Leveraging of single molecule sequencing methods for less invasive cancer detection

Richard Hwarn Yip Yim

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Cancer Biology

University College London

June 23, 2021

I, Richard Hwarn Yip Yim, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

In the field of paediatric neuro-oncology, the positions of tumours within the central nervous system of the patients makes the acquisition of solid tumour biopsies risky. For many tumour types, monitoring of treatment response is restricted to Magnetic Resonance Imaging (MRI), or cerebrospinal fluid (CSF) cytology in the cases with leptomeningeal dissemination. Both of these lack sensitivity, leaving room for improvement.

Recent advances in molecular barcoding sensitivity and error suppression have made the sequencing of DNA derived from liquid biopsies possible. Liquid biopsies offer an alternative to solid biopsies, since the collection of bodily fluids is much less invasive by comparison, and liquid biopsies contain cell-free DNA (cfDNA). In cancer patients, it has been shown that a fraction of the cfDNA in multiple liquid biopsies, such as plasma and CSF, harbour the genetic alterations present within the tumour. This circulating tumour DNA (ctDNA) can be used as a biomarker for diagnosis, stratification, and surveillance of the tumour. The monitoring of treatment response, and the detection of minimal residual disease, is of particular importance in paediatric brain tumours, given the low sensitivity of existing methods.

This project created a versatile system, utilising molecular barcoding, which was able to detect Single Nucleotide Variants (SNVs), Insertions/Deletions and Copy-Number Variants in a single assay. A wet-lab workflow was created and iteratively improved, such that it could handle a diverse range of liquid biopsy types, including plasma, cystic fluid and CSF. This workflow was coupled with a bioinformatic pipeline, designed to process the data for all three variant calling processes simultaneously. For SNV calling, a custom variant caller was created to aid in the suppression of errors in barcoded sequencing, and the system was used in the first documented tracking of Adamantinomatous Craniopharyngioma treatment response using cystic fluid liquid biopsies.

Impact Statement

Many cancers are diagnosed by the analysis of tumour biopsies, which involve the use of invasive procedures. The monitoring of treatment progress is more difficult, as patients are often frail, so minimally invasive techniques, such as Magnetic Resonance Imaging (MRI), are used. These less invasive techniques are not ideal for detecting small numbers of cancerous cells following treatment, so relapses can occur due to a lack of sensitivity. These concerns are a particular problem when diagnosing and monitoring paediatric brain tumours, where tissue biopsies are difficult and risky to obtain.

This project used liquid biopsies, such as blood plasma, as material for the detection of tumours, the collection of which is less invasive than the collection of solid biopsies. The detection of tumour DNA in these liquid biopsies was achieved by developing a workflow with improved sensitivity over traditional DNA detection techniques, and a decreased rate of false-positives. The workflow was able to test for multiple tumour types at once, whilst providing more information to the clinician than scans such as MRIs. Additionally, accurate monitoring of the amount of tumour DNA in the bodily fluid over the course of treatment allowed real-time assessment of the success of the treatment.

The work in this project will benefit academia through the publication of data in the form of journal articles. It will increase the global understanding of DNA assays of liquid biopsies, and build upon the work of other labs. The system was also used as a metric for assessing the efficacy of a new treatment for Adamantinomatous Craniopharyngiomas, and the system can benefit the research into future treatments by acting as a method of comparison between them.

The workflow described within this thesis contains improvements over recently published methods in the similar areas. When combined with these other methods, this work will benefit clinicians and cancer patients in the following ways. Using this work, clinicians will be able to diagnose cancers more accurately and reliably than current methods, whilst still being minimally invasive. The technique's ability to diagnose and monitor multiple tumour types at once means that it is also able to reduce the per-sample cost of running the

equipment. Additionally, the current iteration of the workflow is geared towards multiple rare paediatric brain tumours for which bespoke tests would be expensive to develop. This will make it attractive to health economists, and speed adoption in the clinic.

Acknowledgements

I'd like to thank my supervisors Tim Forshew, Stephan Beck and Javier Herrero for the copious amounts of advice and guidance that they have given me throughout my time at the Cancer Institute. To Aylish Selkirk, my sincere thanks goes out to you for all of the support and understanding that you have given me to get me through these years. Thanks to Alice Gutteridge for helping me to find my feet at the start, when I needed it the most. My heartfelt thanks goes out to Anna Köferle who gave me mountains of advice about LaTeX, wet-lab work, and was always there to share a laugh. Simone Ecker: you are, and have been, at the heart of the lab, and your contributions are too numerous to count. Thanks to Andy Feber for not only being a fountain of knowledge, but for sharing his one-of-a-kind sense of humour. Thanks to Miljana Tanic for maintaining that trademark smile whilst giving out tips and pointers, no matter what the time - the rest of us can't do that. Thank you to all of the members of both the Beck lab and the Flanagan lab for making me feel welcome and for sharing a wealth of knowledge, smiles and banter. Thank you to the members of the BLIC for really nerding out and making me feel at home, whilst sharing nuggets of much needed wisdom. A huge thank you goes out to Pawan Dhami, Heli Vaikkinen, Tony Brooks and Alex McLatchie, who provided so much help with the sequencing. Thanks to the members of the Cancer Institute IT team who tolerated my assaults on the network and let me borrow tools when it really mattered.

Thank you to my collaborators: Tom Jacques, Darren Hargrave, John Apps, Mette Jorgenson, Patricia O'Hare, Alex Virasami, Jessica Eze, Tim Meyer, and TuVinh Luong, without whom none of this would be possible.

You have all made it an absolute joy to work here, and I cannot express how deep the well of gratitude I have for you is.

Lastly, special thanks must be given to Ian Kirker for the thesis template upon which this work is written.

Contents

List of Figures

List of Tables

Chapter 1

Introducton

Over the past decade, liquid biopsies have been increasingly important in the diagnosis and monitoring of tumours.[1–6] This project hoped to capitalise on the emerging technologies of molecularly barcoded DNA Next-Generation Sequencing (NGS) to advance this field of research.

The main focus of this project was to optimise wet-lab workflows and to create data analysis pipelines for the processing of molecularly barcoded DNA sequencing data from a variety of liquid biopsies. Features were added to the pipelines over the course of the project, in accordance with the needs of the author's clinical collaborators, to create a toolkit for the analysis of DNA from liquid biopsies. Finally, the system was tested on samples from a variety of Paediatric Brain Tumours (PBTs) to produce proof-of-concept data for the methodology. PBTs were chosen as a test bed for this project for a wide variety of reasons, including the possible increase in sensitivity and molecular information that barcoded NGS could offer over current techniques such as CSF cytology.[7–11]

The following introduction gives an overview of mutational screening, and an in-depth description of what must be considered when sequencing liquid biopsies. Wet-lab methods of error suppression are discussed, as well as effects these methods have on the bioinformatics pipelines designed to analyse data produced by them. Finally, an overview of the tumour types used to generate proof-of-concept data for the project is also included.

1.1 Solid and liquid biopsies for tumour detection and characterisation

1.1.1 Solid tumour biopsies for diagnosis

Historically, histological analysis/Immunohistochemistry (IHC) of solid tumour biopsies has been used for the diagnosis of cancerous lesions in multiple tumour types, including pancreatic neuroendocrine tumours, tumours of the soft tissue and bone, and tumours of the central nervous system.[12–16] In some tumour types, such as Atypical Teratoid/Rhabdoid Tumours (ATRTs), this technique is still used to determine the specific tumour type.[17] A 2016 World Health Organisation paper on Central Nervous System (CNS) tumour classification exemplifies the shift from histology to molecular analysis for the stratification of tumours, particularly in medulloblastomas.[16]

A shift to molecular stratification of patients for clinical trials occurred during the past decade. This change followed the increasing understanding in the field that conventional trials could not investigate rare sub-populations of tumours individually in an efficient manner. For example, B-cell Precursor Acute Lymphoblastic Leukemia (B-ALL) has 35 cases per million zero to fourteen year olds per year, making it the most common childhood cancer, but *MLL1* rearrangements make up only 1% of these tumours.[18–20] Using conventional trial screening methodologies, a novel treatment targeting *MLL1* rearranged B-ALL would need to screen patients for many years before gathering enough statistical power to form reliable conclusions about treatment efficacy. Meanwhile, a trial on Acute Myeloid Leukemia with *MLL1* rearrangements could be doing the same.

The use of new trial protocols in oncology was increased, and these protocols allowed trials to be grouped under the terms: 'basket trial', 'umbrella trial', and 'platform trial', based on their protocol type.[21–25] Basket trials use relevant genetic or epigenetic aberrations, rather than tumour subtype, to admit subjects who receive a treatment. Such a trial setup would be particularly useful in the above example, and in 2019, a trial on multiple *MLL* rearranged/*NPM1* mutated Acute Leukaemias began recruitment.[26] Umbrella trials follow a more general approach, by stratifying the subgroups of a tumour type by their molecular alterations, and using these markers to direct treatment. By this method, umbrella trials can test multiple treatments at once, with the added complexity of comparing multiple populations of patients who were under different treatment regimes.[24] Platform trials, also known as multiarm, multistage designs, are more dynamic than the former two, with

their ability to add and remove arms or sub-studies.[27, 28] Platform studies usually use the current standard of care as a common control group for all other groups, who undergo other treatment regimes.[27, 28] Some arms can be dropped if the treatment is found to be ineffective at interim analysis points, whilst others can be declared superior if enough evidence for this is accrued by their planned end date.[29] The platform trial can be perpetual, with sub-studies being added when they become available.[27, 28]

The number of technologies used to molecularly characterise tumours has exploded. Early Sanger sequencing assays were supplanted by allele-specific quantitative Polymerase Chain Reaction (qPCR) tests, such as the Cobas mutation detection family, and targeted NGS panels.[30–34] Clinicians have begun to use NGS-based panels for diagnostics as part of clinical trials, including the University of Washington's BROCA and ColoSeq panels.[35] The US Food and Drug Administration (FDA) approved their first gene panel for use on solid tumour biopsies in 2017: FoundationOne CDx, followed by numerous other DNA and RNA panels.[32, 36–38] The UK National Health Service (NHS) first published its National Genomic Test Directory in 2018, and now has a comprehensive and extensive list of panels which are available to clinicians, which is searchable by gene.[39]

1.1.1.1 Challenges with solid biopsies

The genetic mutations in a diverse range of cancer types have been shown to be heterogeneous, both within the tumour, and between tumours which are histologically similar.[40–42] The mutations present in a given tumour can radically alter the care pathway given to the patient. This is exemplified by the spectrum of mutations in the Epidermal Growth Factor Receptor (*EGFR*) gene. Some mutations in the kinase domain sensitise cells to the drug Gefitinib whilst others, particularly T790M, are associated with resistance to the drug.[43] The genetic profiling of tumour material in order to direct the form of care given to a patient has become routine for some cancer types.[44, 45]

Solid tumour biopsies present a number of problems when they are used as the primary material for mutational screening. Primary tumours and their metastases can contain differing sets of mutations, as can different groups of cells within the same lesion. This tumour heterogeneity means that sampling a single or small number of sites may miss cells which harbour mutations that can affect the care pathway.[46–48] Core needle biopsies, which take very small and localised samples of tumour material, are particularly problematic. Additionally, the invasiveness of solid biopsy acquisition can be problematic, especially within

the CNS.[49–51]

1.1.2 Liquid biopsies for diagnosis, stratification and monitoring

Liquid biopsies offer a potential solution to the problems of tumour heterogeneity and invasiveness, owing to the presence of nucleic acids within them. cell-free DNA (cfDNA), a collective term for any DNA which is within a bodily fluid and not within a cell, has been found in a diverse range of bodily fluids, including plasma, amniotic fluid, urine and cerebrospinal fluid (CSF).[52–55] Although the processes which lead to the release of cfDNA are not entirely clear, there is evidence for both apoptotic and necrotic origin of these molecules in the blood.[1, 56–59] Within blood plasma, there has been evidence that much of the cfDNA is haematopoietic in origin.[60] Early work in the 1970s compared cfDNA levels to cancer status, and it has been established that tumour cells do release circulating tumour DNA (ctDNA) into nearby fluids.[61–67] Plasma-derived ctDNA tends to be between 140 and 170bp in length, but there is some evidence that the fragmentation characteristics of ctDNA are slightly different from bulk cfDNA in multiple fluids.[59, 68–71]

Circulating Tumour Cells (CTCs) have been found in liquid biopsies such as blood, urine and CSF, from a variety of tumours including lung, breast, and colon cancer.[72–79] These cells play a role in the dissemination of metastases, and are an important prognostic marker for multiple tumour types, though they are commonly detected in non-metastatic breast cancers.[80, 81] CTCs have been used as a biomarker in CSF as well, where leptomeningeal dissemination has occurred.[75, 82–84]

Liquid biopsies have been utilised in both research and clinical environments for the diagnosis, stratification and monitoring of tumours. FDA approval for the CellSearch Circulating Tumor Cell Kit were granted as early as 2004, and was expanded in 2008, highlighting the need for less invasive diagnostic tests at the time.[85, 86] Sensitive qPCR-based tests have also been approved by the US FDA for diagnostic use on plasma, such as the cobas EGFR Mutation Test v2 and the therascreen PIK3CA RGQ PCR Kit.[87–90] BioRad also gained FDA approval for their droplet digital Polymerase Chain Reaction (ddPCR)-based system for use on patients with chronic myeloid leukemia.[91] With the development of molecular barcoding techniques, the use of cfDNA to detect tumours using NGS panels has become tractable.[47] Since then, NGS of liquid biopsies has entered routine clinical practice, with the FDA approval of the FoundationOne Liquid CDx test for multiple non-blood related cancers in 2020.[92] Within research, some strategies to overcome the sensitivity

limits of NGS have involved the use of NGS on either initial solid biopsies or CSF to guide the development of custom ddPCR assays, which were used for monitoring.[93, 94] This hybrid approach is useful when clonal driver mutations are found, but ddPCR for monitoring misses the emergence of emergent, clinically relevant variants, such as those for treatment resistance.[95–97] This highlights the importance of increasing the sensitivity and specificity of NGS technology such that it is able to detect minimal residual disease, whilst capturing the emergence of variants which could be used to direct treatment.

This project focused on DNA contained within these liquid biopsies, but protein-based assays for tumour surveillance have also been trialled, with mixed results.[98–101]

1.1.2.1 The advantages of liquid biopsies

The use of cfDNA as a liquid biopsy conveys a number of advantages over traditional solid biopsy techniques. ctDNA is, in practice, differentiated from the rest of the cfDNA in a sample by the existence of tumour-specific genetic and/or epigenetic alterations, irrespective of the differences in fragmentation patterns.[58, 59, 70] The rapid clearance of cfDNA from the plasma means that cfDNA has a half-life of 16 minutes to 2.5 hours, and samples are representative of the tumour at the time of sampling.[102–105] Its mixed nature makes ctDNA useful for detecting mutations only present in subclones of a solid tumour, and the biopsy process is far less invasive than the multiple-biopsy protocols used to study solid material.[106–108] This ability to detect subclonal mutations, or mutations present at a single site following metastasis, can give clinicians more confidence in a given treatment's effectiveness over data from solid biopsies.

1.1.2.2 The suitability of liquid biopsies for diagnostics and monitoring

ctDNA correlates with the tumour cellularity in high grade tumours, but the ctDNA levels in patients with low grade tumours can be difficult to detect using standard NGS methods.[105, 109] This is partially because many normal tissues (with wild-type DNA) contribute to the pool of cfDNA in a sample, usually making the proportion of ctDNA in the total cfDNA small.[110, 111] This leads to difficulty in discerning real low-frequency mutations on a background of sequencing artefacts and Polymerase Chain Reaction (PCR) noise.[112, 113] An increase in sensitivity over current NGS methods was required in order for this project to succeed. cfDNA samples derived from CSF can potentially be less problematic than that of plasma samples as the majority of ctDNA from PBTs tends to stay within the CSF.[114] This provides a much richer source of ctDNA over plasma at the cost

of increased invasiveness.[115] The same is true for cystic fluid, which is aspirated from tumours with a cystic morphology, due to the close proximity of the fluid to tumour cells and the lack of flow between the cysts and other fluids. On the other hand, higher volumes of blood can be collected from a patient than CSF, potentially providing more cfDNA to assay.

Plasma-derived ctDNA tends to be between 140 and 170bp in length, which makes it amenable to NGS with out the need for shearing or enzymatic digestion.[1, 68] Previous work on the integrity of CSF DNA has shown that the integrity varies widely.[116] This means that longer fragments can potentially exist in the sample, and these needed to be sheared as a first step towards sequencing.

