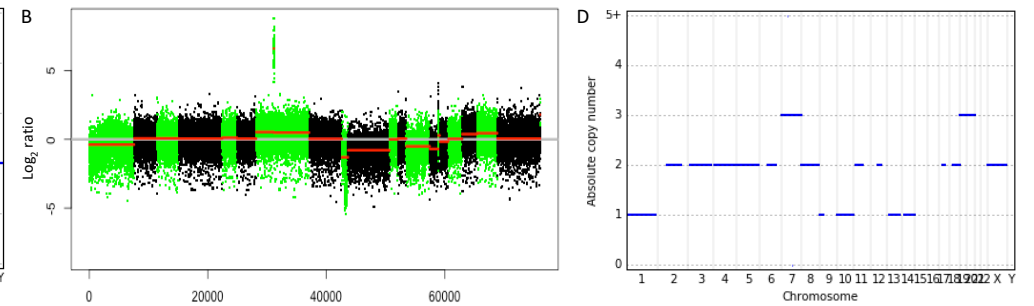
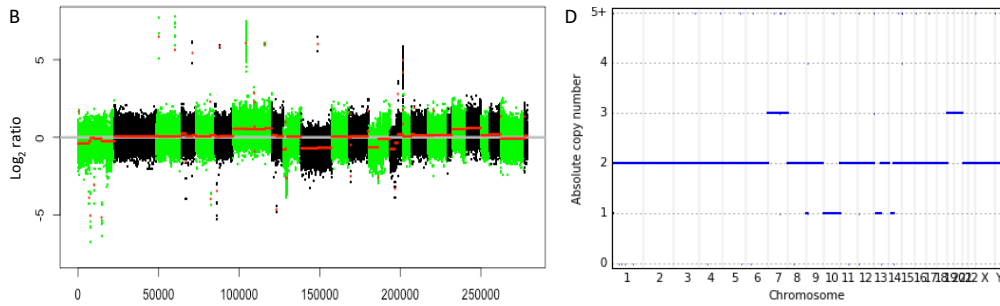
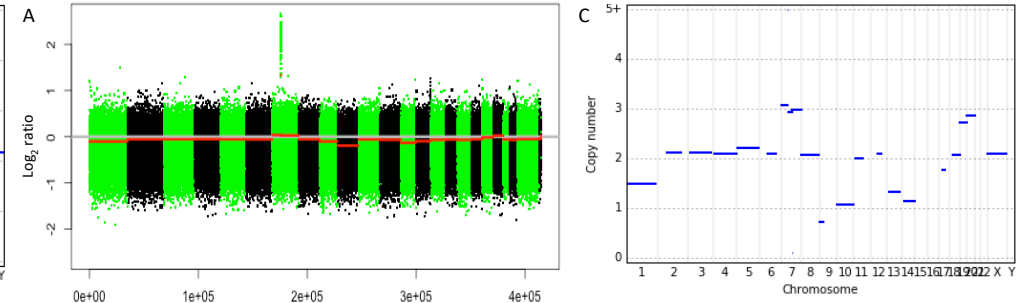


# Patient LIH0347

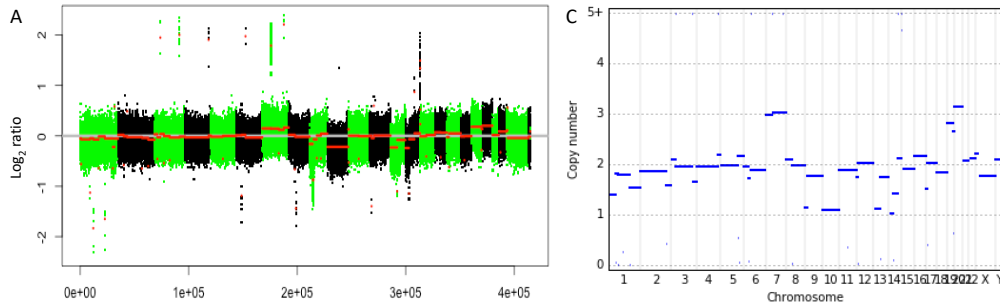
## T347\_Biopsy – purity=0.939



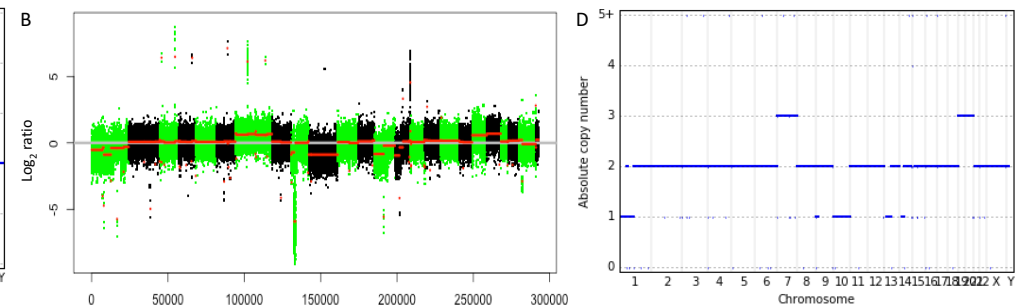
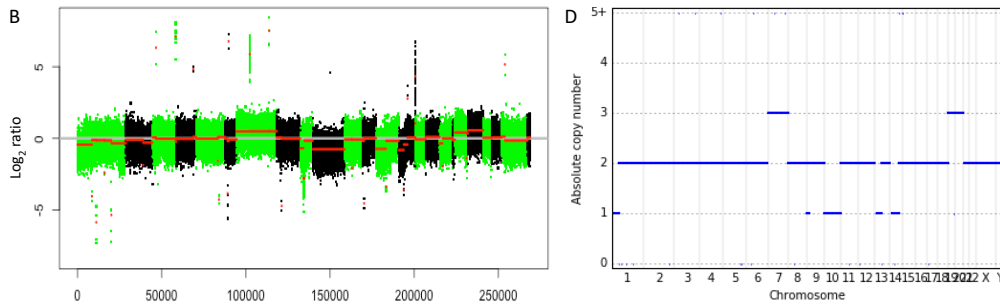
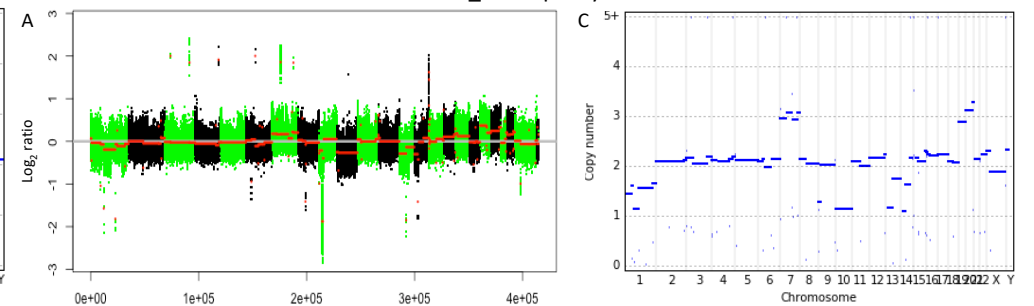
## T347\_PDX – purity=0.865



## T470\_Biopsy – purity=0.852



## T470\_PDX – purity=1.079



I also had access to methylation array data, from which it's also possible to obtain CN and purity information. I did this using the *cnAnalysis450k* (Knoll *et al.*, 2017) and *minfi* (Aryee *et al.*, 2014) R packages, however the data resulted in very noisy log2 ratios (Figure 38) and so I continued with the aCGH method.