The yield of cfDNA from a liquid biopsy is usually small, particularly in plasma where yields of 1-10ng/ml are common. An efficient sample manipulation protocol was needed to ensure that all relevant DNA was sequenced.

Previous work has asserted that mutant DNA in CSF makes a better biomarker for brain tumours than that of plasma.[55] If, however, plasma ctDNA was sufficiently sensitive by the end of the project, the lower invasiveness would make it the preferable biomarker in the clinic. Both were tested in order to give this project the best chance of success.

Prior to the start of this project, there was little known about the potential concentrations of cfDNA in cystic fluid, or the Variant Allele Frequency (VAF) range which could be expected from a sample. As a result, the investigators did not know of the suitability of cystic fluid for treatment monitoring, making this a pilot study.

1.1.2.3 Liquid biopsy terminology

Much of research on cfDNA and ctDNA has been focused on blood plasma. In the case of plasma, the fluid is actively circulated by the heart, and the process of separating plasma from other blood components involves the removal of cellular fractions following centrifugation.[117] This means that any of the commonly used definitions for cfDNA/ccfDNA: cell-free DNA or circulating, cell-free DNA, and the definition for ctDNA: circulating tumour DNA, all hold true for plasma.[3, 52, 109, 110, 118, 119] CSF is not actively circulated, but its flow, which depends on both secretion and drainage, has been studied.[120–122] Depending on whether one's definition of whether a fluid circulates requires active circulation or if passive flow is allowed, the status of CSF as a circulating fluid is debatable. Despite this, the term ctDNA, has been used when discussing CSF in journals

such as Nature and Blood.[123–126] Since the field has settled upon this term, this project continued to refer to the variant-harbouring DNA derived from CSF as ctDNA.

As a further note of relevance to this project: DNA was extracted from the unspun fluid of closed cysts for some samples, and therefore this DNA was neither circulating, nor entirely cell-free. All mentions of this DNA were of the form 'cystic fluid DNA' to differentiate it from any other terminology.

1.2 Molecular barcoding in Next-Generation Sequencing

Molecular barcoding has been used to reduce the noise caused by PCR errors and sequencing errors, by providing ways to ascertain the original template molecule from which a given set of reads arises. A visual representation of how this can be achieved is presented in Figure 1.1. The general principle is to attach a unique string of nucleotides (a barcode) to one or both ends of an individual template molecule at the start of library preparation. After many molecules are PCR amplified and sequenced, it is possible to collapse the sequencing reads made from each molecule into separate families. It must be noted that the term 'molecular barcode' is used in this document to describe a random string of nucleotides which differ between molecules in a single library, whilst an 'index' describes the nucleotide strings used to differentiate between libraries in a sequencing run.

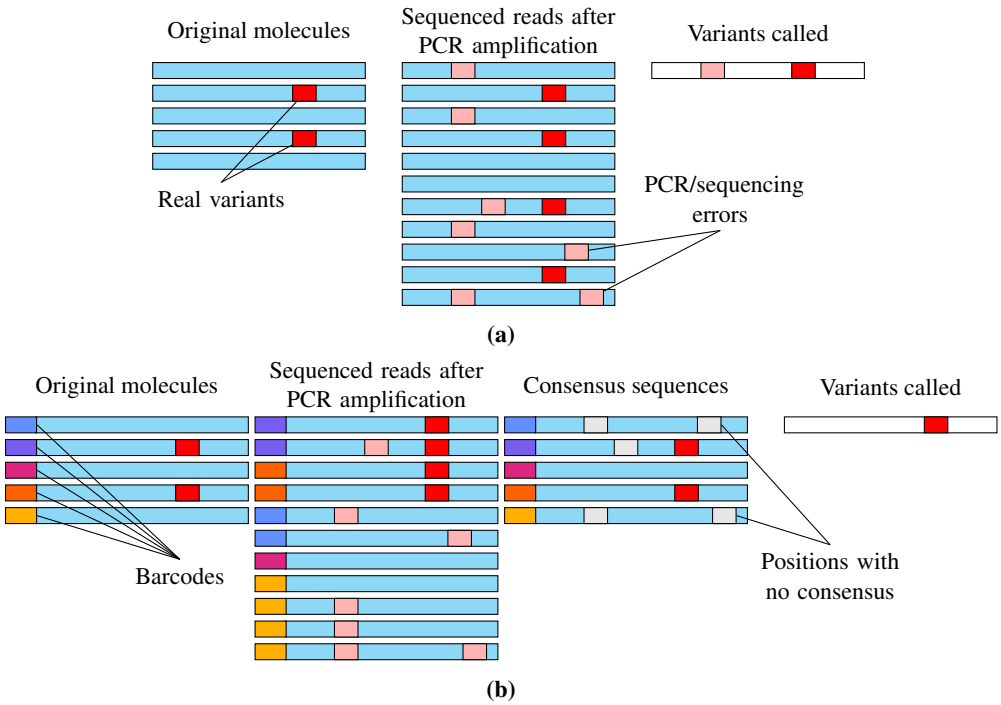


Figure 1.1: How molecular barcoding affects the number of variants called in a sequencing dataset.

1.1a) Without molecular barcoding, PCR and sequencing errors accumulate, and it is difficult to set variant calling thresholds which filter these errors out whilst remaining sensitive to true mutants. **1.1b)** With barcodes attached to each template molecule, it is possible to collapse PCR duplicates into families which represent the original molecules. With stringent parameters, such as limiting consensus formation to positions where all reads in a family agree, fewer false-positive variant calls are made.

1.2.1 Molecular barcoding and other PCR deduplication technologies

Prior to the start of this project, A number of competing methods which sought to reduce the PCR error in Illumina sequencing had been developed. This section provides an overview of some of the technologies available at the time, and their advantages and disadvantages.

TAm-seq: In one of the first published methods of PCR error suppression in NGS, Forshew *et al.* used replicate amplification to reduce noise.[68] Both tumour DNA and control DNA were separately amplified in replicate, with barcodes identifying each amplification reaction. Pooled amplification products were sequenced, the control data was used to model the distribution of errors at each position on the amplicon panel, and tumour variants which were above the noise floor were considered real. This technique was called TAm-seq, and in its original form, it was able to achieve a per base error rate of between <0.1% and <0.6%. TAm-seq was limited in its error-rate reduction since its use of barcodes was to identify each amplification run, rather than each original template molecule. PCR errors are still allowed to accumulate, and whilst comparing the noise to control datasets can reduce the false-positive rate, true molecular barcodes have been demonstrably more effective in error suppression.

Safe-SeqS described two methods for using unique identifiers, or UIDs, in sequencing error suppression.[127, 128] The endogenous method involved the shearing of DNA, followed by the ligation of sequencing adaptors and an inverse PCR step, and sought to use the sequencing start sites in paired-end sequencing as endogenous unique identifiers. This was able to reduce sequencing errors by 70-fold.[127] The utility of this method for cfDNA was hampered by the rate of loss of DNA between input and sequencing which resulted in a high rate of duplication. This method was not practical for cfDNA analysis without improvements, because the amount of starting DNA is a limiting factor. The principle of using the barcode alongside the location of the sequencing start site has been used since, in a number of different workflows including ThruPLEX Tag-seq and the original Duplex Sequencing pipeline.[11, 129–131]

The exogenous method involved the inclusion of PCR primers designed against a target region, one of which contained a barcode. Barcode assignment was performed during the first two PCR amplification cycles, and sequencing adaptors were added for further amplification cycles. This method boasted an ability to convert ~78% of input fragments into sequenced and collapsed read families.[127] One limitation of this, as with all amplicon-

based strategies, was to have defined start and end sites where primers bind. This affected the versatility of the panel in detecting structural variant breakpoints, as a designed amplicon would need to bridge the breakpoint to detect it. With every additional pair of primers in such a panel, the possibility of amplifying an off-target region of the genome increases, and every panel configuration must be evaluated before use.

Circle Sequencing provided a way of limiting the error-affects of PCR pre-amplification or first-round PCR errors. Pre-amplification using rolling circle amplification was followed by PCR the products, which contained the original template sequence repeated approximately three times in tandem.[132] The linear amplification, as opposed to PCR, meant that each new tandem copy of the template molecule was made directly from the template, and errors did not propagate to all products from the same molecule. Subsequent PCR errors and errors during the rolling circle amplification could be filtered out by producing a consensus from the tandem repeats. Since each template molecule was read three times in tandem, the template was required to be less than $\frac{1}{3}$ of the total read length of the sequencing. cfDNA from plasma is 140-170bp long, so the maximum read length on Illumina platforms of 2×250 was needed to sequence these molecules. As the read length of Illumina sequencing increases, the read quality decreases, so concatenated molecules in a single read would hamper noise reduction in this system.

Molecular Inversion Probes (MIPs) provided a simple means of performing library preparation and barcoding, and had a reported error rate of 2.6×10^{-5} per base.[133] A MIP was a single strand of DNA with a 16-24nt genomic target-specific oligo at each end, and harboured a barcode sequence. Both oligos bound the same target strand on the genome as the first step of targeting. The MIP was amplified along with the target sequence in between the oligos to form a circle, then linearised to convert them into sequenceable libraries. When seeking to apply this technology to cfDNA, one must remember that cfDNA fragments are often small. If a piece of cfDNA contains a binding site for one targeting oligo but not the other on a MIP, library preparation cannot proceed. The longer the genomic distance between the arms, the more likely a target cfDNA fragment from a targeted region will only bind a single arm of the MIP. The number of overlapping MIPs needed to adequately target a region becomes high, making MIPs usable but problematic for cfDNA.

Duplex Sequencing was the most promising form of barcoding at the start of the project.[10, 134] The technique had been shown to have the highest potential sensitivity

with an observed error rate of $1 - 4 \times 10^{-7}$ per base and a theoretical minimum error rate of $< 4 \times 10^{-10}$ per base.[134] This level of sensitivity was achieved by ligating a 12nt double stranded barcode onto either end of a double stranded template molecule. After PCR and sequencing, the PCR duplicates originating from one strand of the template had the reverse complement barcodes relative to duplicates originating from the other strand. The consensus sequence from one strand could be matched to its reverse complement based on this, and only variants which were present in both families were counted in a final consensus sequence. This dramatically reduced the error rate. The technique utilised ligation-based library preparation, which meant that it was compatible with cfDNA fragment sizes. Targeting was done by hybrid capture using the Agilent XT or similar system. One potential drawback of Duplex Sequencing in its original form was its overall efficiency in converting template molecules into Duplex Consensus Sequences, as one paper used 750ng of DNA to produce 2.2×10^7 and 9.7×10^7 bases of Duplex Consensus in two experiments.[135] This is not a problem for the sequencing of tumour material where the amount of starting DNA is often adequate, but with the low amounts of cfDNA per millilitre of plasma, this can lead to low sensitivity.[136]

The **ThruPLEX Tag-seq** kit was an off-the-shelf NGS library preparation kit which made use of molecular barcoding. It had an advertised ability to detect variants at a 0.5% VAF, and was based on library preparation technology which was efficient at low inputs.[129, 137] Since this kit was used within the project, the library preparation procedure is presented in detail in 2.2.6.2. Briefly, hairpin adaptors containing barcodes are ligated to either end of a template molecule. The hairpins are broken using a proprietary method, and the resulting single-stranded regions are used to add sequencing adaptors to the libraries. One of the first aims of this project was to determine whether the kit was suitable for the reliable detection of variants in different liquid biopsies.

1.3 Paediatric brain tumours

Cancers of the CNS and intracranial space are the second most common cancer type in children in the UK, with an average of 412 cases per year in children between 0 and 14 years.[138–140] They also lead to the highest rate of mortality of any cancer type in this age group.[141] PBTs have been differentiated into more than one hundred types based on histology, but the majority can be clustered into two broad groups based on origin: Gliomas which arise from glial cells, and Embryonal tumours which putatively arise from embryonic cells which remain after birth (Figure 1.2).[16, 142] Since 2016, and the release of the World Health Organization Classification of Tumors of the Central Nervous System, these tumours have been defined more granularly by molecular aberrations, including gene inactivations, chromosome arm deletions and methylation states.[16, 143] Gliomas are the more common of the groups.[144] Notable tumour types include Pilocytic Astrocytoma (PA) (the most common, making up 17% of all PBTs), and Diffuse Intrinsic Pontine Glioma (DIPG), which have a >90% mortality rate within 2 years).[142] Embryonal tumours are comprised of medulloblastomas, ATRTs, and other primative malignant tumours which were, until 2016, grouped undder the term: Primitive Neuroectodermal Tumours.[16, 145, 146] Johnson *et al.* (2014) provide an overview of the epidemiology of all of these types of tumour.[142]

This project focused on ATRT, DIPG, and Adamantinomatous Craniopharyngioma (ACP), with small numbers of samples from *WNT*-Activated Medulloblastoma (WAM) and PA patients. Information on the common mutations for different types of PBT is located in Section 2.2.3, which describes a targeted capture panel designed against these mutations.

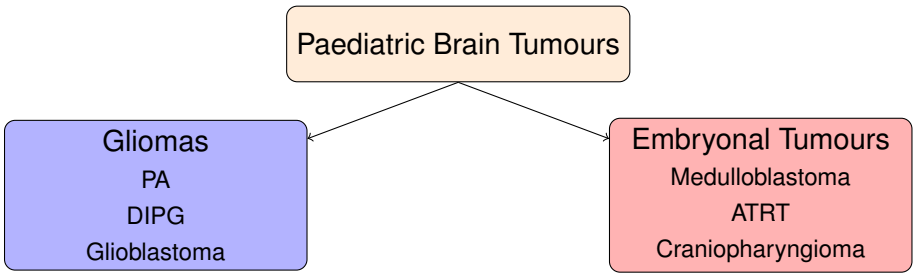


Figure 1.2: The main classes of PBT.

A schematic showing the two main classes of PBT, and the class each tumour type in this project belongs to.

1.3.1 Atypical Teratoid Rhabdoid Tumours

ATRTs are a rare tumour type, occurring in under 18's at a rate of 0.07–0.14 per 100,000, and rarely in adults.[17, 142, 147] ATRTs are associated with poor outcomes, with a median survival time of 11-38 months after diagnosis.[148–150] This is due to rapid growth and large size at diagnosis.[17, 151–155] Monitoring of treatment, including the detection of minimal residual disease, is of great importance with such an aggressive tumour.

ATRTs are genetically very simple. The vast majority of ATRTs have biallelic inactivation of the *SMARCB1* gene, via a combination of Copy-Number Variations (CNVs), SNVs, Insertions/Deletionss (InDels), or Loss of Heterozygosity (LOH) events.[16, 156, 157] When one copy of *SMARCB1* is inactivated in the germline, the result is Rhabdoid Tumor Predisposition Syndrome (RTPS).[158, 159]

Malignant Rhabdoid Tumours (MRTs) are tumours with rhabdoid morphology, which occur outside of the CNS, commonly in the kidneys. These tumours share a mutational profile with ATRTs, and can co-occur with ATRTs in a synchronous or metachronous manner, particularly on the background of RTPS.[160–164]

1.3.1.1 Current diagnosis and monitoring

Patients present with wide ranging combinations of symptoms, including intracranial hypertension, loss of vision and lethargy.[17, 142, 147] A Magnetic Resonance Imaging (MRI) scan of the head or whole neuraxis is obtained from them. An initial diagnosis of the tumour is made using these images, though the diagnosis is speculative due to several PBTs having similar morphologies on such images.[151, 165, 166] Patients with intracranial hypertension usually undergo a gross resection to remove the tumour, and histology or IHC is used to form a firm diagnosis.[17, 151]. Specifically, since IHC was demonstrated as a sensitive test for the loss of *SMARCB1* expression in rhabdoid tumours in 2004, the technique has been used to separate ATRTs and MRTs from other histologically similar tumours.[162, 167, 168] The direct access to tumour material following resection makes diagnosis straightforward. Monitoring and detection of residual disease following treatment, however, is more problematic.

Currently, monitoring is not done using a defined methodology or schedule, but cytology or IHC is performed on CSF samples where possible, in combination with MRI.[7–9] Cytological markers used include the presence of large rhabdoid cells and primitive cells under routine Haematoxylin and Eosin staining, and the loss of nuclear *SMARCB1* stain-

ing under IHC.[7] Both of these techniques rely on the presence of tumour cells within the CSF, so research into assays which do not require so called leptomeningeal dissemination has been undertaken. A protein-based trial on the detection of osteopontin levels in the CSF showed that such a test had limited sensitivity.[100] T2 weighted MRI, the least invasive technique discussed, is routinely used for monitoring, but its resolution precludes its use for the detection of minimal residual disease.[166]

1.3.1.2 Potential for progress

ATRTs are well suited to ctDNA analysis, as almost every tumour harbours genetic alterations which result in the inactivation of a single gene: *SMARCB1*, with very small numbers of tumours being driven by *SMARCA4* inactivation.[16, 156, 157, 169] Molecular characterisation of CSF cfDNA has multiple advantages over the current histological/cytological paradigm, and the current research into protein assays. cfDNA analysis is able to assay the total DNA, whether cellular or cell-free, removing the requirement for the presence of cells in the CSF. This also potentially makes cfDNA analysis more sensitive than either of the two methods above, by assessing a larger pool of targets for biomarkers indicative of a neoplasm.

In the case of RTPS, since ATRT follows Knudsen's two-hit hypothesis, two aberrations in the *SMARCB1* gene can be tracked separately, with the acquired variant being a marker for residual disease.

1.3.2 Adamantinomatous Craniopharyngioma

ACP is a benign, slow-growing tumour type which makes up 1.2%-4.6% of all intracranial tumors, and between 5% and 10% of paediatric intracranial tumours.[170–175] Histologically, there are two main types: Squamous-Papillary Craniopharyngioma, and ACP. ACP is the primary paediatric form of Craniopharyngioma, and is characterised by *CTNNB1* exon 3 mutations which prevent degradation of beta-catenin, resulting in activation of the canonical *WNT* signalling pathway. There are no other known recurrent genetic alterations for this tumour type.[176, 177]

Under MRI, ACP is either a solid or mixed cystic-solid epithelial tumour in the intra/suprasellar regions, which contain hypothalamus and pituitary gland.[170, 173, 176, 178] ACP is often characterised by calcified deposits within the tumour, with the rare occurrence of teeth.[170, 179]

1.3.2.1 Current diagnosis and monitoring

The mass effect of the tumour and its cysts, combined with the location close to the optic chiasma, mean that patients tend to present with visual field defects, deficiencies of anterior pituitary hormones, and symptoms of raised intracranial pressure.[170, 178] A Computerized Tomography X-ray scan (CT) or MRI reveals the location and appearance of cysts, allowing for diagnosis of the tumour.[173]

Surgery is the mainstay of management, with or without radiotherapy, but the location of tumours close to the pituitary gland and other critical structures means that debulking is preferred over full resection.[170, 173, 180, 181] The above have contributed to the high five year survival rate of 91%-98%, but a low long-term quality of life for survivors. Sequelae due to the disease or treatment include obesity, other hormonal deficiencies, and a tendency to relapse.[170, 180] The management of tumour cysts has been a particular challenge. As a result, there has been a search for novel treatments which reduce the effects on surrounding structures. Intracystic radioisotopes, and chemotherapeutic agents such as bleomycin, have been used. Concerns regarding toxicity have, however, limited their uptake. More recently, intracystic injection of Interferon-α (IFN-α), which has the potential for fewer adverse side effects, has been trialled.[182, 183]

1.3.2.2 Immunosuppression and IFN-α

The use of IFN-α was based on its success in treatment of patients with squamous cell carcinoma of head and neck.[184, 185] Profiling of the immune environment of ACP cysts has shown a complex inflammatory profile and IFN-α may modulate this.[176, 186, 187]

Previously, IFN-α treatments have been used with some success in small studies on ACP, but these focused on secondary effects of the treatment, such as tumour size, and the magnitude of effects on hypothalamic, pituitary, and optic function.[176, 183, 188] There has, however, been little work to date on understanding the changes in cyst biology induced by IFN-α. More work is needed to elucidate the mechanism of IFN-α treatment, and to ascertain whether effects on the immune environment lead to increased cell death.

Despite a lack of a understanding of how IFN-α works on ACP at a cellular or a signalling level, multiple small studies have been conducted on this therapy, and they provide preliminary evidence of its efficacy.[182, 183, 189]

1.3.3 Diffuse Intrinsic Pontine Glioma

DIPG is the most common paediatric brain stem tumour.[142, 190–192] The location of DIPGs within the brainstem and their propensity for infiltrative growth make them difficult to resect.[193]

Over 90% of patients die within two years of diagnosis, and this tumour type is the main cause of PBT-related death in children.[142, 194] Survival rates have also remained static for over two decades.[193–195]

The mutational profile of DIPG has been well established, and there are ten to fifteen genes which commonly display genetic alterations. These genes include *TP53*, *H3F3A*, *HIST1H3B*, AVCR1, *ATRX*, and *PDGFRA*.[194, 196–198] There has been research into some of these, including *AVCR1*, *PDGFRA* and *ATRX*, as targets for therapies, but much of the research into targeted therapies is in its early stages.[199–201]

1.3.3.1 Methods of diagnosis

DIPG is currently diagnosed using a combination of MRI of the head and neck, and symptoms at presentation.[192, 194, 195, 202] Historically, biopsies of tumour material have rarely been taken due to the tumour's location near to crucial structures.[192, 195, 202] More recent studies point towards biopsies being justifiable, which opens the door to molecular profiling of tumour material, but this remains an invasive option for the gathering of information.[203, 204] Early studies on using genetic and proteomic biomarkers CSF as a means of molecular diagnosis have been attempted.[205, 206]

There is currently no gold standard method for the monitoring of DIPG other than radiological and clinical observation, due to the tumour location.[207, 208] Studies on using CSF or plasma have been conducted, utilising ddPCR.[209, 210]

1.3.3.2 Potential for progress

CSF biopsies using lumbar puncture are invasive, but they are far less invasive than the brain tissue biopsies which would otherwise be necessary for mutational profiling. There has been work on ddPCR-based assays on CSF, but a single high sensitivity assay, which is able to give a fuller picture of the molecular profile of a given tumour, would be preferable.[210, 211]

The nature of targeted Next-Generation DNA sequencing means that it can assay as many regions of the genome as are desired.[212] Consequently, a single test can provide rich data about the tumour's genetic alterations, allowing clinicians to select personalised,

targeted treatments based on those alterations.

1.3.4 The use of PBTs as a case study for technology development

Despite the general applicability of the proposed system to cancers, the project focused mainly on PBTs as a potential use case, and to highlight versatility. PBTs are individually rare, so the inclusion of regions specific to a single PBT into a general purpose NGS panel would increase sequencing costs of every sample, for small increases in utility of the panel in stratification or monitoring. Additionally, whilst all gathering of solid tumour material is invasive, biopsies of solid material from PBTs represent some of the most extreme cases of this. This has, in the case of DIPG, led to some debate about the possibility of acquiring solid biopsy material.[206, 209] Whilst individual ddPCR assays have been developed for liquid biopsy-based tumour surveillance, in clinical practice, each new assay would need to be validated, for use on a relatively small number of patients. A versatile system, which would need to be validated only once for each panel developed upon it, could provide a PBT speciality clinic with the ability to use a single test for diagnosis and monitoring, whilst minimising the size of the panel and thus sequencing costs. This project sought to focus on PBTs to demonstrate the ability of the system's ability to fulfil this role.

1.4 Overall objectives of the project and hypotheses

The overall aim of this project was to apply new technologies, centred around molecular barcoding and NGS, to the emerging field of liquid biopsies. This involved developing and assessing novel techniques, and building on existing technologies, both to process samples for sequencing, and to analyse the subsequent data. To achieve this overall aim, the following two hypotheses for the project were created:

1. It was possible to create a versatile workflow, which was capable of taking multiple liquid biopsy types and processing them into barcoded, targeted libraries, suitable for Illumina sequencing.

2. It was possible to apply molecular barcoding to liquid biopsies, to improve on cytology: the current gold standard of monitoring in PBTs with leptomeningeal involvement.

The initial objective was to assemble a wet-lab workflow which was able to take samples from a variety of liquid biopsies, and to process them into sequenceable, molecularly

barcoded libraries. This objective was the main focus of Chapter 2. Running concurrently to the work in Chapter 2, the project developed assessed the suitability of existing molecular barcoding software for the data which was being generated by the wet-lab workflow. Once existing software was found to be inadequate or inappropriate, in house data analysis pipelines were created and improved in tandem with the workflow. This iterative approach spanned the work in Chapters 2 and 3. The final objective was to test the pipelines and workflows produced during the project on a range of clinical samples, as a proof-of-concept for the monitoring of treatment in PBT.

Chapter 2

Development of a sample type-agnostic wet-lab workflow for barcoded sequencing

2.1 Introduction

The first step in applying molecular barcoding to cell-free DNA (cfDNA) was to develop procedures for isolating cfDNA from liquid biopsies, processing the fragments into sequencing libraries, and performing appropriate quality checks on the libraries prior to deep sequencing. This workflow development was initially tested on control material, as part of a preliminary study. A set of cfDNA reference standards, with defined Variant Allele Frequencys (VAFs) at specific sites, produced by Horizon Discovery, was used as a base for the preliminary testing. Rubicon Genomics shared a beta-version of their ThruPLEX Tag-seq kit with the author, and this kit, followed by its production counterparts, were used throughout the course of this project. Further testing of the wet-lab workflow was performed on clinical samples, and occurred concurrently with the development of a bioinformatic pipeline, discussed in Chapters 3 and 4.

The overall project was divided into three main studies: a preliminary study on control material, the HiSeq Capture 1 (HC1) study, a small scale pilot study of plasma, cerebrospinal fluid (CSF), and cystic fluid samples, and the HiSeq Capture 2 (HC2) study, a larger pilot study with improved data analysis. The wet-lab workflows for each study had iterative improvements over previous studies in the project. For ease of understanding, the versions of the workflow are referred to as the "preliminary" version, Version 1, and Version 2, corresponding to their project names.

2.1.1 Aims and Objectives

The first objective for this project was to develop methods for isolating DNA from different liquid biopsies, which were compatible with barcoded library preparation kits. In parallel, the author aimed to create a hybrid capture panel of genes and regions which is targeted towards multiple Paediatric Brain Tumours (PBTs). The third aim was to implement accurate library quantification, QC procedures, and normalisation procedures, to optimise the sequencing depth of each sample. Finally, the author aimed to chain all of these novel methods into a pipeline which was capable of taking multiple sample types through from nucleic acid extraction to Next-Generation Sequencing (NGS), with appropriate quantification and pooling steps.

2.2 Materials and methods

There were three main versions of the wet-lab workflow, each created for a corresponding study. All three studies used the same basic workflow, depicted in Figure 2.1, but each study contained iterative improvements to individual elements of the workflow. This methodology section focuses on the novel techniques developed for the project, then gives a detailed description of how the overall workflow improved between the three studies.

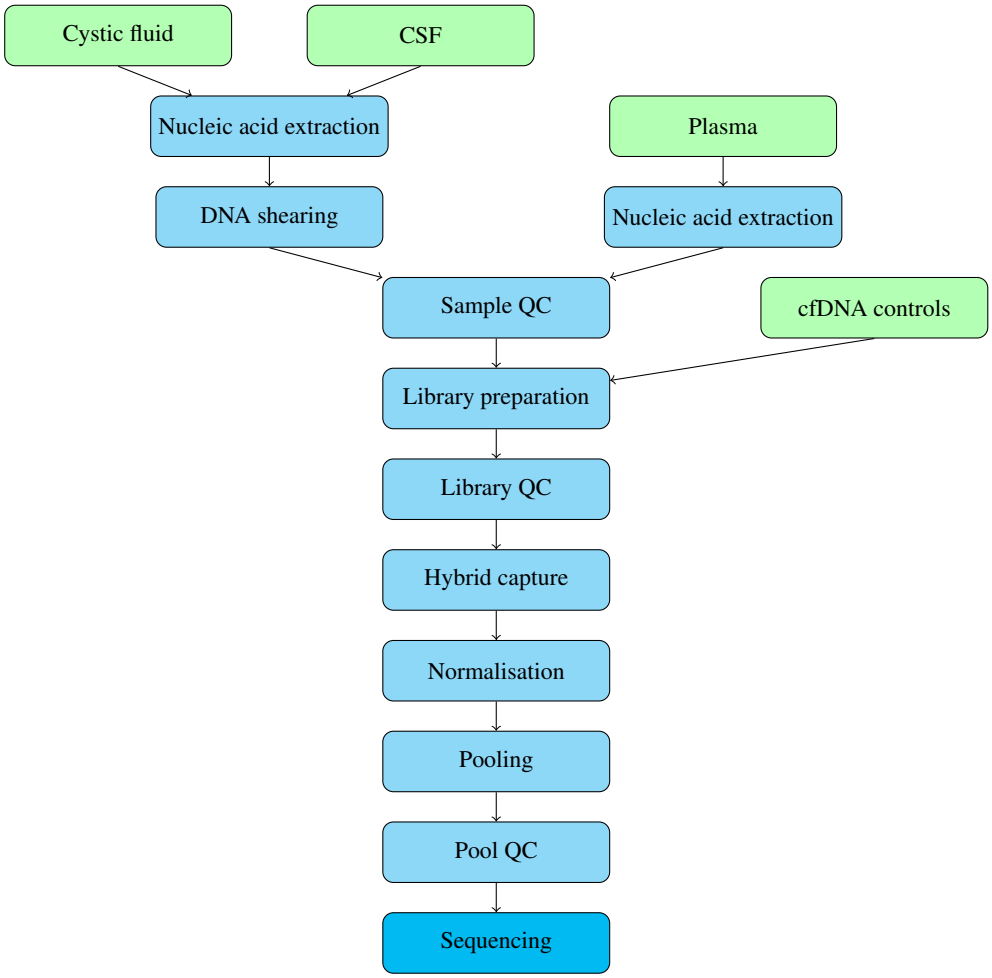


Figure 2.1: The overall wet-lab workflow, used throughout the project.

Green boxes depict different sample types used in the three studies of the project, and blue boxes are the processes to which the samples were subjected.

2.2.1 Pre-study sample handling for the HC1 and HC2 cohorts

During the course of the overall project, one synthetic DNA cohort and three main clinical sample cohorts were used. Details of all of the cohorts in the project are presented in Section 2.3.1. The preliminary study's cohort was a set of purified and sheared control material, so no manipulation of these samples was performed before the study. The HC1 study's HC1 cohort was comprised of plasma samples, CSF samples, and cystic fluid samples, all of which were collected as part of a retrospective study design. Following the success of the HC1 study with CSF and cystic fluid samples, the HC2 and HC2C cohorts were assembled in a similar manner to the HC1 cohort, with a focus on Atypical Teratoid/Rhabdoid Tumour (ATRT) in the HC2 cohort, and a focus on Adamantinomatous Craniopharyngioma (ACP) in the HC2C cohort. The HC2 study's cohorts were subdivided as the HC2C cohort was used to investigate a separate hypothesis to the HC2 cohort, as is described in Chapter 4. The methods described in this section were performed by staff at Great Ormond Street Hospital (GOSH) and the GOSH/UCL Institute of Child Health (ICH).

CSF samples from the GOSH archive were sent for diagnostic cytopathology as part of standard care at the hospital. CSF samples of over 3ml were centrifuged, and any supernatant over 3ml was frozen at -80°C. Whole blood was centrifuged to separate the plasma from other blood components, and the plasma was transferred to a new cryotube. These samples were also frozen at -80°C. Cystic fluid samples from both the HC1 and HC2 cohorts were aspirated from within the cysts of ACP patients using an intracystic catheter. These samples were frozen at -80°C without centrifugation.

All ACP samples were transferred to the author from the Brain UK virtual biobank, via Dr. John Apps, who holds ethical approval for the studies. All other samples were transferred to the author under the ethical approval of the Children's Cancer and Leukaemia Group (CCLG) tumour bank (2014 BS 11).

2.2.2 Development of DNA isolation and sample handling methods for optimal library preparation input

At the start of the project, the aim was to create a workflow which took raw liquid biopsy samples, processed them into DNA samples which were suitable for library preparation with the Tag-seq kit, and took the libraries through the sequencing process. This early development stage necessitated the use of control material, rather than precious patient samples, so no DNA isolation or sample preprocessing was performed for version 1 of the workflow. Once the basic framework for the handling of DNA samples from library preparation to sequencing had been tested, isolation and preprocessing methods were developed for use on the HC1 cohort of CSF, cystic fluid, and plasma samples. Upon analysis of the results from this cohort and Version 2 of the workflow, the sample handling workflow was optimised to minimise the sequencing depth per DNA molecule, whilst maintaining quality of the data produced. This upgraded method was tested, as part of the Version 3 wet-lab workflow, on the larger HC2 and HC2C cohorts of CSF and cystic fluid samples.