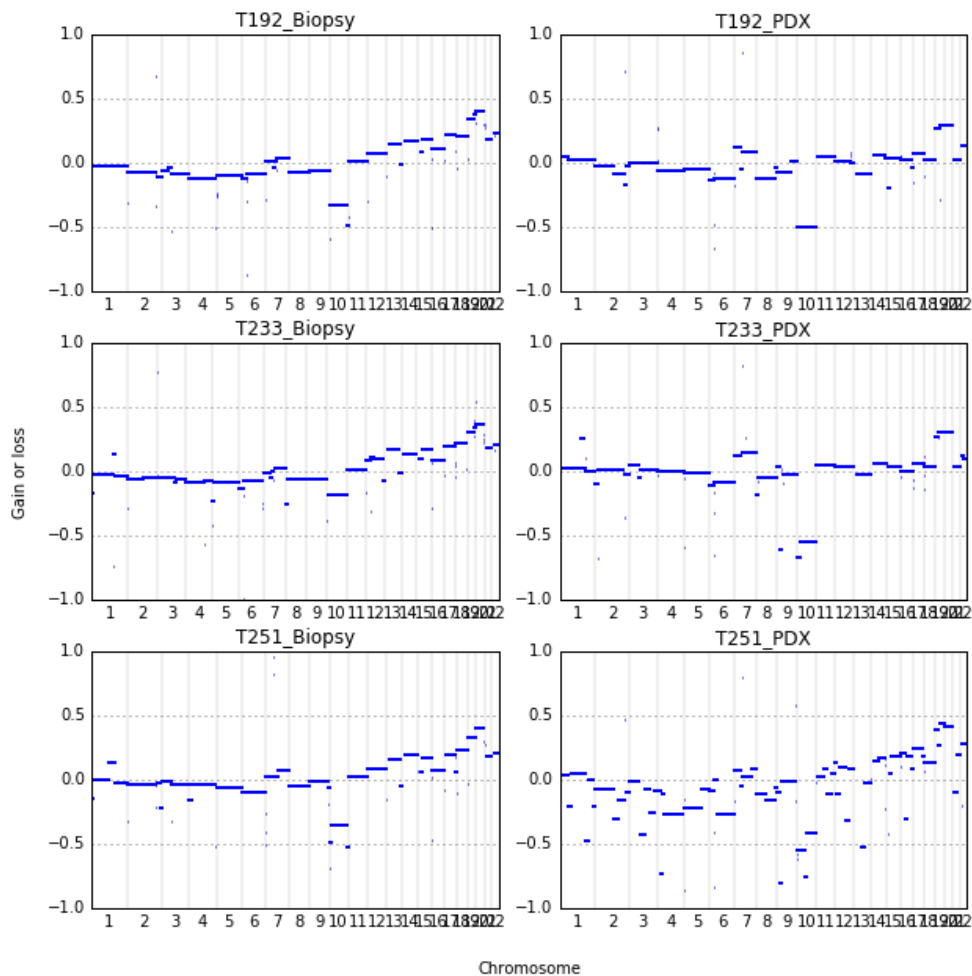


Figure 38. Segmented relative CNs estimated from methylation array for samples from LIH0192.

### 5.2.4 Correlation of CCFs

To estimate CCFs for the called variants, I used PyClone which, unlike other methods, does not require allele specific CNs. While I found PyClone to be of limited accuracy in my benchmarking analysis, jointly analysing multiple samples together is likely to improve results. When run in multi-sample mode, PyClone can utilise information across samples, allowing improved estimation and clustering of CCFs, with clonal frequencies of variants able to be tracked between samples. As a comparison, I used both the multi-sample method, using only variants common to all samples, in combination with the single-sample method, considering all variants in a sample. Another way to determine conservation of clone frequencies is through directly comparing VAFs between biopsy and PDOXs. As CNs were generally well preserved between biopsy and PDOX samples, VAFs should be similar between the two, with any changes in VAFs most likely reflecting differences in clone frequencies. This avoids potential errors introduced by PyClone or CNA

calling. Furthermore, reliably comparing biopsy and PDOX samples is not too dependent on accurate CCF estimates, as long as both samples are analysed with the same pipeline.

I investigated the correlations of these metrics between biopsy and PDOX samples from the three patients separately. In patient LIH0192, the multi-sample method generally showed a good correlation in CCFs between biopsies and PDOXs, with only the second recurrent PDOX (T251) indicating selection of clones from 2 variants that increased in CCF in the PDOX compared to the originating biopsy (Figure 39C). This was also demonstrated by the single-sample method for this sample. However, the single-sample method also suggested selection was present in the first recurrent PDOX (T233), where only a few of the variants had similar CCFs between the originating biopsy and that PDOX (Figure 39B).

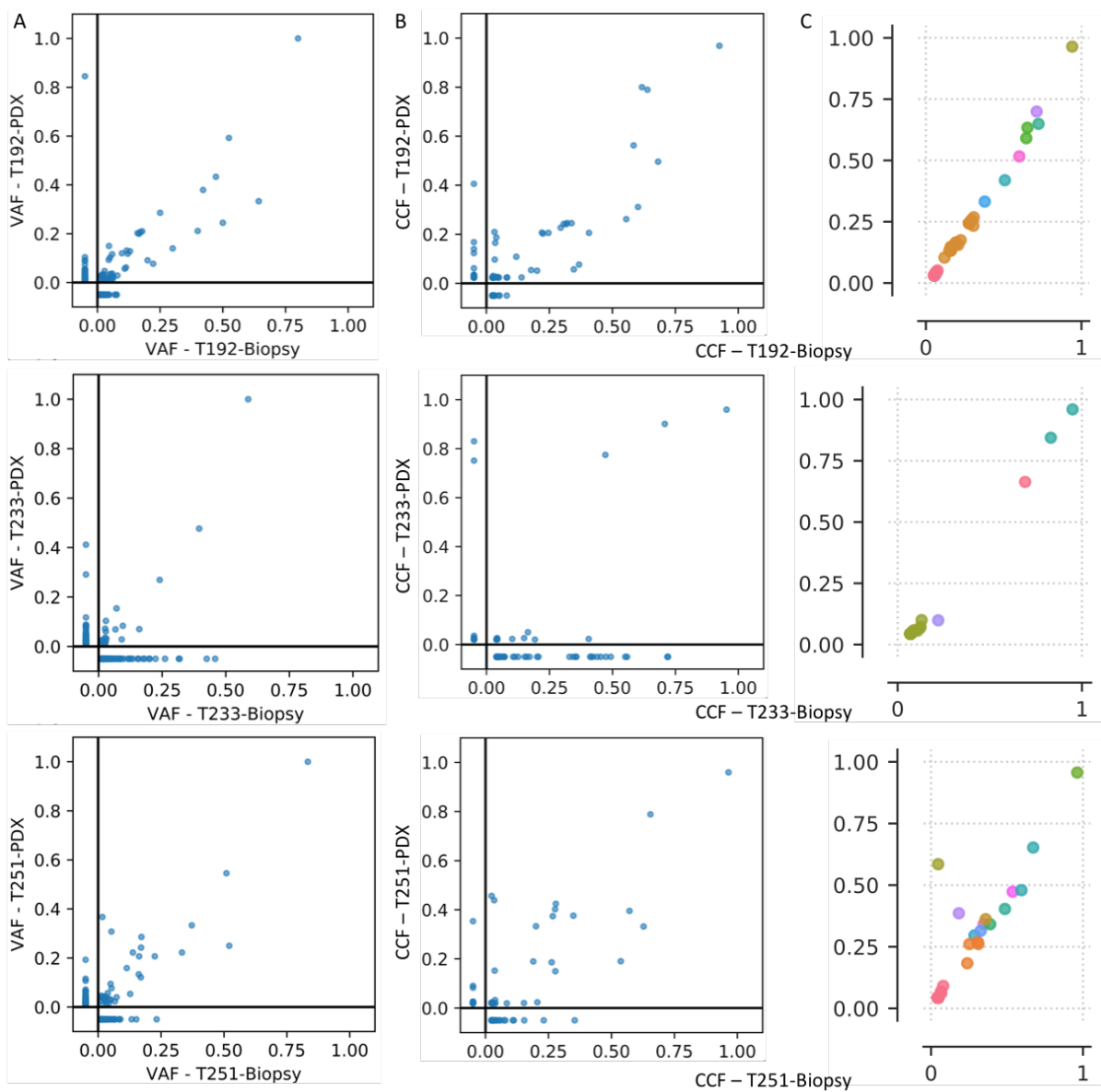


Figure 39. Correlations between corresponding biopsy and PDOX samples from patient LIH0192, in A) VAFs, B) CCFs estimated with single-sample PyClone, and C) CCFs estimated with multi-sample

PyClone. The different colours indicate PyClone's deconvolution of variants into distinct subclone clusters.

To investigate the discrepancy between the single and multi-sample method in this sample, I looked into the sequencing data more closely for any variant that had a CCF difference of  $>0.1$  between T233-Biopsy and T233-PDOX, for either method. This found that numerous variants, using the single-sample method, likely had inaccurate CCF estimates by PyClone (Table 17). For example, the variant in *HDAC9* was predicted to have CCFs of 0.41 and 0.02 in the biopsy and PDOX, respectively, showing a 20-fold difference. This variant had the same CN in both samples, and VAFs that support a 2-3 fold difference given the same CN and order of events, as was predicted using the multi-sample method (Table 17). In addition, there was a strong correlation of shared VAFs between the two samples, providing further support for the multi-sample method for T233. Nonetheless, there was still evidence of 2 variants with high CCFs in the PDOX that were absent from the biopsy, with no indication that these were due to errors in variants calling. One of these was predicted to be deleterious to the *MUC17*, though this is not thought to be expressed in the brain (Uhlén *et al.*, 2015). The other was in *EZH2*, though is not expected to be deleterious. These increases in CCFs suggests there may have been strong selection in the PDOX for clones containing these variants. The majority of variants that appeared lost in the PDOX, with relatively high CCFs in the biopsy, were in the region chr19:53990556-53991104. This is the intergenic region between *CACNG8* and *CACNG6*, and yet was covered by the sequencing target regions. It's likely that the unusually high mutation rate in this region is due to misalignments, as there are a lot of repetitive sequences within it. Many of these variants did have evidence of being present in the PDOX (from similar misalignment) but, due to a dramatic drop in coverage of the region (903 reads in the biopsy, 58 reads in the PDOX), possibly as a result of a PDOX specific CNA that was missed in the aCGH data, they did not reach the 2 supporting read minimum required by VarScan2 to be called. The primary tumour for the same patient (T192) showed a good correlation between VAFs and CCFs from both methods between biopsy and PDOX. The exception to this was a variant in *EGFR* which had an estimated CCF of 0.41 in the PDOX but was undetected in the biopsy. I therefore further investigated *EGFR* in more detail below.

Table 17. Details for variants that differed by  $>0.1$  CCF between biopsy and PDOX for T233, from either single or multi-sample PyClone estimates.

T233 Multi-sample							
Gene	Position	Allele	Biopsy CCF	PDX CCF	Predicted deleterious	Cluster	Evidence for inaccurate CCF estimate or missed variants.
HDAC9	chr7:18585269	C->T	0.22	0.10	No	3	
T233 Single-sample							
Gene	Position	Allele	Biopsy CCF	PDX CCF	Predicted deleterious	Cluster	Evidence for inaccurate CCF estimate or missed variants.
PTEN	chr10:87965536	TG->T	0.15	0.03	No	-	Similar VAF and CN in both samples.
KMT2A	chr11:118499857	A->G	0.71	0.90	No	-	1 in 4 supporting reads in PDX.
KLK1	chr19:50820511	T->G	0.49	0	No	-	
TRIM28	chr19:58542588	C->A	0.20	0	No	-	Reduced coverage in PDX: 6 vs 166.
ARID1A	chr1:26696753	T->C	0.34	0	No	-	Reduced coverage in PDX; 1 vs 83.
GNAS	chr20:58900071	A->G	0.47	0.77	No	-	2 in 4 supporting reads in PDX.
SETD2	chr3:47017539	GA->G	0.19	0.02	No	-	
FGFR3	chr4:1802955	C->G	0.72	0	No	-	VAFs and CNs suggests a 2 or 3 fold decrease.
HDAC9	chr7:18585269	C->T	0.41	0.02	No	-	
CHEK2	chr22:28687871	T->C	0.11	0	No	-	
MUC17	chr7:101033231	T->G	0	0.83	Yes	-	
EZH2	chr7:148819030	T->C	0	0.75	No	-	
There were an additional 19 variants in CACNG6 between chr19:53990556-53991104 where the CCF decreases significantly from the Biopsy to the PDX (from 0.11-0.72 to 0 in 18 variants and from 0.16 to 0.05 in the other variant). None of these were predicted to be deleterious.							

Using the multi-sample method with all six biopsy and PDOX samples from patient LIH0192 together, eight variants were included that were common to all. These clustered into three groups; near clonal variants, that remained consistently high across samples, and two subclones, that both fluctuated slightly in frequency between samples (Figure 40).



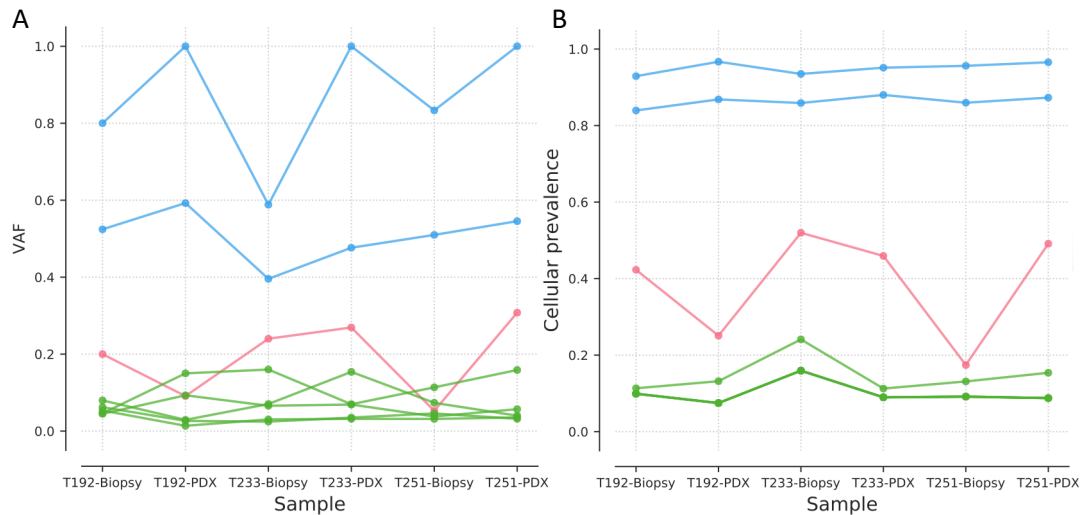


Figure 40. VAFs (A) and CCFs (B) of the 8 variants shared across all samples in patient LIH0347, using the multi-sample PyClone method. The different colours indicate PyClone’s deconvolution of variants into distinct subclone clusters.

The primary tumour in patient LIH0347 (T347) showed good correlation of CCFs between the biopsy and PDOX, especially in the multi-sample method. However, the recurrence (T470) had a poorer correlation (Figure 41), and I therefore looked into those variants further in the sequencing data. This again showed evidence of some inaccuracies with the single-sample method (Table 18), but overall, both the single and multi-sample support the presence of altered clonal frequencies in the PDOX compared to the biopsy.

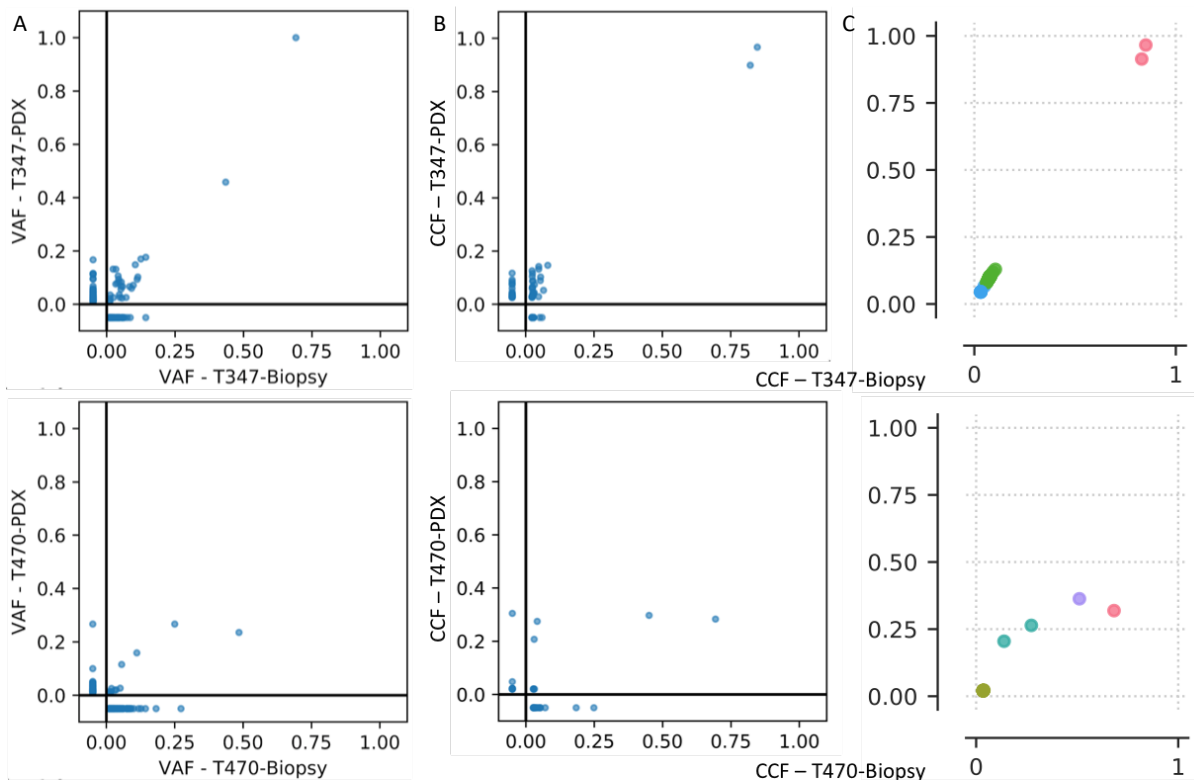


Figure 41. Correlations between corresponding biopsy and PDOX samples from patient LIH0347, in A) VAFs, B) CCFs estimated with single-sample PyClone, and C) CCFs estimated with multi-sample PyClone. The different colours indicate PyClone’s deconvolution of variants into distinct subclone clusters.