2.2.2.1 Development of DNA isolation and shearing protocols for patient CSF and cystic fluid samples

Cell-free CSF DNA can be as short as that of plasma DNA, which can be sequenced without shearing.[68, 71] Depending on the collection methodology, which was not controlled by clinical collaborators for these retrospective studies, circulating tumour cells within the CSF could leave near full length DNA within a liquid biopsy sample, and DNA is sequenced efficiently by Illumina sequencing when sheared.[213] The QIAamp Circulating Nucleic Acid kit was chosen for DNA extraction, and a representative of QIAGEN informed the author that DNA was sheared to approximately 10kb as it passed through this kit's columns. For maximum sensitivity, a method for shearing long DNA down to below 500bp, whilst maintaining the integrity of DNA which started at below 200bp in length, was created.

To test how a shearing methodology affected short DNA, a 182bp fragment of soybean DNA, referred to as sDNA, was used in tests of shearing conditions. The sequence of this DNA fragment, designed by Dr. M. Tanic (unpublished), based on the work by Pallisgaard *et al.*, is available in Table 2.2a.[214] The fragment was PCR amplified according to the parameters in Tables 2.2b and 2.2c, to create enough material for the shearing tests. Five 25µl amplification reactions were run, and the products were purified by performing a bead cleanup with 37.5µl of Agencourt AMPure XP beads. The purified products from all reactions were pooled, quantified using a Qubit dsDNA HS Assay kit, and diluted to 0.25ng/µl. The effects of shearing conditions on the integrity of samples of sDNA could now be tested.

Item	Sequence (5'-3')
Template	CATGGTCCACTTCCTCAGGTAAACCATAGGTTCTTGCTGTCTATTTGTATAATG GTATTGTAGGGCAGTCAGTATTTAATGTTATGATCACATCACTAGATCAGCGTG ACTTAGATGTTTCTCATTCCTTATTTGAATCTATAAACTTTTAATCTTCAGCTT GTGGAAAATTATTGATGGGA
Forward primer	CATGGTCCACTTCCTCAGGT
Reverse primer	TCCCATCAATAATTTTCCACAA

(a)

Reagent	Volume (µl)
Water	18.2
5X Herculase buffer	5
Herculase polymerase	0.2
100mM dNTPs	0.2
10µM Forward primer	0.6
10µM Reverse primer	0.6
Template	0.2

(b)

Temperature (°C)	Time (s)	Cycles
95	300	1
95	30	
50	30	25
72	60	
4	Forever	1

(c)

Figure 2.2: Details of the manufacture of the sDNA 182bp control material.

2.2a) Sequences of the PCR template and primers, designed by Dr. M. Tanic. **2.2b)** PCR reaction setup using a Herculase II Fusion DNA Polymerase Kit. **2.2c)** Cycling conditions for sDNA amplification.

The optimum shearing conditions, which were able to preserve sDNA integrity, were required to shear the ~10kb DNA eluate from QIAGEN columns down to a length suitable for sequencing. To test the ability of sets of shearing conditions to do so, longer control material, which simulated the product of running genomic DNA through a column from a Qiagen QIAamp Circulating Nucleic Acid Kit, was created. To make this "lDNA", Bioline Human Genomic DNA was mixed with extraction buffers from the Circulating Nucleic Acid Kit according to Table 2.1, and bound to a column mounted on a vacuum manifold. Once bound, the DNA was washed with 600µl of Buffer ACW1, 750µl of Buffer ACW2, and 750 µl of absolute ethanol, according to the extraction kit handbook.[215] The lDNA was finally eluted in 200µl of Invitrogen TE buffer. The eluate was quantified using a Qubit dsDNA HS Assay kit, and diluted to 0.5ng/µl with nuclease-free water. Shearing conditions which preserved the integrity of sDNA could now be tested for their effectiveness in bringing lDNA down to sequenceable lengths.

Table 2.1: The mixture of control genomic DNA and QIAGEN buffers used to create the lDNA simulated nucleic acid extraction product.

Reagent	Volume (µl)
Nuclease-free water	1065
Bioline Human Genomic DNA	10
Buffer ACL	800
Buffer ACB	1800

A total of thirty nine sets of shearing conditions for the Covaris E220 Evolution sonicator were tested during development, and they are presented in Appendix C. Initially, testing was performed using MicroTUBE-15 tubes, which had a 15µl capacity, since the ThruPLEX Tag-seq library preparation kit was only able to accept 10µl of input. It quickly became apparent that the amount of DNA in 10µl of QIAamp column eluate was not sufficient to provide the required sequencing depth for rare variant detection, so further tests were performed in MicroTUBE-50 tubes, with a maximum 55µl capacity. The larger MicroTUBEs allowed for the entire 50µl eluate from a QIAamp column to be sheared at once, providing more material for library preparation. The extra material could either be concentrated, and loaded into a single reaction of the library preparation kit, or the sheared eluate could be loaded into 5 reactions, if a suitable concentration methodology could not be found.

Initial condition sets were taken from a Covaris guide for shearing using E220 series sonicators. As testing continued, the incident power, duty factor, and cycles per burst settings was gradually lowered, to reduce the damage to sDNA. The treatment time was gradually increased to allow the gentler conditions to successfully shear the lDNA down to sequenceable lengths. This culminated in the conditions described in Table 2.2.

Table 2.2: Finalised Covaris E220 Evolution shearing parameters for CSF and cystic Fluid DNA

Parameter	Value
Tube type	microTUBE-50 AFA Fiber Screw-Cap
Incident power	15W
Duty factor	15.00%
Cycles per burst	200
Treatment time	950s
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	yes

2.2.3 FLCP-1 - A capture panel for targeted sequencing

The motivation for the creation of Forshew Lab Capture Panel 1 (FLCP-1) was to create a general purpose capture panel for genomic regions which are commonly altered in PBTs. The panel was under development as the HC1 cohort was being selected, so it was made to be able to capture both regions unique to tumour types and those which were more common to PBTs. The Agilent XT platform was selected in part because extra regions could be added to the panel, should more regions become of interest to the project. In its initial form, FLCP-1 was part of a proof of concept for the overall wet-lab workflow, upon which a clinical test could be built. At the beginning of the project, the possibility of making this panel capable of handling Hepatocellular Carcinoma (HCC) samples was explored, so regions which are commonly altered in HCCs, such as *TERT* were also included in the panel.

FLCP-1 was designed through literature review, analysis of the COSMIC database, and with clinical expert guidance from Thomas Jacques, professor of paediatric neuropathology at GOSH.[216] Candidate regions to target were selected in order to balance the likelihood of mutation detection in PBT and HCC samples against the size of the total genomic region

covered by the panel. Given the rarity of PBTs, mutations common to many types of paediatric cranial tumour were preferred in order to maximise chances of a given sample yielding detectable mutations.

Individual regions within a candidate gene were selected for targeting in the panel, based on the distribution and type of genetic alteration reported in the literature. If exonic Single Nucleotide Variants (SNVs) or short exonic Insertions/Deletionss (InDels) were reported, the gene was viewed on the COSMIC database, and regions where alterations were frequently reported were targeted.[217, 218] If mutations were widely distributed, all coding exons were targeted. In doing so, the Agilent SureDesign software was configured to take all exons marked as coding from all splice variants listed in the Ensembl and RefSeq databases.[219, 220] Where non-coding mutations, fusions or genomic rearrangements were reported, PBT- and HCC-related mutations were found in the literature. Relevant introns or promoter regions were added to the list of targeted regions. Table 2.3 outlines the regions that were targeted in FLCP-1.

PBTs are not only histologically diverse, but are also mutationally diverse. DIPG commonly has mutations in *H3F3A*, *HISTH1H3B*, *PIK3CA*, *TP53*, *PTEN*, *ACVR1* and *PPMID*.[197] The Chromosome 19 microRNA Cluster (C19MC) is either amplified or fused to other genes, and is involved in 'embryonal tumors with multilayered rosettes (C19MC-altered)', which were formerly grouped under primative neuroectodermal tumours.[15, 16, 221, 222] C19MC open reading frames spread throughout the cluster were targeted to potentially capture variations in read depth.[222, 223] Single bases within open reading frames in the C19MC were targeted, with the knowledge that the RNA baits that would be selected for the capture kit were 120nt long, and reads overlapping this larger region would be captured. *RELA* fusions in ependymomas tend to occur in regions up to the end of exon 3, so these regions were chosen in FLCP-1.[224] In order to cover glioblastomas, *BRAF*, *IDH1*/2, *PTEN*, *EGFR*, *MET* and *PDGFRA* were targeted.[48] *BRAF* fusions related to Pilocytic Astrocytomas (PAs) were targeted by the inclusion of introns and exons where which commonly harboured the breakpoints.[225] A wide range of targets was used in order to aid in detection and monitoring of as many types of PBT as possible.

The final panel's targeted regions totalled 88,849bp, and the RNA baits covered a total of 117,646bp.

Table 2.3: Genomic regions targeted by the FLCP-1 capture panel (hg19)

Gene/region	Target type(s) and relevant transcript	Mutation type(s)	Targeted regions (unless all coding exons targeted)
ACVR1	all coding exons (9)	point	
ATRX	all coding exons (36)	point	
BCOR	all coding exons (15)	point	
BCORL1	all coding exons (14)	point	
BRAF	exon 15, start of intron 8-9 to end of exon 11 in transcript ENST00000288602.6	point, fusion	chr7:140453075-140453193, chr7:140481376-140494107
C19MC	open reading frames	amplification	chr19:54177273, chr19:54186474, chr19:54186475, chr19:54191367, chr19:54192903, chr19:54210496, chr19:54210519, chr19:54211903, chr19:54217644, chr19:54219769
CTNNB1	hotspot on exon 3	point	chr3:41266096-41266138
EGFR	hotspot exons 18 to 21 in transcript ENST00000275493.2	point	chr7:55241614-55241736, chr7:55242415-55242513, chr7:55248986-55249171, chr7:55259412-55259567
H3F3A	all coding exons (3)	point	
HIST1H3B	all coding exons (1)	point	
IDH1	hotspot exon 4 in transcript ENST00000415913.1	point	chr2:209113093-209113384
IDH2	hotspot exon 4 in transcript ENST00000330062.3	point	chr15:90631819-90631979
KRAS	hotspot exons 2 to 4 ENST00000311936.3	point, rearrangement	chr12:25398208-25398329, chr12:25380168-25380346, chr12:25378548-25378707
MET	all coding exons (21)	amplification	
MYB	all coding exons (21)	amplification, focal deletion	
MYC	all coding exons (3)	amplification	
MYCN	all coding exons (3)	amplification	
NF1	all coding exons (63)	point	
NRAS	hotspot exons 2 and 3 in transcript ENST00000369535.4	point	chr1:115258671-115258798, chr1:115256421-115256599
NTRK1	10 bases upstream of exon 8 to 20 bases downstream of exon 10 in transcript ENST00000524377.1	fusion	chr1:156843414-156844439
PDGFRA	all coding exons (24)	point, amplification	
PIK3CA	hotspot exons 2, 3, 5 to 8, 10 and 21 in transcript ENST00000263967.3	point	chr3:178916538-178916965, chr3:178917478-178917588, chr3:178921332-178921577, chr3:178922291-178922376, chr3:178927383-178927488, chr3:178927974-178928126, chr3:178935998-178936122, chr3:178951882-178952100
PIK3R1	all coding exons (19)	point	
PPM1D	all coding exons (8)	point	
PTCH1	all coding exons (26)	point	

PTEN	all coding exons (9)	point	
RELA	exon 1 to 251 bases into intron 4-5 in transcript ENST00000406246.3	fusion	chr11:65429408-65430565
SMARCA4	all coding exons (36)	point	
SMARCB1	all coding exons (9)	point	
SUFU	all coding exons (13)	point	
TERT	promoter hotspots	point	chr5:1295228-1295229, chr5:1295250-1295251
TP53	all coding exons (14)	point	

2.2.3.1 Implementation of FLCP-1 with an Agilent XT hybrid capture kit

Prior to the project, Rubicon Genomics had developed modifications to the Agilent XT hybrid-capture protocol, which made the system compatible with the ThruPLEX Tag-seq library preparation kit.[226, 227] Capitalising on this prior work, the Agilent protocol was implemented with Rubicon Genomics' modifications. Since the system worked well during testing, no modifications to the protocol were made during the course of the project.

The purified libraries from library preparation were first mixed with blocking oligos, to reduce the effects of daisy-chaining during capture.[228, 229] The libraries were then heated to 95°C, to denature the double-stranded DNA, then allowed to hybridise to biotinylated RNA probes targeted to the regions in FLCP-1 at 65°C.[227, 230] At the end of a 24h hybridisation, the DNA which bound to the biotinylated RNA probes was captured by affinity to streptavidin-coated magnetic beads. The beads were pulled to the side of the tube using a magnetic rack, and washed with proprietary buffers. The bead-bound DNA was then PCR amplified using Herculase II Fusion DNA polymerase, and associated buffers.

During the preliminary trial, the magnetic rack used for purification of bound DNA was an unbranded rack meant for 96-well plates. This rack was able to separate the beads from suspension, albeit slowly. During the HC1 and HC2 trials, a stronger custom magnetic rack was designed, 3D printed and used during AMPure purification steps, and during the hybrid capture step (Figure 2.3). The faster separation of beads during the 65°C wash steps of the hybrid capture allowed the beads to be maintained at 65°C more reliably.

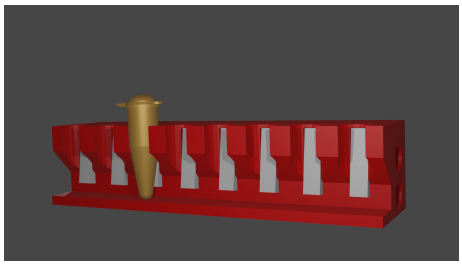


Figure 2.3: A 3D render of the high-strength magnetic tube rack used for the quick separation of beads during hybrid capture.

The red shape was 3D printed, the white cuboids are neodymium magnets, and the gold is an example 0.2ml tube.

2.2.4 Calculating the capture efficiency from sequencing data

To achieve the aim of optimising the sequencing depth of each sample, an accurate method of measuring the proportion of reads which were within the capture panel's covered regions was necessary. This 'capture efficiency' was measured by creating a .bed file which listed all regions covered by FLCP-1, and finding the number of bases which fell within this .bed file as a proportion of all sequenced. Since a capture kit would capture DNA molecules which partially overlapped the probes of the panel, the covered regions were padded by 150bp on either side using BEDTools to create extended covered regions (Figure 2.4).[231] Any read families which overlapped the regions of this file were considered on-target. An inverse .bed file, which listed regions that were not in the extended covered regions of FLCP-1, was also created.

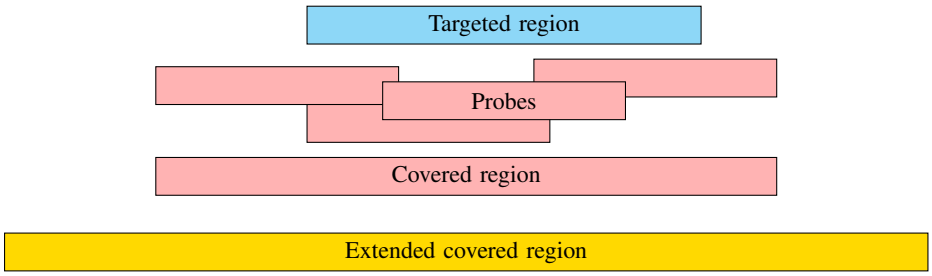


Figure 2.4: The relationship between a targeted region, the RNA probes which cover this region, and regions used for capture efficiency analysis.

The targeted region is covered by probes which extend outwards from either end of the targeted region. The covered region is, therefore, larger than the targeted region. An extra 150bp was added to either end of the covered region for the analysis of capture efficiency.

For the HC1 and HC2 studies, the capture efficiency was computed from the Quality Check/Control (QC) MiSeq data, downloaded from Illumina's BaseSpace platform onto one

of the Legion or the Myriad High Performance Computing Clusters at UCL. Each pair of FASTQ files from each replicate was run through the prototype SNV calling pipeline from trimming to Connor (described in 3.3.2). Briefly, the raw reads were trimmed using Trim Galore!, aligned using BWA mem, post-processed using Samtools, and collapsed into read families by barcode using Connor. Samtools was used to create a .bed file which contained the family depth at each position of the extended covered regions of FLCP-1, for each collapsed file. The same was done for the positions which were not in the extended covered regions. The family depths in each file were summed, and divided by the total family depth at all positions in the genome, to give the on-target percentage and the off-target percentage.

2.2.5 Implementation and improvement of library normalisation and pooling procedures

A main aim of the project was to optimise the number of reads per sample, balancing the need to sequence each DNA molecule in the sample multiple times, whilst maximising the number of samples per run. This was dependent on the percentage of a sample's original DNA molecules that the wet-lab workflow was able to sequence, the capture efficiency of the XT kit, and the accuracy of quantification during library normalisation and pooling.

During the preliminary study, the percentage of molecules which the wet-lab workflow was able to sequence, and the capture efficiency was not known. The decision was taken to deliberately over-sequence the molecules, potentially creating read families of hundreds of reads, to see how many molecules were able to pass through the wet-lab workflow. Following post-capture amplification, the libraries were quantified using a Bioanalyzer 2100 with High Sensitivity DNA kit. Based on these quantifications, the libraries were normalised to 8nM. These libraries were pooled equally, and the result was sequenced on the NextSeq 500.

Following the preliminary study, the sample types were changed, necessitating a change in extraction methodology, and the resulting libraries were sequenced on the HiSeq 2500 platform. Normalisation was developed into a two-step procedure, to increase the accuracy of the amount of each library in a pool. First, libraries were quantified using a Qubit dsDNA HS Assay kit, and run on a Bioanalyzer 2100 with High Sensitivity DNA kit. The mean fragment sizes, read from the Bioanalyzer traces, and the ng/μl Qubit results were used to calculate the molarity of the libraries in nM. Each library was then normalised to 3nM, and the result was quantified again on the Qubit. The Qubit results were converted

into molarity as before, and the result was used to calculate the volume of each library to add to the pool (in this study, the volumes were between 2.45µl and 3.82µl).

The HC2 study was conducted on samples with highly variable DNA inputs. The efficiency of the initial molecules' passage from extraction to sequencing in the HC1 study showed that 10-15% of molecules in the sample produced sequenced read families. The aim was to create a normalisation procedure which ensured that each sequenceable molecule from the input was sequenced an average of 5 times at least. The initial input amount in ng of each sample was used to estimate the number of bases' worth of sequencing required to sequence it. A HiSeq 2500 with a Rapid cluster kit v2 and a Rapid SBS v2 reagent kit could produce 1.2 billion paired-end reads per flow cell, and could handle read lengths of up to 250bp/read. From these pieces of information, samples were assigned to one of two pools, each of which was to be sequenced on a Rapid v2 kit on a single flow cell. The percentage of the pool taken up by each sample's library was tuned such that each molecule should be sequenced more than 5 times, to allow for lower data outputs resulting from overclustering, or pipetting errors. Similarly to the HC1 study, libraries were run on the Qubit and Bioanalyzer 2100, and their molarities were calculated. The libraries were normalised to 6nM, rerun on the Qubit, and their new molarities were used to direct pooling. The final volume of the pool was increased, so that the volume of each library added to the pool was maximised. This was to reduce the effects of pipetting errors, which were likely the cause of the read count variability found during the HC1 study.

2.2.6 Creation and improvement of the overall pipeline

One of the aims of the project as a whole was to create a workflow for targeted sequencing of liquid biopsy cfDNA. Taking advantage of work that had been done before and concentrating on truly novel aspects of the project meant that off-the-shelf kits and methodologies were used, where applicable. Within the first 4 months of the project, the author was approached by Rubicon Genomics to beta test the barcoded library preparation kit which would later be known as the ThruPLEX Tag-seq kit. Initial trials of this kit were performed according to the manufacturer's protocols, on control material, and the resulting libraries were sequenced, and assessed for suitability as a component of the project. The results of the trial showed promise, but that there were challenges to overcome regarding the data analysis solutions. This library preparation system was used for the rest of the project, whilst efforts focused on developing bioinformatics solutions to overcome these deficiencies.

The course of development necessitated the replacement of some of some stages of the workflow, but the overall form of the pipeline remained the same as in Fig. 2.1.

2.2.6.1 Pre-library preparation sample handling

The library preparation kits required purified DNA, fragmented to lengths suitable for Illumina sequencing where necessary. The number of library amplification PCR cycles to use during library preparation were dependent on the amount of DNA input, so quantification was necessary. To achieve this, the raw CSF and cystic fluid samples in the HC1 and HC2 cohorts were extracted, sheared, then quantified. The plasma samples were extracted, then quantified without shearing. The following describes how this methodology changed between the studies.

To assess suitability of the ThruPLEX Tag-seq kit as part of this project's wet-lab workflow, control material was needed. The Horizon Discovery Multiplex I cfDNA Reference Standard Set (Cat. HD780) was chosen as input DNA for the preliminary study. These controls contain variants at known VAFs (5%, 1%, 0.1%, and 0%) at specific sites in the genome, which can be used to test the sensitivity and specificity of variant allele detection.[232] HD780 DNA is extracted from human cell line DNA, and sheared to an average of 160bp, allowing for the elimination of DNA extraction and shearing from the preliminary workflow. Since the Horizon Discovery DNA was provided at known concentrations, these samples were not quantified.

The HC1 cohort was partially made up of patient CSF and cystic fluid samples, for which extraction and shearing was necessary. The HC1 cohort also included plasma samples, which required extraction only. The volume of each CSF, cystic fluid or plasma sample was made up to the nearest millilitre using phosphate-buffered saline, and the sample was extracted using a QIAGEN Circulating Nucleic Acid Kit and a vacuum manifold, according to the kit's instructions. The purified DNA was eluted in 50µl of nuclease-free water. CSF samples were then sheared in Covaris microTUBE-50 AFA Screw-Cap tubes, according to the parameters in Table 2.2. Finally, the samples were quantified using a Bioanalyzer 2100 with a High Sensitivity DNA kit.

Following the results of the HC1 study, it was clear that the amount of DNA in each sample was such that taking 10µl of the 50µl extraction eluate was not sufficient for sensitive variant detection. The Version 2 methodology included a concentration step before quantification, to maximise the amount of DNA passed to the library preparation kit. The

raw samples were extracted and sheared using the Version 1 methodology. After shearing, samples were placed in custom filtered 0.5ml tubes in an Eppendorf Concentrator 5301, and concentrated down to a volume below 10µl at 30°C. The volume of each sample was then made up to 11.5µl with nuclease-free water. Initially, the concentration of DNA for 3 ATRT CSF samples was determined by running them on a Bioanalyzer 2100 with a High Sensitivity DNA kit. This gave lower DNA outputs from library preparation than expected. A different quantification method was sought to accurately quantify the DNA before library preparation. A suite of droplet digital Polymerase Chain Reaction (ddPCR) assays had been developed during the HC1 study, and due to the availability of the assay, the concentration of the rest of the samples was found by analysing 1µl using ddPCR Assay Pair 1. These assays are described in detail in Section 3.3.4. Briefly, the dual channel ddPCR assay targeted *SMARCB1* and RPP30, and the possibility of a deletion in *SMARCB1* affecting the results from the *SMARCB1* CNV assay meant that, in the event of disagreement between the RPP30 assay and the *SMARCB1* assay, the RPP30 assay alone would be used for quantification. All subsequent HC2 study samples were quantified using this method.

2.2.6.2 Library preparation

During the preliminary study, the Horizon Discovery control samples were tested on a beta test version of Rubicon Genomics' ThruPLEX Tag-seq Kit (96 barcode pair, dual indexed version), according to the manufacturer's specifications. Due to the closed-source nature of the reagents within this kit, no optimisations of this process were made by the author. The information contained in the rest of this paragraph, concerning the library preparation process, has been published as part of the product literature for the ThruPLEX Tag-seq kit, and is illustratd in Figure 2.5.[129, 233, 234] A 10µl sample of input DNA was subjected to end-repair, which terminated each double-stranded molecule with a blunt end. This was followed by the blunt ligation of ThruPLEX Tag-seq hairpin adaptors, each of which contained a 6nt molecular barcode, to both ends of each template molecule. The ligated molecules were subjected to a proprietary hairpin breaking step, followed by successive rounds of PCR with primers which contained Illumina P5 and P7 adaptor sequences. The final product was amplified libraries containing P5 and P7 Illumina adaptors, sample-specific indexes, barcodes, and spacer "stem" sequences.

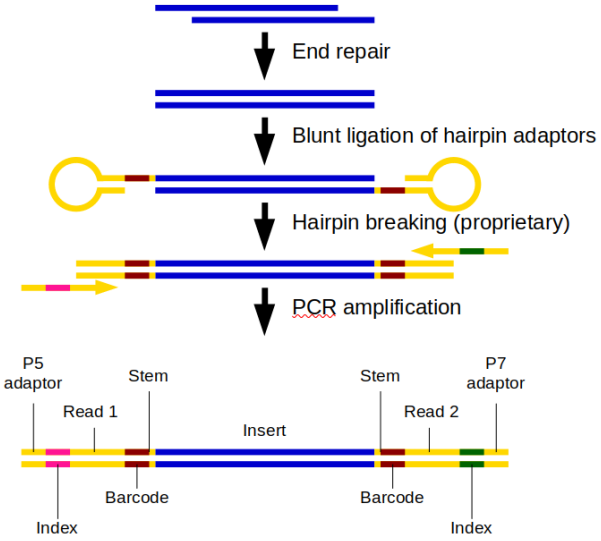


Figure 2.5: Rubicon ThruPLEX Tag-seq Kit library preparation, and final library structure.

An overview of the library preparation process, based on information in the ThruPLEX Tag-seq kit protocol. The P5 and P7 adaptors, dual indexes, and Read 1 and Read 2 start sites are compatible with Illumina sequencers. Each read contains a 6nt barcode, and an 8-11nt stem sequence which varies in length at the 5' end, but not sequence. Each read from the sequencer produces data which contains the barcode at the 5' end, an 8-11nt stem, followed by the sample sequence.

The HC1 study was a small-scale trial of the overall wet-lab workflow, from initial sample handling by staff at the GOSH/UCL ICH to sequencing on the Illumina HiSeq 2500. The aims of this study included the improvement of the hybrid capture efficiency step of the workflow, and the assessment of the new capture efficiency of this Version 1 workflow. These changes to the wet-lab workflow, combined with the high sequencing depth of the HiSeq 2500, meant that it was necessary to run control material through before the HC1 cohort's samples. Two replicates of the Horizon Discovery 5% Reference Standard were used in this assessment. For each replicate, 30ng of the standard was made up to 10µl with nuclease-free water, and used as input for a Rubicon Genomics (now Takara Bio USA) ThruPLEX Tag-seq kit (48 barcode, single indexed version). Library preparation was performed according to the same instructions as in the preliminary study, and the libraries were purified and quantified similarly. For the main HC1 study, 10µl of each sample was used as input for the same ThruPLEX Tag-seq kit, and processed similarly.

The HC2 study was a larger study, with incremental improvements to the wet-lab workflow. Due to the number of samples, a 96 barcode production version of the ThruPLEX Tag-seq kit was used. Each sample was concentrated down to 11.5µl, quantified using a

ddPCR assay, and the quantification results guided the number of PCR cycles for library amplification. Where the sample concentration was above 10ng/µl, 100ng was used as an input to the library preparation process, and the sample volume was made up to 10µl using nuclease-free water. If not, 10µl of the sample was used for library preparation. The amount of DNA added to each library preparation reaction is displayed in Tables D.1 and 2.6.

2.2.6.3 Steps following library preparation

In all three studies, the amplified libraries were mixed with equal volumes of Agencourt AMPure XP beads in a 0.2ml tube, and mixed by inverting. A magnetic rack was used to pellet the beads and the supernatant was removed. The pellets were washed twice with 200µl of 70% ethanol, then allowed to dry at room temperature for 3-10min. 30µl of nuclease free water was added to the pellets and the mixture was vortexed, to elute the DNA from the beads. The beads were removed using a magnetic plate, and the eluate was quantified using an Agilent Bioanalyzer 2100 with a High Sensitivity DNA kit.

Where the yield of a given sample was below 500ng, libraries were pooled in a manner which was proportional to their library preparation inputs. This meant that the number of sequenced reads from each library was likewise proportional to the amount of DNA used as library preparation input, and that the hybrid capture step was free from any deviations from the expected DNA concentration. This was possible, since the libraries' molecules were indexed prior to this stage. Pooled and separate libraries were entered into the hybrid capture stage, as described in Section 2.2.3.1, then normalised and pooled, as described in Section 2.2.5.

2.2.6.4 Quality control and data production sequencing runs

During the preliminary study, it was not known whether the sequencing output provided by a NextSeq 500 in High Output mode was enough for the project, as the efficiency of the wet-lab workflow had not yet been determined. This was determined experimentally, by sending the pool of control libraries from the preliminary study to the UCL Genomics facility for use on their sequencer. The results of the preliminary study on control material showed that a higher sequencing depth was needed to recover the maximum number of read families from input DNA from larger cohorts in a cost-effective manner. The HC1 study's sequencing step was switched over to the Illumina HiSeq 2500.

The results from the preliminary study showed that there was improvement to be made in terms of pooling accuracy. To reduce the probability of needing to re-sequence a library

due to poor pooling accuracy during the HC1 study, a QC run on an Illumina MiSeq sequencer with a MiSeq Reagent Nano Kit v2 (300-cycles) was included. The number of reads corresponding to each library was used to assess the pooling accuracy ahead of the HiSeq run. The 3nM pool was treated as a 4nM pool for loading onto an Illumina MiSeq sequencer with a MiSeq Reagent Nano Kit v2 (300-cycles). Following quality control, the libraries were run on an Illumina HiSeq 2500 using HiSeq rapid SBS v2 kits (200 cycles + 50 cycles) and a HiSeq Rapid PE cluster kit v2. This was done in Rapid, paired-end mode with a read length of 139 per end.

The results of the HC1 study showed that the bottleneck in the conversion of DNA molecules into sequenced read families was not the sequencing step, but the input. The sequencing platform allowed for a greater number of libraries to be sequenced in a single run than were in the HC1 cohort, so neither the QC step nor the HiSeq methodology were changed between the HC1 and HC2 studies.

2.3 Results

2.3.1 The samples and cohorts of the preliminary, HC1 and HC2 studies

2.3.1.1 The preliminary study

The samples used in this study were from the Horizon Discovery Multiplex I cfDNA (cat. HD780) control set. They would aid in sensitivity assessment since they contain known mutations at a range of VAFs. As a part of a beta test, a pre-release version of the Rubicon Genomics ThruPLEX Tag-seq Illumina library preparation kit was used for library preparation. The libraries produced by this kit were purified using AMPure XP beads and quantified using an Agilent Bioanalyzer. Table 2.4 details the samples and controls and input amounts used for library preparation, and the yields after purification.

Table 2.4: The samples used in the preliminary study, and their yields after library preparation.

Sample	Sample	Input (ng)	Yield (ng)
1	Horizon Discovery cfDNA Reference Standard 0% VAF	30	1400
2	Horizon Discovery cfDNA Reference Standard 0% VAF	30	1300
3	Horizon Discovery cfDNA Reference Standard 5% VAF	30	1400
4	Horizon Discovery cfDNA Reference Standard 5% VAF	30	1500
5	Horizon Discovery cfDNA Reference Standard 1% VAF	30	1700
6	Horizon Discovery cfDNA Reference Standard 1% VAF	30	2200
7	Horizon Discovery cfDNA Reference Standard 0.1% VAF	30	1200
8	Horizon Discovery cfDNA Reference Standard 0.1% VAF	30	1400

2.3.1.2 The HC1 study

The HC1 study was comprised of two runs. The first run was a test, used to verify the functioning of the Version 1 wet-lab workflow between library preparation and MiSeq QC sequencing, as developed during the preliminary study. This run was made up of two technical replicates of the Horizon Discovery 5% Reference Standard, separated into independent 30ng library preparation reactions, and treated as separate samples from then on. As described later, in 2.3.3, the capture efficiency of this run was 37.56% $\pm$6.02%. This limited the number of samples which could be sequenced per run on the HiSeq for the second run, so a small cohort of CSF, cystic fluid and plasma samples was selected to test the ability of barcoded NGS to detect variants in these liquid biopsies. This cohort was named HC1, and is presented in Table 2.5. These samples were collected retrospectively by collaborators

at GOSH, with no defined collection protocol, and were used as an exploratory dataset for methods development. This is reflected in the variable, and often small, amounts of DNA available for sequencing, after extraction. Two of the samples failed library preparation, but the overall wet-lab workflow success rate of 89% was promising, with library preparation failures explained by their low inputs.

Table 2.5: The clinical samples used in the HC1 study.