Table 17. Details for variants that differed by >0.1 CCF between biopsy and PDOX for T470, from either single or multi-sample PyClone estimates.

T470 Multi-sample							
Gene	Position	Allele	Biopsy CCF	PDX CCF	Predicted deleterious	Cluster	Evidence for inaccurate CCF estimate or missed variants.
PTEN	chr10:87961002	TG->T	0.68	0.32	Yes	0	
GSE1	chr16:85654423	T->G	0.36	0.51	Yes	3	
T470 Single-sample							
Gene	Position	Allele	Biopsy CCF	PDX CCF	Predicted deleterious	Cluster	Evidence for inaccurate CCF estimate or missed variants.
PTEN	chr10:87961002	TG->T	0.69	0.28	Yes	-	1 in 4 supporting reads in the PDX  Similar VAF and CN in both samples  VAFs and CNs suggests a 2 fold increase.  1 in 18 supporting reads in the Biopsy
SNORD8	chr14:21397813	T->C	0.18	0	No	-	
GSE1	chr16:85654423	T->G	0.45	0.30	Yes	-	
GNAS	chr20:58900071	A->G	0.25	0	No	-	
FOXO3	chr6:108561803	G->A	0.04	0.27	Yes	-	
NOTCH2	chr1:119915219	C->T	0.03	0.21	No	-	
AKT2	chr19:40236049	A->C	0	0.30	Yes	-	

When using all four samples from patient LIH0347 together in the multi-sample method, four shared variants were clustered into three groups, which altered in frequency between the primary and recurrent tumour, though were mostly conserved between biopsies and corresponding PDOXs (Figure 42).

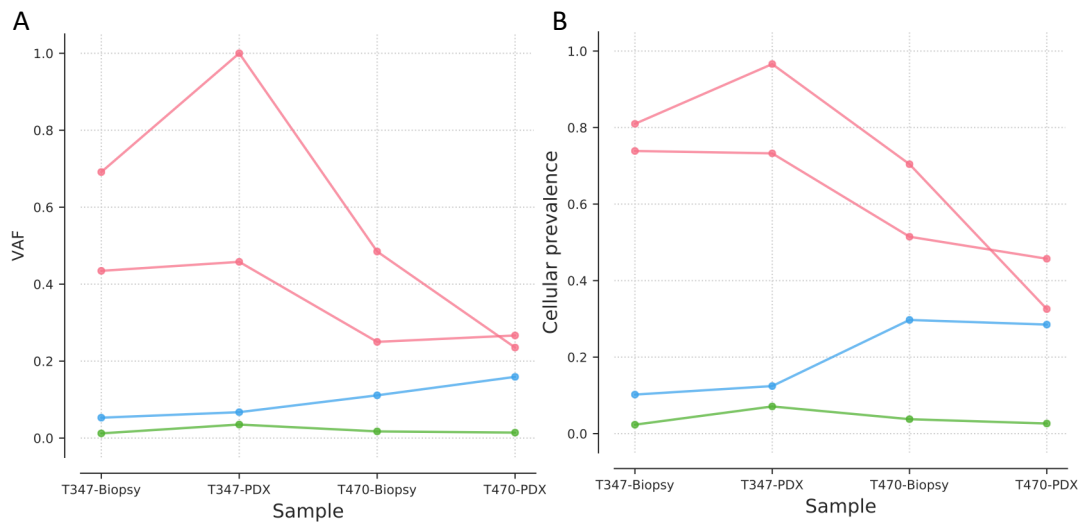


Figure 42. VAFs (A) and CCFs (B) of variants across samples in patient LIH0192, using the multi-sample PyClone method with only variants common to all samples. The different colours indicate PyClone’s deconvolution of variants into distinct subclone clusters.

Patient LIH0158 showed very low CCF correlation between the primary (T158) biopsy and PDOX across both methods (Figure 43). Again, the single-sample method had many probable inaccuracies in CCF estimates, but with none found using the multi-sample method (Table 18). Therefore, altered clonal frequencies between biopsy and PDOX is indicated.

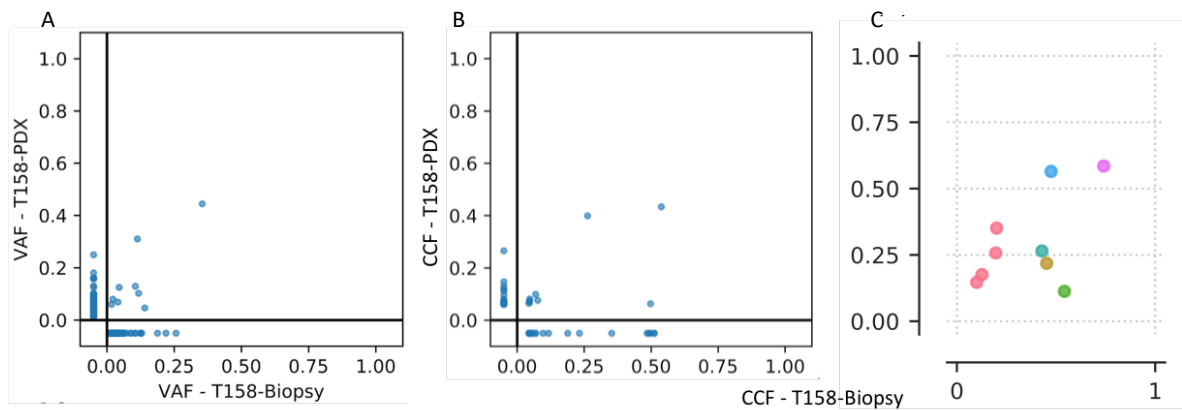


Figure 43. Correlations between corresponding biopsy and PDOX samples from patient LIH0347, in A) VAFs, B) CCFs estimated with single-sample PyClone, and C) CCFs estimated with multi-sample PyClone. The different colours indicate PyClone’s deconvolution of variants into distinct subclone clusters.

Table 18. Details for variants that differed by  $>0.1$  CCF between biopsy and PDOX for T158, from either single or multi-sample PyClone estimates.

T158 Multi-sample							
Gene	Position	Allele	Biopsy CCF	PDX CCF	Predicted deleterious	Cluster	Evidence for inaccurate CCF estimate or missed variants.
ZBTB45	chr19:58542517	A->G	0.43	0.26	No	3	
ZBTB45	chr19:58542521	T->C	0.45	0.22	No	1	
FOXO3	chr6:108561803	G->A	0.54	0.11	Yes	2	
AKAP9	chr7:92000991	A->T	0.74	0.58	Yes	5	
NOTCH1	chr9:136505644	C->T	0.20	0.35	No	0	
T158 Single-sample							
Gene	Position	Allele	Biopsy CCF	PDX CCF	Predicted deleterious	Cluster	Evidence for inaccurate CCF estimate or missed variants.
PTEN	chr10:87965536	CT->C	0.48	0	No	-	Reduced coverage in PDX; 9 vs 156
CIC	chr19:42287787	G->A	0.51	0	No	-	
CACNG6	chr19:53990626	A->G	0.23	0	No	-	
CACNG6	chr19:53990870	C->T	0	0.12	No	-	Reduced coverage in Biopsy; 4 vs 37
CACNG6	chr19:53990969	T->C	0	0.11	No	-	Reduced coverage in Biopsy; 9 vs 38
PLCG1	chr20:41166758	CTTG->C	0.26	0.40	No	-	
GNAS	chr20:58900071	A->G	0.19	0	No	-	1 in 8 supporting reads in the PDX
NF2	chr22:29604032	A->T	0.50	0	Yes	-	Reduced coverage in PDX; 27 vs 136
ZBTB20	chr3:114339432	G->A	0.49	0	No	-	Reduced coverage in PDX; 16 vs 72
PIK3R1	chr5:68293310	G->A	0.51	0	Yes	-	
FOXO3	chr6:108561803	G->A	0.50	0.06	Yes	-	
HDAC9	chr7:18585269	C->T	0.12	0	No	-	1 in 25 supporting reads in PDX
AKAP9	chr7:92000991	A->T	0.54	0.43	Yes	-	
BCOR	chrX:40050038	T->G	0.35	0	No	-	Reduced coverage in PDX; 18 vs 94
CHD8	chr14:21437061	C->G	0	0.13	No	-	
COL6A3	chr2:237387842	A->G	0	0.15	Yes	-	
FGFR3	chr4:1805767	G->T	0	0.27	Yes	-	

### 5.2.5 Further investigation of *EGFR*

Due to the unusual pattern seen for *EGFR* in T192, where it had a CCF of 0.41 in the PDOX but was undetected in the patient biopsy, contrasting all other variants which showed a very strong correlation between the two samples, I investigated this gene further.