Sample	Diagnosis	Patient	Sample type	Input (ng)	Status
1	ATRT	A1	CSF	0.69	Sequenced
2	ATRT	A2	CSF	12	Sequenced
3	ATRT	A2	CSF	0.61	Sequenced
4	ATRT	A5	CSF	37	Sequenced
5	ATRT	A5	CSF	Undetected	Library prep failed
6	DIPG	A6	Plasma	6.1	Sequenced
7	DIPG	A8	Plasma	4.9	Sequenced
8	DIPG	A9	Plasma	2.4	Sequenced
9	DIPG	A10	Plasma	2.0	Sequenced
10	Medulloblastoma	A11	Plasma	0.18	Sequenced
12	ACP		Cystic fluid	45	Sequenced
13	ACP		Cystic fluid	12	Sequenced
14	ACP		Cystic fluid	20	Sequenced
15	ATRT	A2	CSF	0.12	Library prep failed
16	ATRT	A2	Plasma	4.6	Sequenced
17	DIPG	A7	Plasma	4.4	Sequenced
18	ACP		Plasma	3.1	Sequenced
19	ACP		Plasma	1.8	Sequenced

2.3.1.3 The HC2 and HC2C cohorts of the HC2 study

This study focused on CSF and cystic fluid. Samples of a variety of tumour types were chosen, and the cohort was biased towards ATRT and Malignant Rhabdoid Tumour (MRT) samples. This was to capitalise on success in detecting *SMARCB1* alterations during the HC1 study, as presented in Chapter 3. A total of 41 CSF samples were used in this phase. This included 30 ATRT samples (11 patients), 6 MRT samples (3 patients), 1 PA sample, 1 DIPG CSF sample, and 3 *WNT*-Activated Medulloblastoma (WAM) samples (2 patients). In an attempt to improve on the results of the plasma sequencing from Chapter 3, 3 DIPG

plasma samples (3 patients) were also sequenced. Due to the size of the cohort, details of the samples are in Appendix D.

Following the success with cystic fluid during the HC1 study, the HC2 study included the HC2C cohort: ten cystic fluid samples, five each from two patients (Table 2.6). These samples were collected over the course of a 22-25 day Interferon-α (IFN-α) treatment programme, and at some time points, there were multiple samples. These samples were the aspirates taken from the cysts immediately prior to injection of IFN-α to the cyst. Patient C1 reportedly responded well to treatment, whilst Patient C2 showed no change. These samples were not collected for a prospective study, so the collection methodology was not standardised. These samples not only allowed for the testing of reliability of the workflow, but also allowed for the assessment of whether cellular fractions of cystic fluid yielded differing amounts of mutant DNA from their cell-free counterparts. Details of the samples are presented in Table 2.6.

The samples of both the HC2 and HC2C cohorts were run on the Version 2 wet-lab workflow concurrently, as described in the Methods section of this chapter.

Table 2.6: Details of the samples in the HC2C cohort, and the mean family depth sequenced from those samples.

Patient	Sample	Type	Treatment day	Library preparation input (ng)	Mean family depth
C1	HC2C-1	Unspun cystic fluid	1	100	1716
C1	HC2C-2	Supernatant	1	100	3038
C1	HC2C-3	Supernatant	4	100	3166
C1	HC2C-4	Supernatant	10	100	2399
C1	HC2C-5	Supernatant	25	95.4	2460
C2	HC2C-6	Supernatant	1	100	2296
C2	HC2C-7	Cellular fraction	1	100	2638
C2	HC2C-8	Supernatant	11	32.3	789
C2	HC2C-9	Cellular fraction	22	100	3215
C2	HC2C-10	Supernatant	22	100	3195

For both cohorts, initial pre-library preparation DNA quantification was performed using a Bioanalyzer 2100. These quantification results were used in the determination of the number of PCR cycles performed during library amplification, for the first batch of

samples run through the workflow. Unexpected results during the first batch of fourteen samples (HC2-16, HC2-18, HC2-19 and HC2-28 and all ten HC2C samples) during library preparation caused a revision of the pre-preparation quantification method. Sample HC2-19 performed as expected during library preparation, but the three other CSF samples (HC2-16, HC2-18, and HC2-28) yielded lower outputs of DNA than expected. The entire HC2C cohort of ACP samples were unaffected, likely due to their high DNA inputs. To attempt to rescue the three first batch CSF samples and use them for further analysis, the samples were subjected to further rounds of PCR (4, 5, and 5 respectively) using Agilent's post-capture amplification protocol.[235] This yielded an optimal 750ng for HC2-16, but samples HC2-18 and HC2-28 yielded only 150ng and 58ng respectively. This indicated that either the Bioanalyzer quantification before library preparation was inaccurate, or that there was significant degradation when the samples were stored at 4°C between quantification and library preparation. To overcome both possible challenges, all remaining samples of the HC2 cohort were quantified using ddPCR, then frozen before library preparation. The resulting calculated inputs to library preparation are presented in Figure 2.6. As shown, the amount of DNA in most CSF samples was below 10ng, which posed a challenge for the detection of rare variants in these samples. Library preparation succeeded for all ddPCR quantified samples, apart from a single MRT sample.

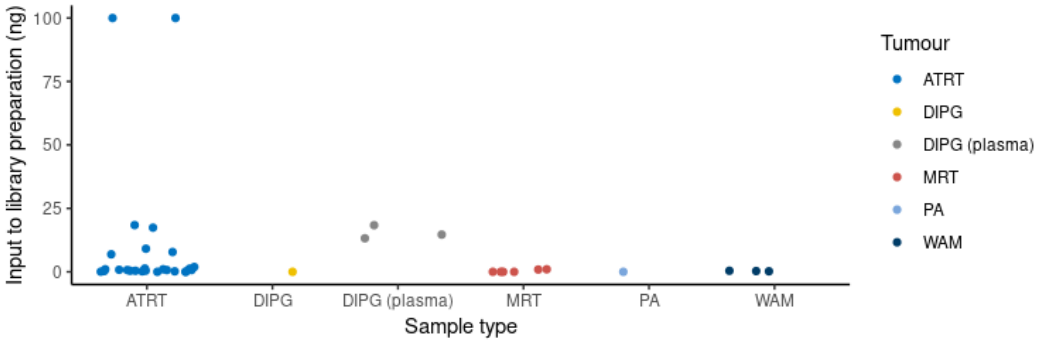


Figure 2.6: DNA inputs for library preparation of the HC2 cohort.

Input amounts were calculated ddPCR results from 1µl of each sample, and the four CSF samples prepped using Bioanalyzer quantification are omitted from this figure. Unless otherwise indicated, all sample types were CSF. A zoomed version of this figure (Figure D.1) is available to highlight the differences between the low input samples.

2.3.2 Developing procedures for conversion of diverse liquid biopsies into short DNA fragments, suitable for sequencing

Work for this began at the beginning of the HC1 study, in preparation for the cohort of CSF and cystic fluid samples. The cystic fluid samples were aspirated from ACP cysts, and frozen. This meant that the samples, once thawed for processing, would contain genomic DNA from any cells which lysed during the freeze-thaw process. Similarly, the unspun CSF samples could contain near full-length genomic DNA. Since circulating tumour cells within the CSF are used for disease detection, and there may have been mutation-harbouring cells within the cystic fluid, it was prudent to render the long DNA into a sequenceable state.[7, 8] The Covaris E220 Evolution sonicator was chosen as a means of shearing DNA due to its random nature, as opposed to endonuclease digestion, and high advertised shearing accuracy.

Due to the lab's favourable experience with the QIAGEN Circulating Nucleic Acid extraction kit, and its inclusion in contemporary methodologies, this kit was chosen for the isolation of liquid biopdy samples.[109, 236] Previous work showed that some of the CSF DNA was much shorter than full-length DNA, and personal communication with QIAGEN revealed that genomic DNA which passed through the columns supplied with the Circulating Nucleic Acid Kit was sheared to approximately 10kb.[116] It was necessary to test whether the shearing conditions suggested by Covaris were able to preserve short DNA, whilst shearing longer DNA to a length which was compatible with Illumina sequencing.

Two different types of control material were created for this work: one which represented short DNA in the 100-200bp range (named sDNA), and one which represented the output from nucleic acid extraction (named lDNA). Initial testing with the Covaris parameters revealed that under conditions which brought the lDNA down to sequenceable lengths, sDNA yields were very low. This led to the exploration of the shearing parameter space, first by trialling the more gentle condition sets provided by Covaris, then by reducing the severity of the conditions in a stepwise fashion. A total of 39 conditions were trialled, leading to the adoption of the conditions in Table 2.2. These shearing conditions were able to preserve 77% of the sDNA, whilst shearing lDNA down to between 100 and 400bp (Figure 2.7).

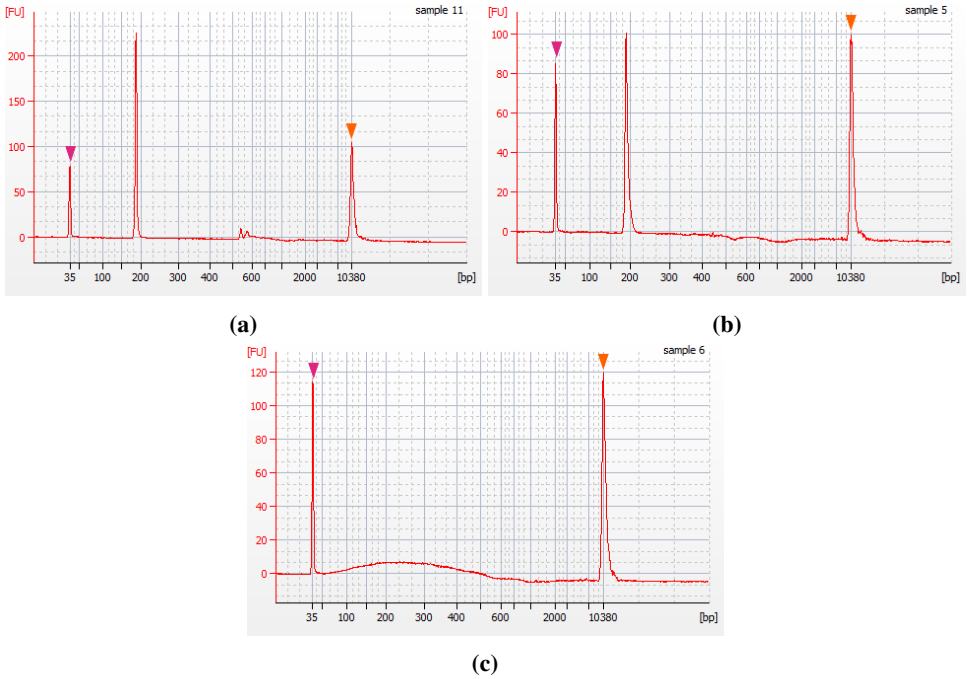


Figure 2.7: Bioanalyzer traces showing the effects of shearing on 182bp sDNA and ~10kb lDNA.

2.7a) Unsheared 182bp sDNA. **2.7b)** sDNA following shearing, showing a 23% loss of DNA. **2.7c)** Sheared lDNA showing a peak between 100 and 500bp, centred at 200-250bp. In all traces, pink arrows indicate lower markers, and orange markers indicate upper markers.

2.3.3 FLCP-1 - a hybrid capture panel targeting Paediatric Brain Tumours

The original development of FLCP-1 was done during the preliminary study. This panel was primarily targeted at a range of PBTs, with some regions unique to single or small numbers of PBTs, and some which were more common. The aim of the panel was to be a compromise between the chance of disease detection, the ability to distinguish between tumour types, and the amount of sequencing needed to perform both.

The panel focused on DNA from targeted regions of the genome by binding these to magnetic beads, and washing away DNA from other regions. The efficiency by which off-target DNA was removed from the libraries was never 100%. The proportion of the DNA in a sample which was targeted would therefore affect the depth to which each sample would need to be sequenced, for there to be enough reads from each molecule to form reliable families. Moxie Genomics' Curio platform, upon which the analysis of the preliminary dataset was based, was unable to give a value for the on-target percentage at its beta stage

of development. Since there were changes to the wet-lab workflow between the preliminary study and the HC1 study which necessitated the running of control material through the Version 1 workflow, the decision was taken to wait for the HC1 control data before assessing the on-target percentage. These samples were sequenced on the MiSeq, and the resulting data was aligned to the GRCh38 genome. The mean on-target percentage calculated for this run was 37.56% $\pm$6.02%. This rate was lower than expected, and could have been caused by inefficiencies in the magnetic rack used during the 65°C wash steps causing drops in temperature and capture efficiency, or by a poorly calibrated thermocycler maintaining a suboptimal temperature. With a finite number of reads possible in a sequencing run, the number of samples which could be put on a sequencing run were limited.

The HC1 cohort was run through the same wet-lab workflow as the HC1 control material, but a new magnetic rack, and a new, well-calibrated PCR thermocycler promised an increase in the on-target percentage. Prior to the calculation of this run's on-target percentage, the initial pipeline detailed in Chapter 3 had been developed, and the HiSeq data was collapsed into families. This removed PCR duplicates, and resulted in an on-target percentage calculation which better reflected the final data of a run. This yielded a mean on-target percentage of 49.49% $\pm$20%, and a median of 49.44%. Based on the more favourable capture results from the HC1 cohort, larger numbers of samples were loaded onto the runs which made up the HC2 study. The sequencing datasets were analysed by the pipeline detailed in Chapter 4, resulting in collapsed families, similar to those from the HC1 cohort run. Overall, the run showed a mean on-target percentage of 58.01% $\pm$19%, and a median of 63.16%. These statistics are presented visually in Figure 2.8.

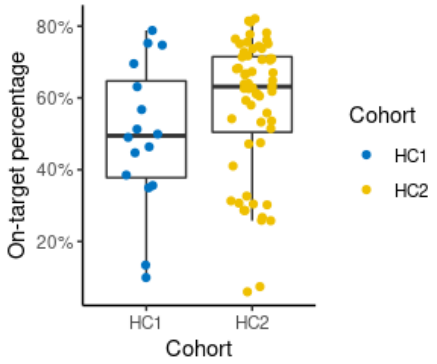


Figure 2.8: The on-target percentages achieved with the HC1 and HC2 cohorts.

The horizontal lines of the boxes in this plot are defined as the upper quartile, median, and lower quartile of the on-target percentages of each plot. The whiskers extend to the highest and lowest datapoints, except when the datapoints are more than 1.5x the inter-quartile range from the closest quartile line.

2.3.4 Implementation and improvement of library normalisation and pooling procedures, and quality control

The optimisation of the number of sequenced reads per molecule of sequenceable template DNA depended on a number of factors which were unknown at the inception of this project. An iterative approach was taken, in which each study produced results which influenced the design of the next version of the workflow. There were four pieces of information which were needed for accurate pooling: the original amount of DNA in the sample, the number of molecules lost in the production of captured and amplified libraries, the hybrid capture efficiency, and the accurate concentration of the captured libraries for pooling purposes. A final limiting factor, which imposed an upper limit on the amount of each library in the pool, was the number of clusters that a given sequencing flow cell was capable of producing. During the preliminary trial, none of the pieces of information were known, so an arbitrary target of equimolar pooling, followed by over-sequencing, would guarantee the maximum data for the chosen flow cell. The resulting data would then be used to determine all of the above parameters.

The preliminary study's libraries were quantified post-capture using a Bioanalyzer 2100, and Figure 2.9a highlights the need for proper normalisation. The Bioanalyzer normalisation results (Figure 2.9b) did not, however, show a strong correlation with the number of reads for the sample (Figure 2.9c). The final numbers of reads across the cohort (Figure 2.9d) had a mean of 39.0 million, and a coefficient of variation of 17.7%. This clearly showed that both quantification and normalisation were portions of the workflow which could be improved.