The above copy number analysis shows that focal amplification of *EGFR* in chr7 is present in all samples, apart from T158\_PDX. Sequencing data confirmed that coverage across *EGFR* (chr7:55,017,021-55,210,080) was nearly 5 fold lower in the PDOX (46151 reads) compared to the biopsy (220245), thereby confirming the lack, or reduction, of amplification of *EGFR* in the PDOX.

It was shown by our collaborators, through Western blot and aCGH, that LIH0192 and LIH0347 have structural *EGFR* variants that were retained in the respective PDOXs. The LIH0192 primary (T192) expressed VII ( $\Delta$ 14-15), whereas the second recurrence (T251) instead expressed VIII, with both expressing low levels of wildtype (wt). LIH0347 expressed  $\Delta$ 2-15 in both the primary (T347) and first recurrence (T470), although wt expression was the most prevalent in both.

I first sought to provide further evidence of these variants, and confirm the correlation between biopsy and PDOX samples, by comparing sequencing depths across exons in *EGFR* (Figure 44). LIH0192 showed a high frequency deletion of exons 14-15 in the biopsy and PDOX from the primary, confirming the expression of a high frequency VII variant. Reduced coverage of exons 2-7 was apparent in the first and second recurrences, supporting the presence of VIII (Figure 44A). In LIH0347, evidence of *EGFR* structural variants in any sample was not obvious from the sequencing reads, presumably due to their low frequencies compared to wt expression (Figure 44B). This analysis supports previous observations that these PDOX models retain the *EGFR* structural variants present in the biopsies.

I next looked at point variants in *EGFR* in samples from LIH0192. As expected from the aCGH results, *EGFR* had a very high coverage in all samples (Table 19), which may suggest the presence of extrachromosomal DNA (ecDNA). I therefore compared the allele frequencies of both germline and somatic variants in this region to investigate patterns of conservation of these potential ecDNAs, both longitudinally and from biopsy to PDOX, in patient LIH0192. In general, *EGFR* variants fluctuated in VAF considerably between samples (Table 19). Two variants in particular stood out; chr7:55168634 G>C had 3/984 reads supporting the alternative allele in the primary biopsy, which increased to 988/1169 in the primary PDOX. No supporting reads were found in the recurrent samples. Another variant, chr7: 55154017 C->T, had a similar VAF in the primary biopsy (330/5596) and PDOX (375/6287) samples, but was absent, or near absent, in the recurrent samples (Table 19). This variant was filtered out of the final list of somatic variants included in further analysis, as it was incorrectly (evidenced by the lack of supporting reads in some samples) classed as a germline variant due to being at a known polymorphic site.

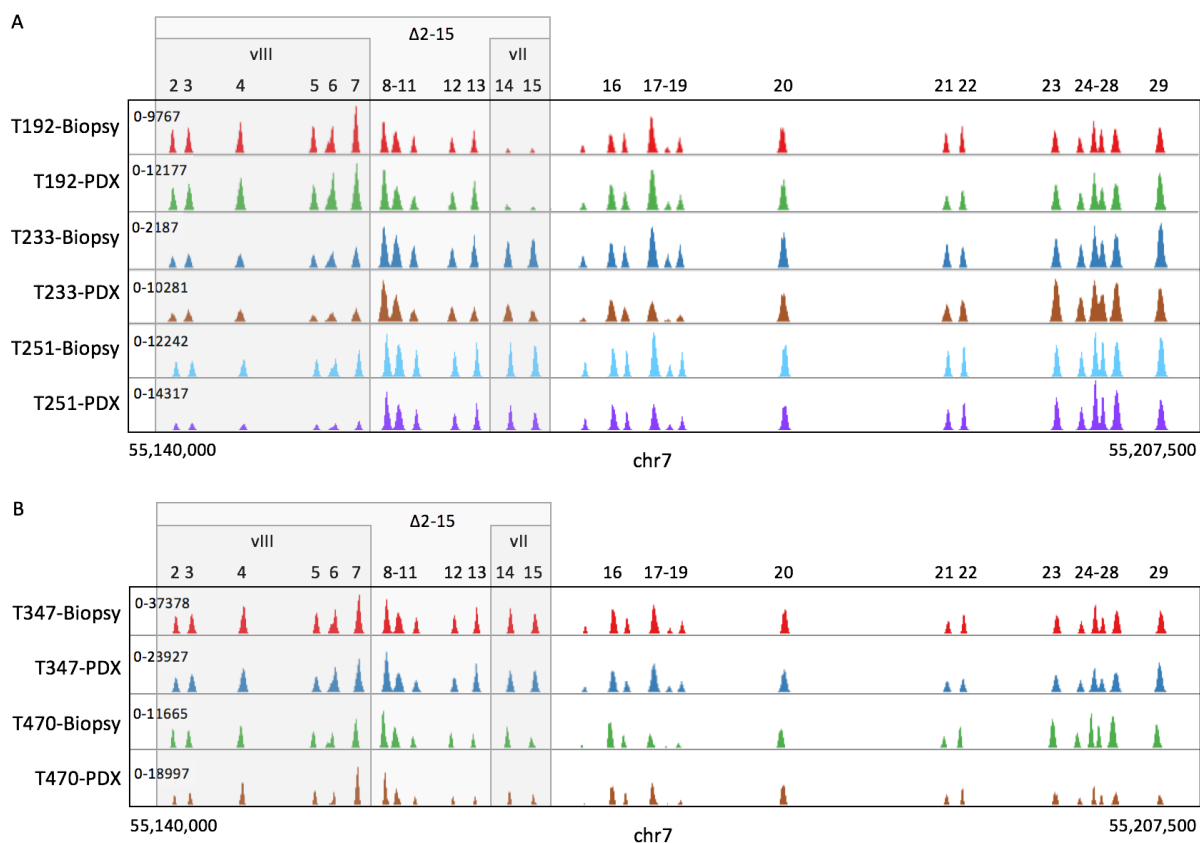


Figure 44. Sequencing coverages across exons in EGFR, demonstrating the relative frequencies of VII (deletion of exons 14-15), VIII (deletions of exons 2-7) and  $\Delta$ 2-15 (deletion of exons 2-15) variants, in samples from patients A) LIH0192 and B) LIH0347.

Table 19. Total read depth and number of alternate reads for two variants in EGFR with high VAF fluctuations in samples from LIH0192.

Variant EGFR	chr7:55168634 G>C		chr7: 55154017 C->T	
	Read depth	Alt reads	Read depth	Alt reads
T192_Biopsy	984	3	5966	330
T192_PDOX	1169	988	6287	375
T233_patient	401	0	705	0
T233_PDOX	625	0	1346	0
T251_patient	2245	0	4422	2
T251_PDOX	1826	0	1865	1

### 5.3 Discussion

PDOX tumour models are used by researchers as a way to experiment on representations of patient tumours, without such limitations of tissue availability. Their use, however, is dependent on the assumption that the models sufficiently recapitulate relevant aspects of patient tumour biology, a factor that has proven problematic with previous models (Bolhaqueiro *et al.*, 2019; Knouse *et al.*, 2018; Ben-David *et al.*, 2017). To address this issue, our collaborators have developed PDOX models derived from short term

culture of mechanically minced GBM tumour tissue, and found that they do conserve much of the biology of the originating patient biopsies. In this study, I provided further evidence of maintained biology between most, but not all, patient GBM biopsies and corresponding mouse PDOXs, through analysis of copy number alterations and variant frequencies.

I used PyClone to perform subclonal deconvolution and assess the maintenance of CCFs from biopsies to corresponding PDOXs. Through benchmarking in Chapter 3, I show that PyClone has limited accuracy when run on single samples, but it's likely that this improves when including multiple samples. To investigate this, I ran it on both single and multiple samples, and compared the results. Unlike the single-sample method, the multi-sample method showed a good correlation of CCFs between biopsies and PDOXs for most samples, as did a direct comparison of VAFs. Therefore, despite the poor performance of PyClone in Chapter 3, it's likely that the multi-sample method allows for a reliable assessment of CCF correlations between samples due to being able to pool information across them.