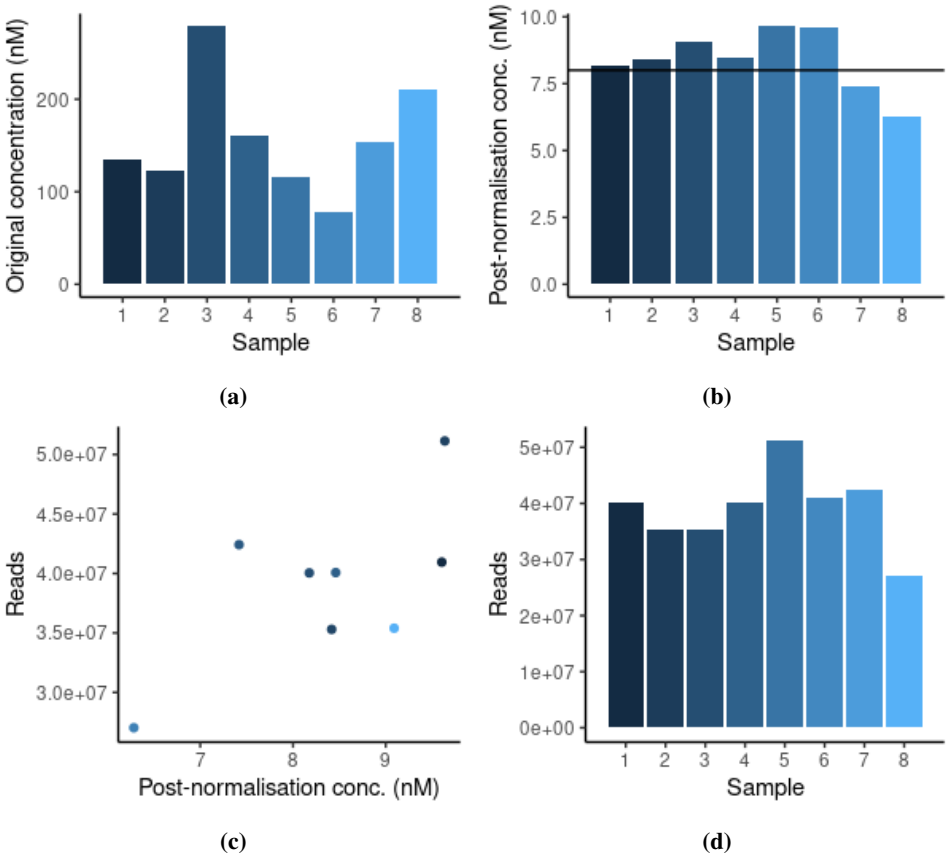


Figure 2.9: Pooling of samples in the preliminary study.

2.9a) The concentration of each library as produced by the post-capture amplification step. These samples collectively had a coefficient of variation of 39.9%. **2.9b)** The concentration of each library after normalisation to a target of 8nM (horizontal line). The coefficient of variation was reduced to 13.5%, with a mean of 8.4nM. **2.9c)** A scatter plot showing the weak correlation between Bioanalyzer quantification and number of reads in the R1 FASTQ file. The Pearson's correlation coefficient was 0.67. **2.9d)** The numbers of reads in the R1 FASTQ file, showing a coefficient of variation of 17.7%.

Following the results of the preliminary study, the HC1 study's methodology incorporated the two-step normalisation procedure, as described in 2.2.5. The number of reads per sample was also directly measured by the MiSeq QC sequencing run, bypassing the use of the Bioanalyzer, which had been shown to be unreliable for this purpose. The change of wet-lab methodology and final sequencing method meant that this cohort was, like the preliminary cohort, pooled equimolarly and over-sequenced. The result was a 15.2% coefficient of variation in the read numbers; a 14% decrease in variability versus the preliminary study's results.

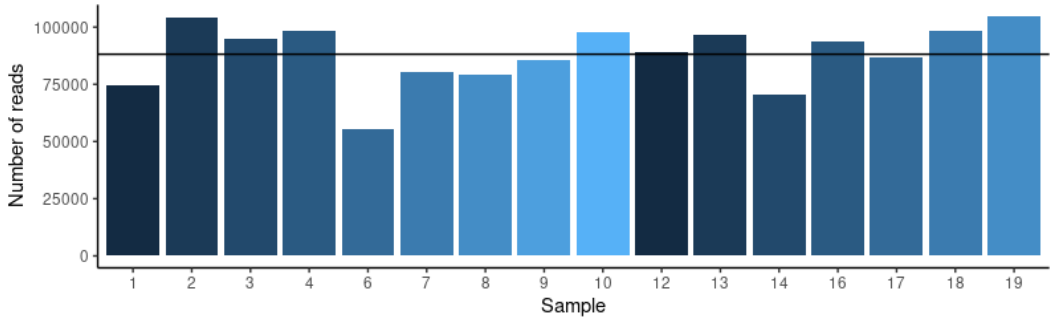


Figure 2.10: Pooling results following a MiSeq QC for the HC1 study.

The raw read counts for each sample in the MiSeq QC run. The coefficient of variation in the read counts is 15.2%, and the mean is 88134 reads (horizontal line).

A newer method of pooling samples for sequencing was used during the HC2 study. This method was designed to enable users to utilise the minimum number of reads to ensure that each starting molecule had enough reads to make a family. All samples were pooled using this method, and each pool was run on an Illumina MiSeq for QC. Figure 2.11 shows the high degree of correlation between the projected percentage of reads for each library in its run, and the actual percentage.

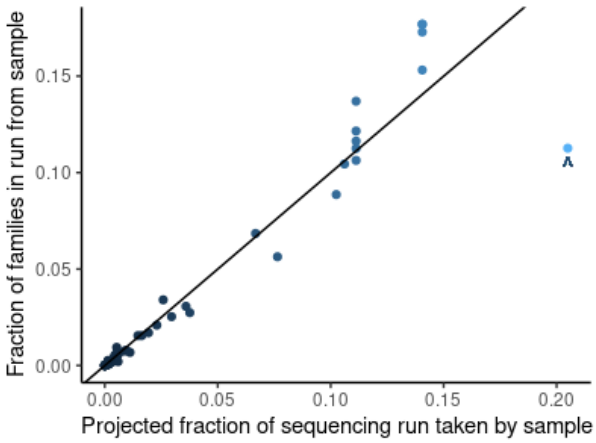


Figure 2.11: Pooling results following a MiSeq QC for the HC2 study.

Based on the amount of each sample's DNA added to each sequencing pool, a projected percentage of reads which should belong to the sample was calculated. The actual reads associated with each sample were divided by the total number of reads in the sequencing run to give the actual pool percentage. The black line with equation (y = x) represents theoretical perfect normalisation and sample pooling. A "^" marks a single sample where the actual read percentage deviated from the projected read percentage substantially. Including the anomalous value, a Pearson's correlation coefficient of 0.96 was achieved using this pooling method.

2.3.5 A united workflow for the barcoded sequencing of multiple sample types

The novel components discussed so far in this chapter were combined with off-the-shelf components, to create a united workflow. The overall workflow changed between the three studies, and this section details how the differences between the versions of the workflow changed its results.

2.3.5.1 The preliminary study

The libraries in this study were normalised and pooled. The pool was diluted to 1.7pM and sequenced on an Illumina NextSeq 500. This resulted in slight over-clustering with 260,000 per mm^2 as opposed to the recommended target of 210,000 per mm^2. Based on

the specifications of the NextSeq 500 on High Output Mode, it was predicted that each sample would produce 44 million reads.[237] This run produced a mean of 39.0 million ±6.9 million reads per sample due to the over-clustering, as shown in Figure 2.9d. The over-clustering resulted in an average 12% loss of reads, compared to the predictions.

2.3.5.2 The HC1 workflow

The assessment of the effectiveness of this workflow occurred in parallel with the creation of an in-house pipeline, which is presented in Chapter 3. Briefly, this pipeline took raw sequencing data, aligned the reads to the genome, and collapsed the reads down into read families, based on their barcodes and alignment position. These families would remove PCR duplicates from the data, and represent the original input DNA molecules better than the raw reads. The raw read depths for each sample are available in Table E.1. Once the reads were collapsed, the mean family depth from each sample of the HC1 cohort was compared to the number of genome copies used as input to library preparation. The data in Figure 2.12 shows that the wet-lab workflow was unusually efficient at taking molecules from low-input samples through to sequencing. Sample HC1-10, the Medulloblastoma sample, was excluded from Figure 2.12 because it had a mean family depth per genome copy of 275%. This value was clearly artefactual, and likely to be due to an underestimation of the amount of DNA in the original input.

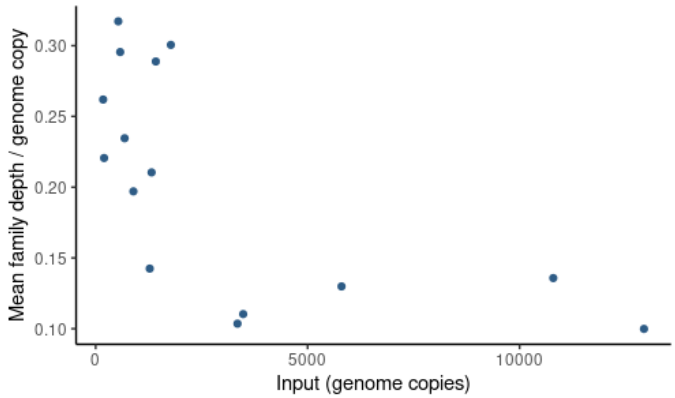


Figure 2.12: How the efficiency of the wet-lab workflow varies depending on the amount of input DNA

Genome copies were calculated from Bioanalyzer quantifications of the input DNA. Mean family depths were the mean across all targeted regions of the panel.

2.3.5.3 The HC2 workflow

In order to ensure that most molecules in a sample were sequenced enough to form families, the optimistic assumption was made that 30% of library preparation input molecules would be sequenced. A lower efficiency than assumed in the pipeline would mean that each molecule was be sequenced more times than was optimal. The raw sequencing read depths are available in Table E.2. As shown in Figure 2.12 from the HC1 phase of this study, for inputs of over 4000 copies (~14ng), an efficiency of 10-15% was more likely. At this point in the project, it was suspected that the unusually high efficiency of the low input data was due to underestimation of the DNA inputs during the initial Bioanalyzer quantification, or that these were artefacts from over-sequencing of the samples. The aim now was to assess the ddPCR quantification, used for this workflow version, and the efficiency of the workflow as a whole. Each sample's data was run through the Cerberus pipeline, described in Chapter 4, up to the point where the reads were collapsed into read families. The following data is based on the families made using Connor with the stringent settings associated with SNV and CNV calling.

Fifty seven of the inputs to library preparation contained detectable levels of DNA. One MRT sample failed, leaving fifty six samples. The number of families generated by Connor for each sample was divided by the number of haploid genome copies added to the library preparation. These efficiencies were plotted against the inputs in Figure 2.13. This new data shows that for inputs where the HC1 study suggested an efficiency above 15%, the efficiency was steady: irrespective of the input amount, 70% of samples had an efficiency of between 4.5% and 17%. The data only deviated from these bounds in samples with input amounts of less than 580 copies (2ng), where DNA quantification was likely to be less accurate. The three samples in the box represent the samples which produced unusually low library preparation yields. It follows that these samples also produced fewer families than was expected from their inputs.

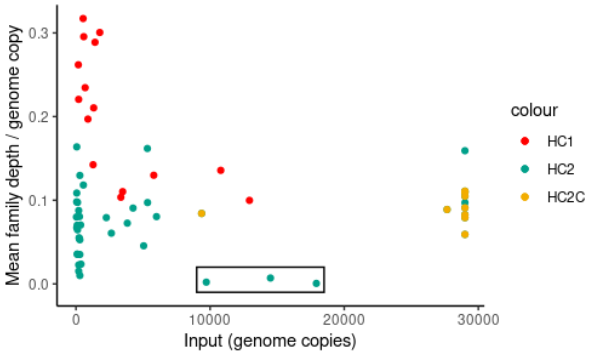


Figure 2.13: The efficiency of the wet-lab workflow and bioinformatics pipelines in taking molecules from extraction to sequencing, for the cohorts of the HC1 and HC2 studies.

The HC1 cohort, the HC2C cohort and the three samples in a box (samples HC2-16, HC2-18 and HC2-28) were quantified by Bioanalyzer. The three points in the box exhibited unusually low efficiency based on their input, so quantification was switched to ddPCR. The remaining points of the HC2 cohort, which were ddPCR quantified, show a smaller overall range in efficiencies at low inputs than those quantified by Bioanalyzer.

2.4 Discussion

Throughout the project, an iterative approach was taken in the development of the wet-lab workflow, with each study building on the knowledge gained from the previous study's results. The preliminary study focused on the assembly of components from library preparation onwards. The Version 1 workflow from the HC1 study added DNA extraction, and shearing; and it improved on the normalisation protocol. The Version 2 workflow from the HC2 study took normalisation from a flat process, and fully optimised the process for each individual sample.

2.4.1 Sample preparation methods

Sample preparation, before library preparation, developed mainly from the HC1 study onwards, as the preliminary study utilised extracted and sheared control DNA material. The sample shearing methods development phase of the HC1 study utilised control material processed by the author, which simulated the two main types of DNA in a CSF sample. The "sDNA" 182bp PCR fragment was used as a stand-in for short, cell-free DNA within the CSF, whose length was comparable to this fragment. Previous work showed that cell-free CSF DNA was approximately 100-200bp in length, so the control sDNA was of a length which represented its real-world counterpart.[71] The "lDNA" control, representing processed full-length cellular DNA, was made to mimic its real-world counterpart by being treated using the same methods as the clinical CSF samples were.

During the early stages of the shearing optimisation, 15µl MicroTUBEs were used, as the maximum input volume to the ThruPLEX Tag-seq kit was 10µl (Appendix C). Once the concurrent analysis of the preliminary project's results began to suggest that the raw DNA input amount could be a limiting factor of sensitivity, development was changed to utilise the MicroTUBE-50. These MicroTUBE-50s were capable of handling the entire eluate from a QIAamp extraction column, so the entire eluate was ready to be used for library preparation if necessary.

The final results of the shearing experiments (Figure 2.7) showed that the parameter set which was settled upon was capable of shearing the lDNA down to an acceptable length for sequencing, at a 23% loss of sDNA. This was a good compromise for the project, allowing for the total DNA in a sample to be sequenced. The nature of the cystic fluid samples which were included in both the HC1 and HC2C cohorts may have benefited particularly from this gentle shearing. Personal communication with the collaborators who extracted the fluid, as

well as visual inspection of the tubes, showed a high degree of cellularity in the samples, whilst no previous work had determined the amount of cfDNA in the sample.

The parameter set in Table 2.2 was the first set encountered, which performed adequately at shearing the lDNA, whilst preserving the sDNA. This does not preclude the existence of parameter sets which could do so more efficiently. In particular, the Covaris shearing protocols for the E220 Evolution sonicator recommended up to 100W incident power at 30% duty factor, meaning an average power of 30W.[238] The set used in this project was 15W incident power at 15% duty factor (2.25W average power). Future work could assess the effectiveness of increasing the duty factor and decreasing the treatment time, whilst avoiding excessive heat build-up in the sample. The minimisation of treatment time was not critical for this project, as a proof-of-concept, but the application of such shearing protocols in clinical settings where sample throughput is key, would make this an important avenue of development.

2.4.2 Hybrid capture - the FLCP-1 panel and its implementation

A main aim of this chapter was to develop a capture panel for genes and regions which was targeted towards a diverse range of PBTs. The use of prior work by Guichard *et al.*, and collaboration with Prof. Tom Jacques at GOSH made the creation of this panel straightforward.[216] The choice of the Agilent XT kit as a platform for this meant that the panel itself could be tailored to the samples which the author had access to, whilst enabling the panel to be expanded by the addition of extra biotinylated RNA baits to the capture panel.

The design of the panel made use of all coding exons from all splice variants in the RefSeq and Ensembl databases. This likely included many exons which were never expressed within the brain, making them unlikely to be useful when searching for SNVs or InDels. An interrogation of each splice variant within the panel could have been used to decrease the number of regions in the panel, but the 118kb panel was already narrow, and the extra captured regions could be used as normalisers during Copy-Number Variation (CNV) calling. As a result, the decision was made not to cull these during the preliminary study, as their usefulness outweighed the benefit of reduced sequencing.

The implementation of the panel in the workflow improved over time. Initial calculation of the on-target percentage using control material, during the HC1 study, resulted in an on-target percentage of 37.56%, which had major implications on the number of samples

which could be sequenced per run. The use of an improved magnetic rack, and the use of a recently calibrated thermocycler increased the on-target percentage to 49%, paving the way for larger numbers of samples per run in future studies. This improvement was built upon during the HC2 study. Although no major improvements to the capture procedure were undertaken between the HC1 and HC2 studies, more care was taken during the washing steps to keep the 0.2ml tubes at 65°C throughout, which could explain the improvement in on-target percentage. This is because the specificity of annealing between single-stranded oligonucleotides is dependent on temperature.[239, 240]

2.4.3 library quantification, normalisation, and QC

The normalisation and pooling of libraries prior to sequencing was achieved using a wide range of techniques at the UCL Cancer Institute, and personal communication with members of the Genomics Core facility suggested that optimisation of the methods was required for every wet-lab workflow. The library preparation kit was in its beta-testing phase at the beginning of the project, and there were no established pooling protocols for similar workflows, meaning that optimisation fell within the bounds of this project. The preliminary study's results demonstrated that a single Bioanalyzer quantification insufficiently normalised the libraries, producing widely variable numbers of reads per sample, demonstrating the necessity of improvement. The implementation of a two-step quantification procedure during the HC1 study was a major improvement, as shown in Figure 2.10. This was significant when combined with other factors. The more reads which collapse into a single family, the more reliable that family is, but with diminishing returns, as sequencing depth is increased. If a minimum number of reads per family is set, then the sequencing depth should be tuned such that the mean family size is slightly above that, so the vast majority of original template molecules which form libraries are sequenced to a sufficient depth to form families. In this case, the pooling accuracy added an uncertainty to the prediction of sequencing depth, and the main way to counteract this was to add more of each library to a pool, and to have fewer samples per pool. The high accuracy of pooling meant that the uncertainty factor was low, and the number of samples per sequencing run was able to be high.