Such an analysis indicated that clonal frequencies were well preserved between biopsy and PDOX for two primary tumours (T192 and T233) and one recurrent tumour (T347). There was reasonable correlation in another recurrence (T251), though with possible selection indicated by two variants, and poor correlation in a primary and recurrence (T470 and T158), indicating significant changes in clonal frequencies. However, it is not known whether these deviations result from differential selection of clones in the PDOXs, or instead an effect of sampling bias. The latter is possible if the sequencing was from a distinct part of the biopsy than was used to create the PDOX, resulting in a different make up of clones due to spatial intratumour heterogeneity (Sun *et al.*, 2017; Siegmund and Shibata, 2016). If sampling bias is the cause of the poor correlation between some biopsies and PDOXs, it would not be a large concern, as the same issue would be encountered with any application of the biopsy. On the other hand, if the cause were instead due to differential selection of subclones, this would be more problematic as it reflects an alteration due to the mouse micro-environment. It would be interesting to apply the SubClonalSelection model (Williams *et al.*, 2018), described in Chapter 4, to the variants from these samples, in order to assess their mode of evolution. This would allow us to determine if selection is more prevalent in the PDOX samples than their corresponding biopsies, which would indicate if differences in clonal frequencies are a result of selection due to xenografting, or instead due to sampling bias. Unfortunately, there are insufficient numbers of variants from the targeted sequencing (with an appropriate number of supporting reads) to generate a reliable result from the model. While samples from the GLASS dataset used in Chapter 4 were also not ideal for its use either (though they generally had a more appropriate number and distribution of VAFs than samples in this chapter), the aim of investigating the overall pattern of selection across samples in a large cohort meant that noise in the model's results were less of an issue in that analysis. However, when looking at individual cases, it is important to be confident that each result is reliable, and so I was not able to use the model to compare the mode of evolution in biopsies and PDOXs.

T192 showed a very high correlation of CCFs between biopsy and PDOX, though with the exception of *EGFR* that increased dramatically in the PDOX, prompting me to investigate this further. Coverages across *EGFR* exons showed the presence of structural variants in the biopsies from LIH0192, and these were conserved in the corresponding PDOXs. In contrast, and despite showing very high correlation of CCFs for other genes, allele frequencies of point variants in *EGFR* demonstrated very different proportions of copies of *EGFR* between the T192 biopsy and PDOX. Such high coverage, combined with high variation in frequencies of point variants across *EGFR*, suggests the presence of ecDNA, which is particularly common in aggressive tumours such as GBM (Kim *et al.*, 2020; Turner *et al.*, 2017) and known to commonly include additional copies of *EGFR* (Vogt *et al.*, 2004; Decarvalho *et al.*, 2018). Divergent inheritance of ecDNA, in spite of conservation of chromosomal genetic clonal frequencies, has been previously seen in GBM (Decarvalho *et al.*, 2018), and represents an additional level of heterogeneity. The results in this study, assuming that the *EGFR* amplification in T192 is due to ecDNA, support this observation, and show that numbers of ecDNA are not always recapitulated in corresponding PDOX models, even when all other variants show a strong conservation of clonal frequencies. This could have implications on how accurately PDOXs reflect the originating biopsies, given that oncogenes on ecDNA are associated with increased aggressiveness of tumours (Kim *et al.*, 2020), though this again may be explained by sampling bias.

This study had several limitations that may have affected its accuracy. The lack of a germline reference meant that I was not able to reliably distinguish between somatic and germline point variants, instead relying on whether positions were previously known to have germline polymorphisms. Whilst this should have resulted in the removal of the majority of germline variants from the called set, some may still have been left, which could impact PyClone's calculations due to conflicts with the purity estimates. Additionally, some somatic variants may have been incorrectly classed as germline and therefore missed, as was likely the case with chr7: 55154017 C->T, evidenced by a lack of supporting reads in the T233 and T251 samples. PyClone requires absolute bulk CN estimates for variants. As discussed in previous chapters, this is an unrealistic scenario when GBM is known to have subclonal CNAs. While this isn't much of an issue when pairs of biopsy and PDOXs are analysed with equally inaccurate CNs, rounding to the most likely absolute value allows the introduction of error that could differ between samples and lead to inconsistent results. Another potential limitation of the study is that the aCGH data had substantial noise from both frequent signal dropouts and probes filtered out due to poor quality. Signal dropouts, likely resulting from insufficient probe hybridisation in regions where the patient doesn't match the reference perfectly, result in an apparent CN of 0 for those probes. If surrounding probes are not present to help counteract that lack of signal then the segmentation step results in much larger regions with apparent CNs of 0. While this was largely overcome with a more stringent segmentation step combined with an override of CN=2 in the worst affected chromosomes, many regions with a CN of 0 still remained in some samples.



Purity estimates were calculated using a custom method. Whilst this relied on uncertain assumptions based only on prior knowledge of GBM CNAs and visual inspection of data across samples, it seemed the most reliable method in these cases. Although not feasible in larger cohorts, this manual curation of results is appropriate when only analysing a small number of samples, especially given that all samples are analysed with the same script, thereby minimising bias. The strong correlation seen between some biopsy and corresponding PDOXs suggests that any noise or error introduced when processing the data were not sufficient to heavily impact the analysis.

In conclusion, these results provide evidence that the PDOX models, created without cell dissociation or in vitro passage, largely preserve the clonal structure of the originating GBM patient biopsies in some tumours. Others showed significant differences in clonal frequencies, which might be explained by either sampling bias or from PDOX specific clonal selection.

## 5.4 Methods

### 5.4.1 Variant calling

Samples were previously sequenced using targeted panels of 181-234 genes, adapted from the Heidelberg brain tumour panel (Sahm *et al.*, 2016), with 75bp read lengths and average sequencing depth between 100x-250x. These reads were previously depleted of mouse specific reads and aligned to the hg38 human reference genome (Golebiewska *et al.*, 2020). I then processed these with Picard MarkDuplicates (Broad Institute), and GATK (v3.8) indel realignment and base recalibration (Van der Auwera *et al.*, 2013). Variants were called using Samtools' (v1.9) mpileup (Li *et al.*, 2009) and Varscan 2's (v2.4.4) pileup2snp and mpileup2indel commands (Koboldt *et al.*, 2012), using a p-value threshold of 1, and otherwise default parameters (although I specifically set these as I've found other Varscan functions' defaults sometimes differ from what is stated in the manual): `--strand-filter 0 --min-freq-for-hom 0.75 --min-coverage 8 --min-reads2 2 --min-avg-qual 15 --min-var-freq 0.01 -p-value 1`. A custom script was used to filter out variants listed in the dbSNP database of known polymorphic sites (Sherry *et al.*, 1999). Bedtools (v2.29.2) (Quinlan and Hall, 2010) was then used to reduce variants to only those covered by the 150 gene panel target regions, which was converted from hg19 to hg38 using hgLiftOver (UCSC hgLiftOver). Coverage graphs were created with the use of the Interactive Genome Viewer (Thorvaldsdóttir *et al.*, 2013).

Deleterious variants were identified using the Ensembl Variant Effect Predictor (McLaren *et al.*, 2016), by those annotated with "IMPACT=HIGH", or where SIFT (Vaser *et al.*, 2015) or PolyPhen-2 (Adzhubei *et al.*, 2010) predicted a variant was "deleterious", "deleterious\_low\_confidence", "probably\_damaging", or "possibly\_damaging".

## 5.4.2 Copy number calling

Log2 ratios from aCGH probes were segmented using DNACopy (v1.52.0) (Seshan E. and Olshen, 2019) using default parameters, with the exception of the addition of 'undo.splits="sdundo",undo.SD=4' to the segment command, which was required to reduce a high level of noise from frequent signal dropouts. Absolute CNs were then estimated from these using the methods described in the results. Probe positions were converted from hg19 to hg38 using hgLiftOver (UCSC hgLiftOver).

Methylation array data was processed using cnAnalysis450k (Knoll *et al.*, 2017) (v0.99.26) and minfi (Aryee *et al.*, 2014) (v1.24.0), with 119 publicly available normal brain samples (Capper *et al.*, 2018) used as controls for normalisation. The genomic coordinates of CpG probes were taken from the GPL13534 reference in the NCBI Gene Ontology Omnibus database (Barrett *et al.*, 2013).

## 5.4.3 Correlation of CCFs

PyClone (v0.13.1) was run under default parameters, with the addition of '--prior total\_copy\_number', through the 'run\_analysis\_pipeline' command to run the full workflow. Purities were taken from the aCGH custom estimates for the biopsies, and set to 1 for the PDOX samples as these are not likely to contain normal human cells, and sequencing reads from mouse normal cells had previously been removed.

## 5.5 References

- Adzhubei, I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
- Aryee, M.J. *et al.* (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. **30**, 1363–1369.
- Auton, A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
- Van der Auwera, G.A. *et al.* (2013) From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.*
- Barrett, T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.*, D991–D995.
- Barthel, F.P. *et al.* (2019) Longitudinal molecular trajectories of diffuse glioma in adults. *Nature*, **576**, 112–120.
- Ben-David, U. *et al.* (2017) Patient-derived xenografts undergo mouse-specific tumor evolution. *Nat. Genet.*, **49**, 1567–1575.
- Bolhaqueiro, A.C.F. *et al.* (2019) Ongoing chromosomal instability and karyotype evolution in human colorectal cancer organoids. *Nat. Genet.*, **51**, 824–834.
- Broad Institute Picard Tools - By Broad Institute.
- Capper, D. *et al.* (2018) DNA methylation-based classification of central nervous system tumours. *Nature*, **555**, 469–474.
- Decarvalho, A.C. *et al.* (2018) Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. *Nat. Genet.*, **50**, 708–717.
- Gerstung, M. *et al.* (2020) The evolutionary history of 2,658 cancers. *Nature*, **578**, 122–128.
- Golebiewska, A. *et al.* (2020) Patient-derived organoids and orthotopic xenografts of primary and recurrent gliomas represent relevant patient avatars for precision oncology. *Acta Neuropathol.*, **10**, 16.
- Joo, K.M. *et al.* (2013) Patient-Specific Orthotopic Glioblastoma Xenograft Models Recapitulate the Histopathology and Biology of Human Glioblastomas In Situ. *Cell Rep.*, **3**, 260–273.

- Kim,H. *et al.* (2020) Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat. Genet.*, 1–7.
- Knoll,M. *et al.* (2017) cnAnalysis450k: an R package for comparative analysis of 450k/EPIC Illumina methylation array derived copy number data. *Bioinformatics*, **33**, 2266–2272.
- Knouse,K.A. *et al.* (2018) Chromosome Segregation Fidelity in Epithelia Requires Tissue Architecture. *Cell*, **175**, 200-211.e13.
- Koboldt,D.C. *et al.* (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
- Körber,V. *et al.* (2019) Evolutionary Trajectories of IDH WT Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis. *Cancer Cell*, **35**, 692-704.e12.
- Lai,Y. *et al.* (2017) Current status and perspectives of patient-derived xenograft models in cancer research. *J. Hematol. Oncol.*, **10**, 1–14.
- Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–9.
- McLaren,W. *et al.* (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.
- McNulty,S.N. *et al.* (2019) Beyond sequence variation: assessment of copy number variation in adult glioblastoma through targeted tumor somatic profiling. *Hum. Pathol.*, **86**, 170–181.
- Quinlan,A.R. and Hall,I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
- Sahm,F. *et al.* (2016) Next-generation sequencing in routine brain tumor diagnostics enables an integrated diagnosis and identifies actionable targets. *Acta Neuropathol.*, **131**, 903–910.
- Seshan E.,V. and Olshen,A. (2019) DNACopy: DNA copy number data analysis.
- Shen,H. *et al.* (2013) Comprehensive Characterization of Human Genome Variation by High Coverage Whole-Genome Sequencing of Forty Four Caucasians. *PLoS One*, **8**, e59494.
- Sherry,S.T. *et al.* (1999) dbSNP - database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.*, **9**, 677–679.
- Siegmund,K. and Shibata,D. (2016) At least two well-spaced samples are needed to genotype a solid tumor. *BMC Cancer*, **16**, 250.
- Sun,R. *et al.* (2017) Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.*, **49**, 1015–1024.
- Thorvaldsdóttir,H. *et al.* (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
- Turner,K.M. *et al.* (2017) Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature*, **543**, 122–125.
- UCSC hgLiftOver.
- Uhlén,M. *et al.* (2015) Tissue-based map of the human proteome. *Science (80-. )*, **347**.
- Vaser,R. *et al.* (2015) SIFT missense predictions for genomes.
- Vogt,N. *et al.* (2004) Molecular structure of double-minute chromosomes bearing amplified copies of the epidermal growth factor receptor gene in gliomas. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 11368–11373.
- Williams,M.J. *et al.* (2018) Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.*, **50**, 895–903.
- Yoshida,G.J. (2020) Applications of patient-derived tumor xenograft models and tumor organoids. *J. Hematol. Oncol.*, **13**, 1–16.

## Chapter 6 - Discussion

Glioblastoma (GBM) is one of the deadliest of all cancers (Johnson and O'Neill, 2012; Stupp *et al.*, 2017). This is largely due to the almost inevitable recurrence of tumours as a result of therapy resistance. In this study, I aimed to use genomic datasets from matched primary and recurrent tumours, to identify cellular processes that may influence this resistance in GBM, and potentially provide us with new targets for treating it.

Such a goal requires delineating the genetic intratumour heterogeneity (ITH) of tumours, that results from their continuous gain of new mutations and evolution over time. As it was unclear what the most suitable pipelines were for achieving this, I sought to benchmark a range of methods involved in the process, to identify those most accurate. This required developing the programs HeteroGenesis and w-Wessim, to allow simulation of realistically complex artificial whole exome sequencing (WES) datasets, with known ground truths. HeteroGenesis is a much more flexible and advanced somatic genome simulator than previously available options. It should therefore prove useful to other researchers wanting to simulate somatic genomes. The *in silico* WES sequencer w-Wessim has important improvements from its predecessor, Wessim (Kim *et al.*, 2013), that enable it to model CNVs as well as providing more realistic read distributions. This, also, is likely to be useful to other researchers. Additionally, I plan to make datasets, generated with these methods, available for researchers to download and use for their own purposes. To my knowledge, no other genomes or WES datasets for such realistically complex artificial bulk tumour samples have been created elsewhere.

The simulated reads generated by w-Wessim have a limitation in that the quality scores assigned to error bases are higher than those of real data. It's possible this resulted in the higher than expected false positive variant calls. While this was largely accounted for by subsampling the false positives in the call sets before using them with subclonal deconvolution methods, it would be beneficial to improve the simulated datasets to address this issue. This could be achieved either by manually adjusting the trained error model that I used with w-Wessim, or through collaborations with authors of newer *in silico* sequencers, which could be adapted to include the WES specific features of w-Wessim. I also plan to minimise w-Wessim's resource requirements by trialling it with reduced probe input numbers.

In Chapter 3, I carry out benchmarking of subclonal deconvolution methods, as well as mutation calling methods required to generate their inputs. This identified Mutect2 as the most suitable point variant caller, FACETS as the most accurate copy number aberration (CNA) caller, and Ccube as the most accurate subclonal deconvolution method, when using a single tumour sample with a matched normal. However, even the best subclonal deconvolution method showed poor accuracy in estimating variant cancer cell fractions (CCFs). Combined with the fact that single tumour samples are unlikely to be representative of the

whole tumour, this suggests that use of single samples for subclonal deconvolution is not to be recommended (Bhandari *et al.*, 2018; Siegmund and Shibata, 2016; Watkins and Schwarz, 2018; Sun *et al.*, 2017; Chkhaidze *et al.*, 2019).

Further benchmarking to assess to what extent multi-sample set-ups improve the accuracy of subclonal deconvolution, is required. Such analysis would be able to use the datasets created in this study, with simple adjustments; Reads from each subclone within a tumour can be merged in varying proportions to form samples representing multiple distinct regions of a tumour. Additionally, HeteroGenesis allows for easy re-calculation of the mutation profiles of individual subclones, to reflect samples with differing subclone frequencies.

One of the challenges with identifying treatment resistance mechanisms in GBM, is the scarcity of sequencing datasets of matched primary and recurrent tumours. This results from the fact that only 25% of recurrent GBMs are eligible for a second surgery, and those that are often contain extensive necrosis, rendering them unsuitable for sequencing (Kraboth and Kalman, 2020). Our group had accumulated paired GBM data for 18 in-house patient samples, combined with a further 42 patients from other cohorts, for which WES data was available. I planned to analyse these once I had determined the most suitable pipelines. However, whilst I was working on the benchmarking analysis, a novel and much larger dataset became available, consisting of 94 high quality IDHwt GBM patients who underwent standard therapy (GLASS Consortium, 2018). This was a valuable resource and, despite the lack of raw data availability, I adapted my plans in order to utilise it in this study.

By applying the SubClonalSelection model (Williams *et al.*, 2018), I showed that selection in GBM is not associated with therapy across the GLASS cohort. However, I was able to identify variants that may be driving increased resistance to therapy in a minority of patients. Such observations are still possible alongside the findings of the model, which is unable to detect all instances of subclonal expansions, and may also have suffered from poor accuracy owing to sub-optimal data quality in many of the included samples.

The power of the pathway analysis was limited by the sample size, but nonetheless, it highlighted some promising areas for further investigation. In particular, alteration of the bile acid synthesis was identified as a promising candidate for conferring therapy resistance, whereas alteration to the tight junction interactions pathway was identified as a candidate for sensitising GBM cells to therapy. Computational investigations with additional omics datasets, followed by functional assays, are warranted to further investigate the highlighted pathways for their role in GBM survival through therapy. The pathway analysis also provides the foundation for future work planned for our group, whereby our in-house paired samples will be fed through the same pipeline. This will include the original collection of WES samples, combined with ~40 new pairs, for which whole-genome sequencing has recently been provided to us by MD Anderson

as part of their cancer moon shot programme. By having access to the raw data for these samples, we are able to apply optimised, and likely more accurate pipelines for inferring changes in CCFs. This may either include a multi-sample subclonal deconvolution method, if one is identified as being highly accurate in the planned future benchmarking, or alternatively, not applying subclonal deconvolution and instead 1) applying FACETS (identified as the best CNA caller in Chapter 3) to call CNAs on the recurrent samples, whilst using primary samples as matched normal samples, 2) masking variants in regions where CNA differences exist, and 3) directly inferring CCFs from purity adjusted VAFs. This is likely an effective approach given that recurrent GBMs rarely show extensive new structural changes since the primary (Barthel *et al.*, 2019), and may provide further support for the pathways that stood out in the current study, as potentially influencing GBM progression through therapy.

A further important use for characterising genetic ITH in tumours, is in assessing whether tumour models used in the lab accurately recapitulate the subclonal architecture of patient biopsies. In Chapter 5, I performed variant subclonal deconvolution to show that subclone frequencies are largely similar between biopsies and patient-derived orthotopic xenograft mouse models (PDOXs), created without cell dissociation or in vitro passage. Previous evidence supports that these models also maintain much of the non-genetic biology of the patient biopsies. However, as the transcriptional profiles of PDOXs clustered separately to those of patient biopsies, there should still be a level of caution in interpreting results from the use of the models. Nonetheless, they represent a significant improvement from those previously available, and have been shown to reflect the treatment response of the original patient tumour.

Ultimately, it is possible that no single target will be found by researchers that can be moderated by drugs to prevent GBM recurrence. In such a scenario, we must turn to other approaches to improve prognosis for patients. Novel tumour evolution-based strategies have been proposed for the multiple cancer types that inevitably recur, with a focus on maintaining control of resistant cell populations (Walther *et al.*, 2015). The traditional treatment approach of aiming to eliminate as many cancerous cells as possible, specifically remove treatment-sensitive cells, leaving behind resistant cells with reduced competition for space and resources. This “competitive release” enables resistant cells to rapidly proliferate and dominate the tumour. In contrast, evolution-based strategies aim to keep resistant cell numbers down whilst preventing growth of the overall tumour. One such approach is termed ‘adaptive therapy’ and involves continuous cycles of low-dose therapy so that proliferation alternates between resistant and sensitive cells; when the drug is on, sensitive cells slowly reduce in number and allow resistant cells to slowly increase, and when the drug is off, the sensitive cells outcompete the resistant cells which have a lower fitness in the absence of therapy (Yamamoto *et al.*, 2018; Enriquez-Navas *et al.*, 2015; West *et al.*, 2020; Gatenby *et al.*, 2009). This reduced fitness comes from the fact that drug resistance is often a costly process to the cell and so, in the absence of therapy, reduces cell proliferation. In continuously targeting these different cell populations, adaptive therapy maintains tumour size. Furthermore, the reduced therapy dosage means patients

experience less toxicity and side effects. Adaptive therapy has been successful in phase 2 clinical trials of abiraterone management of metastatic castration-resistant prostate cancer, where median progression free survival increased from 16.2 months under standard treatment, to at least 30 months, and with less than half the drug usage (Zhang *et al.*, 2019, 2017). Another evolution-based strategy is ‘evolutionary herding’, and involves using one drug to induce the tumour into developing resistance to it, but in doing so, results in hypersensitivity to another; a scenario known as ‘antagonistic pleiotropy’ (Zhao *et al.*, 2016; Noorani *et al.*, 2020; Acar *et al.*, 2019). For example, in acute myeloid leukaemia, resistance to the bromodomain inhibitors through polycomb repressive complex 2-NSD2/3-mediated MYC regulation, induce sensitivity to BCL2 inhibitors (Lin *et al.*, 2020; Fiskus *et al.*, 2019). Such evolution based treatment approaches further highlight the need to be able to accurately characterise genetic ITH in tumours.

Overall, this study provides: 1) Resources for simulating artificial tumour genomes and WES sequencing datasets. These will be of use to other researchers wanting to test any of a number of different types of genome analysis methods, thereby facilitating method improvements and allowing researchers to determine the best pipelines to use. 2) A guide for researchers on the accuracy of mutation calling and subclonal deconvolution methods for characterising genetic ITH in single tumour samples with matched normals. This will aid researchers in planning their experiments, not only in identifying the most suitable analysis methods to apply, but also in highlighting the potential importance of prioritising multi-sampling approaches when possible. 3) A list of pathways that are candidates for conferring increased resistance or sensitivity to treatment in GBM. These warrant further computational investigation followed by validation in the lab of their role in therapy resistance in GBM. As current approaches in treating GBM show little survival benefit and result in rapid tumour adaptation, it is vital that the issue of therapy resistance is addressed in future treatment approaches. The mechanisms identified in this study, that are candidates for driving or reducing such therapy resistance, may therefore lead to the much needed breakthrough in treatment of GBM.

## References

- Acar,A. *et al.* (2019) Exploiting evolutionary herding to control drug resistance in cancer. *bioRxiv*, 566950.
- Barthel,F.P. *et al.* (2019) Longitudinal molecular trajectories of diffuse glioma in adults. *Nature*, **576**, 112–120.
- Bhandari,V. *et al.* (2018) Quantifying the Influence of Mutation Detection on Tumour Subclonal Reconstruction. *bioRxiv*, 418780.
- Chkhaidze,K. *et al.* (2019) Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *bioRxiv*, 544536.
- Enriquez-Navas,P.M. *et al.* (2015) Application of evolutionary principles to cancer therapy. *Cancer Res.*, **75**, 4675–4680.
- Fiskus,W. *et al.* (2019) Superior efficacy of cotreatment with BET protein inhibitor and BCL2 or MCL1 inhibitor against AML blast progenitor cells. *Blood Cancer J.*, **9**.
- Gatenby,R.A. *et al.* (2009) Adaptive therapy. *Cancer Res.*, **69**, 4894–4903.
- GLASS Consortium (2018) Glioma through the looking GLASS: Molecular evolution of diffuse gliomas and

- the Glioma Longitudinal Analysis Consortium. *Neuro. Oncol.*, **20**, 873–884.
- Johnson,D.R. and O’Neill,B.P. (2012) Glioblastoma survival in the United States before and during the temozolomide era. *J. Neurooncol.*, **107**, 359–364.
- Kim,S. *et al.* (2013) Wessim: a whole-exome sequencing simulator based on in silico exome capture. *Bioinformatics*, **29**, 1076–1077.
- Krboth,Z. and Kalman,B. (2020) Longitudinal Characteristics of Glioblastoma in Genome-Wide Studies. *Pathol. Oncol. Res.*, **26**, 2035–2047.
- Lin,K.H. *et al.* (2020) Using antagonistic pleiotropy to design a chemotherapy-induced evolutionary trap to target drug resistance in cancer. *Nat. Genet.*, **52**, 408–417.
- Noorani,I. *et al.* (2020) PiggyBac mutagenesis and exome sequencing identify genetic driver landscapes and potential therapeutic targets of EGFR-mutant gliomas. *Genome Biol.*, **21**, 181.
- Siegmund,K. and Shibata,D. (2016) At least two well-spaced samples are needed to genotype a solid tumor. *BMC Cancer*, **16**, 250.
- Stupp,R. *et al.* (2017) Effect of tumor-treating fields plus maintenance temozolomide vs maintenance temozolomide alone on survival in patients with glioblastoma a randomized clinical trial. *JAMA - J. Am. Med. Assoc.*, **318**, 2306–2316.
- Sun,R. *et al.* (2017) Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.*, **49**, 1015–1024.
- Walther,V. *et al.* (2015) Can oncology recapitulate paleontology? Lessons from species extinctions. *Nat. Rev. Clin. Oncol.*, **12**, 273–285.
- Watkins,T.B.K. and Schwarz,R.F. (2018) Phylogenetic Quantification of Intratumor Heterogeneity. *Cold Spring Harb. Perspect. Med.*, **8**, a028316.
- West,J. *et al.* (2020) Towards multidrug adaptive therapy. *Cancer Res.*, **80**, 1578–1589.
- Williams,M.J. *et al.* (2018) Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.*, **50**, 895–903.
- Yamamoto,Y. *et al.* (2018) Intracellular cholesterol level regulates sensitivity of glioblastoma cells against temozolomide-induced cell death by modulation of caspase-8 activation via death receptor 5-accumulation and activation in the plasma membrane lipid raft. *Biochem. Biophys. Res. Commun.*, **495**, 1292–1299.
- Zhang,J. *et al.* (2017) Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. *Nat. Commun.*, **8**.
- Zhang,J. *et al.* (2019) Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer (mCRPC): Updated analysis of the adaptive abiraterone (abi) study (NCT02415621). *J. Clin. Oncol.*, **37**, 5041–5041.
- Zhao,B. *et al.* (2016) Exploiting Temporal Collateral Sensitivity in Tumor Clonal Evolution. *Cell*, **165**, 234–246.