The HC2 study took pooling to its logical conclusion, using the prior knowledge of the percentage of input molecules which formed amplified libraries, the capture efficiency, and the pooling accuracy. The highly accurate ddPCR quantification was the primary data

used to predict the number of template molecules' products were in the final libraries. This, combined with the advertised number of reads per sequencing kit, informed pooling, and produced the highly accurate results of Figure 2.11. The single anomaly in this figure was HC2C-1 - an unspun cystic fluid sample, quantified using ddPCR, with a 100ng input into library preparation. This sample was run in a batch alongside other samples whose predicted and actual family depths were close, so a partial failure of a reaction in the workflow was unlikely. The most likely explanation was a pipetting error during pooling, or during the first step of library preparation. Despite the anomaly, the strong correlation between the predicted and actual family depths showed that the method was robust, and reliable from sub-nanogram to 100ng library preparation inputs. During the normalisation of the HC2 libraries, some libraries required the accurate pipetting of volumes in the region of 2µl. Future use of this normalisation and pooling method would benefit from ensuring that the volumes were higher, for increased reliability.

2.4.4 The complete workflow

During the preliminary study, there was no literature on how efficient a workflow of this type was at converting input DNA molecules into double-stranded libraries, so one of the first objectives when creating the overall workflow was to estimate this efficiency. There were drastic changes to the workflow between the preliminary and HC1 studies, however, which may have rendered any efficiency value of the preliminary workflow irrelevant when predicting the Version 1 efficiency. Due to the prohibitive cost of running control material on the HiSeq, the decision was made to process the HC1 cohort before assessing the efficiency, instead of running the two control libraries generated at the beginning of the HC1 study. This went hand in hand with the ability to assess the accuracy of the improved pooling method, as discussed above, and the samples were over-sequenced.

This data showed a hockey stick-shaped curve of efficiency (Figure 2.12). The curve likely showed that at low concentrations, the initial quantification was less accurate, and under-represented the number of molecules in low input samples, leading to higher efficiency values. Further evidence to support this came in the form of the HC2 efficiencies, which were based on the ddPCR quantification (Figure 2.13). Through the region where the efficiency rose above 15% in the HC1 study, the HC2 samples' efficiencies varied more than at higher concentrations, but did not show the same rise in efficiency. This increased variation was expected, as the DNA concentrations approached the limit of accuracy of the

quantification method.

The overall efficiency of the workflow was tested on retrospective samples, in the HC1, HC2 and HC2C cohorts. These samples were often small in volume, and in DNA content, as evidenced by Figures 2.6 and 2.13. The ability of the final workflow to generate sequenceable libraries from disparate sample types, with variable inputs demonstrates the robustness of the methodologies and parameters which were optimised over the course of this project.

Chapter 3

Development of a prototype pipeline for the detection of Single Nucleotide Variants in barcoded sequencing data, and initial detection of copy-number variants from liquid biopsy DNA

3.1 Introduction

One of the main objectives of the overall project was to create a suitable data analysis pipeline for use with data produced by the wet-lab workflow. This chapter details the early work on bioinformatic pipeline development which occurred during the preliminary and HC1 studies, which brought the project closer to this goal. The work in this chapter was performed with the preliminary and HC1 cohorts as development datasets, processed using the preliminary and Version 1 wet-lab workflows respectively, as presented in Chapter 2. The work culminates in the detection of Single Nucleotide Variants (SNVs) in Adamantinomatous Craniopharyngioma (ACP) samples, the detection of Copy-Number Variation (CNV) in Atypical Teratoid/Rhabdoid Tumour (ATRT) samples, and the orthogonal validation of CNV variants by droplet digital Polymerase Chain Reaction (ddPCR).

At the time of the preliminary study, there was no bioinformatic solution for handling Tag-seq data, so testing of the of the Duplex Sequencing pipeline (v2.0) was performed to assess the suitability of the pipeline in the analysis of the data.[10] During the course of the study, Curio Genomics produced a web-based platform for the analysis of Tag-seq data,

and this was used to generate preliminary results for assessment of its suitability for the project's aims. The Curio platform was in its beta-testing stage, so the number of features available was limited, and the implementation of these features was done through Curio Genomics. The decision was made to develop an in-house pipeline, which was able to run on UCL's High Performance Computing (HPC) platforms, and was more easily modified by the author. The initial development of this pipeline, supporting work, and validations of the pipeline's first successful results are presented in this chapter.

3.2 Aims and objectives

Initially, the aim was to determine the suitability of any pre-existing pipelines and software packages, developed against other barcoding schemes, for SNV calling of Tag-seq data. When the Curio platform was released in beta-test form, it was added to the list of pre-existing software. Once it became clear that Curio, in its beta form, was sub-optimal, the main aim was to develop a prototype in-house pipeline for the detection of SNVs. Concurrently, the project aimed to investigate the detection of CNVs within the data, and to verify any findings using orthogonal methods.

3.3 Materials and methods

3.3.1 Determination of the suitability of existing pipelines and software packages

3.3.1.1 Initial analysis of sequencing data without utilising barcode information

When assessing the suitability of pipelines and packages for the analysis of Tag-seq data, a traditional analysis serves as a baseline against which the new method can be compared. An advantage of this kind of baseline is the ability to remove the barcodes and stem sequences from the raw reads of a barcoded dataset, to create an unbarcoded dataset.

Paired-end FASTQ files, produced by a Next-Seq 500 at the UCL Genomics facility, were downloaded from Illumina's BaseSpace platform to UCL's Legion HPC cluster. The data for each sample was in 4 separate file pairs, one pair for each lane of the NextSeq flow cell. A script was developed to concatenate each set of Read 1 files and Read 2 files for each sample into a single file pair, such that a read on line (x) of the output Read 1 FASTQ file was still paired with the read on line (x) of the Read 2 file. Copies of these concatenated files were made, for use in barcoded analysis, as described later in this chapter. The barcodes and stem sequences were trimmed from the 5' end of each read using Trim Galore. This produced a raw set of FASTQ files which were ready for non-barcoded analysis.

The analysis of the datasets began with alignment against the GRCh38 version of the human genome using BWA mem 0.7.2.[241] Samtools 1.2 was used to compress the output, then to remove duplicate bit flags in the BAM files, and to sort and index the files.[242] Picard MarkDuplicates was used to remove duplicate reads.[243] Samtools mpileup was used to make an mpileup file for regions within Forshew Lab Capture Panel 1 (FLCP-1), and Varscan 2.3.9 was used to detect SNVs relative to the GRCh38 reference.[244] Filter-based variant annotation was performed using Annovar with the settings outlined in Table 3.1.[245] There were two competing factors which led to the 4.7% Variant Allele Frequency (VAF) used to filter variants. The first was the number of variants detected at the advertised positions in Samples 3 to 8, the pairs of Horizon Discovery cfDNA control samples with advertised variants at 5%, 1% and 0.1% respectively. The second was the total number of variants in Samples 1 and 2, the cfDNA control samples which contained no advertised variants. This VAF was chosen as a compromise, minimising the former whilst maximising the latter, and was in line with standard variant calling detection thresholds in Illumina DNA sequencing.[246]

Table 3.1: Parameters used for Annovar filter-based annotation of variants

Parameter	Value
Genome build version	GRCh38
Variant allele frequency	Minimum 4.7%
Refseq Gene variant exonic function[220]	Exclude synonymous SNVs
HRC R1 database allele frequency[247]	Exclude above 0.1%
1000 Genomes August 2015 database allele frequency[248]	Exclude above 0.1%
Kaviar 23 September 2015 database allele frequency[249]	Exclude above 0.1%
ESP 6500siv2 database allele frequency [250]	Exclude above 0.1%
ExAC 03 database allele frequency[251]	Exclude above 0.1%

3.3.1.2 Testing the Duplex Sequencing Pipeline on ThruPLEX Tag-seq data

During the early stages of the preliminary study, there was no off-the-shelf pipeline which was designed to process Tag-seq data. Before investing development time into the creation of a custom pipeline, the Duplex Sequencing pipeline was assessed for its ability to process Tag-seq data. The concatenated raw NextSeq files from the non-barcoded analysis were saved before barcode removal.

On the Legion HPC cluster at UCL, version 2.1 of the Duplex Sequencing pipeline was downloaded from `https://github.com/loeblab/Duplex-Sequencing`.[10] All of the scripts which form the Duplex pipeline are executed by a master Bash script in the default. In order to run on UCL's Legion cluster and the Sun Grid Engine, the Bash script was split into five parts and a launcher script was written for each part. It was discovered that the single stranded consensus-maker from the Duplex Sequencing pipeline was not able to run with sequences of a different length to a stated constant, so any reads which were shorter than 151 bases were padded to the required length with "N" sequence characters and "#" Sanger quality characters.

Part 1 consisted of a launcher Perl script which searched for sets of paired-end FASTQ files based on filename, and submitted a qsub script to the cluster's job queue. The qsub script concatenated the 6nt barcodes for both paired-end reads and stored them in the FASTQ header for each read. An 11nt sequence immediately 3' of the barcode, containing the 8-11nt stem, was removed from each read. Quality scores were modified similarly to the sequence strings.

Parts 2 to 5 consisted of a related set of launcher Perl scripts and four different qsub

scripts which performed the data processing on the cluster's nodes. The launcher scripts checked in log files and the cluster queue. If a previous part was running for a given sample, the launcher waited for the previous process to finish before continuing. If the preceding part of the pipeline exited with errors for a given sample, if input files for the current part were missing, if the current part was already running, or if a subsequent part in the pipeline had been started, the launcher would not allow the current part to start. Once all checks were passed, the current part was submitted to the cluster for the given sample.

Part 2 ran BWA aln (version 0.7.12) on each processed FASTQ file to generate an aln file.[252] Part 3 took the processed FASTQ files and the two aln files, and processed them into a paired-end SAM file using BWA sampe. In part 4, Samtools view and sort (version 1.2) were used to convert the SAM file into a BAM file in which records were sorted by genomic location.[242] Part 5 launched the single-stranded consensus maker with the sorted BAM file and created a paired-end BAM file which contained consensus sequences from reads with the same genomic starting location and barcode string.

3.3.1.3 Testing beta release of the Curio platform on preliminary Tag-seq data

During the investigation of existing technologies, Curio Genomics developed a cloud-based platform for the processing of barcoded data which was explicitly compatible with Thru-PLEX Tag-seq data. This platform was released in 'beta', and the author joined the program.

The concatenated paired-end FASTQ files which were used in the unbarcoded analysis were uploaded to the Curio platform through their web portal. The platform was used to generate alignments for each FASTQ file pair using the Bowtie 2 aligner and the parameters in Table 3.2.[253] At the time, the hg19 assembly was the only one available on the platform, so this was used with the intention of lifting the variants over to GRCh38 at the end of analysis. Reads with an alignment quality of lower than 99% were removed, and individual bases with an alignment quality below the threshold were changed to N's.

Allele frequencies were calculated next. Barcodes, known as Unique Molecular Tags on the platform, were used to group reads into read families. At a given position on the genome, each family which overlapped with the position was checked for consensus at the position. If the percentage identity of the sequences comprising the family was more than 99% and there were 3 or more reads in the family, the consensus base was used in allele frequencies at the position.

Table 3.2: Parameters used for alignment using Bowtie 2 on the Curio Genomics NGS analysis platform

Parameter	Value
Alignment mode	Local
Aligner	Bowtie 2
Genome assembly	hg19
Minimum read length	100
Multiseed heuristic preset	Very sensitive
Trim 3' ends with Phred quality below	28
Barcode length	6
Maximum stem sequence length between barcode and read	11
Extended mates	Detect as discordant
Discordant alignments	Suppress alignments
Unpaired alignments	Allow alignments

At the time of this test, variant calling had not yet been implemented on the Curio platform, so the BAM files of collapsed read families were downloaded. The original sample material, for all samples in the preliminary cohort, were comprised of DNA from multiple cell lines, mixed to give certain VAFs at specific genomic locations.[232] Outside of these defined positions, there were unadvertised variants, which were a product of the same mixing of cell line DNA, which could not be verified. As a result, variant calling was only advisable at the advertised positions. Samtools was used to index the BAM files for ease of access, and VAFs were manually read from the data using the Integrated Genomics Viewer (IGV).[254]

3.3.2 A prototype pipeline for detection of Single Nucleotide Variants

Following the results of the preliminary study and the lack of features available in the Curio platform in its 'beta testing' form, the decision was made to develop an in-house pipeline, using open-source tools. This would give the author the flexibility to develop the pipeline to suit the data, and to add components which highlighted or removed artefacts.

The main stages of the prototype pipeline are summarised in Figure 3.1. Compressed FASTQ files of data from the HC1 cohort were downloaded from Illumina's Basespace platform onto the UCL Legion HPC Cluster. The data for each sample was in two pairs of files, one pair from each lane on the HiSeq Rapid flow cell, and these were combined into a single pair of files using the same custom script developed for the preliminary study. Each pair of files, containing the forward and reverse reads for a sample respectively, was subjected to Quality Check/Control (QC) using FastQC, and Trim Galore was used to remove barcode and stem sequences from the 3' end of sequences where read-through occurred. The reads were aligned to the GRCh38 version of the human genome using BWA mem, and the resulting BAM files were sorted, filtered to remove unaligned/poorly aligned reads, and indexed by Samtools.

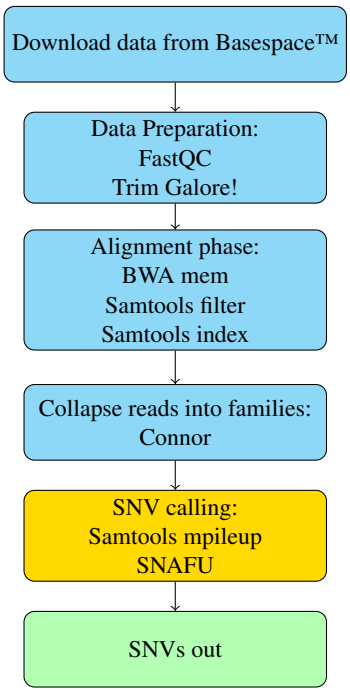


Figure 3.1: The prototype Single Nucleotide Variant calling pipeline.
The main stages of the prototype pipeline, and the principle software packages and commands used at each stage. The gold box indicates the variant calling phase.

3.3.2.1 Collapsing reads into families using Connor

During the evaluation of the Curio platform, an Connor, an open-source tool for the collapsing of reads into families, was developed. This tool, available at https://github.com/umich-brcf-bioinf/Connor, used a method which was reportedly compatible with the Tag-seq barcoding scheme, since Connor collapsed reads based purely on barcode and alignment start position.[129, 130] Groups of read pairs with the same alignment start sites on the genome for both reads, and barcode pair (as described in Figure 2.5), were collapsed into a single family. Due to the possibility of sequencing errors when sequencing the barcode, a degree of promiscuity in barcode sequence was necessary. Hamming Distance (HD) is the minimum number of substitutions necessary to convert one string into another string of equal length. For example: the string "AATAAT" can be converted into "AAAAAC" with two substitutions, so the first string has a HD from the second of two. When collapsing reads into families, all barcode pairs with a HD of one or zero from each other were collapsed together, and a consensus sequence was made from these reads. In order to minimise the prevalence of PCR or sequencing errors in the collapsed output, a consensus threshold of 0.8 was used. Any position in a family where fewer than 80% of the constituent reads agreed on a base call were replaced by N's in the consensus sequence. Additionally, a family was only written to the output if it was made of a minimum of three reads.

3.3.2.2 Early Single Nucleotide Variant calling using SNAFU v1.0

Artefacts discovered by visualisation of Connor-collapsed BAM files, described in the Results section, necessitated the development of a custom variant caller which was tolerant to these artefacts. This new variant caller, called Single Nucleotide Alteration Filtering Utility (SNAFU), took the family counts at each position of the mpileup file to find putative variants, and filtered them based on the families which overlapped the position. The logic for SNAFU v1.0 is displayed in Figure 3.2.

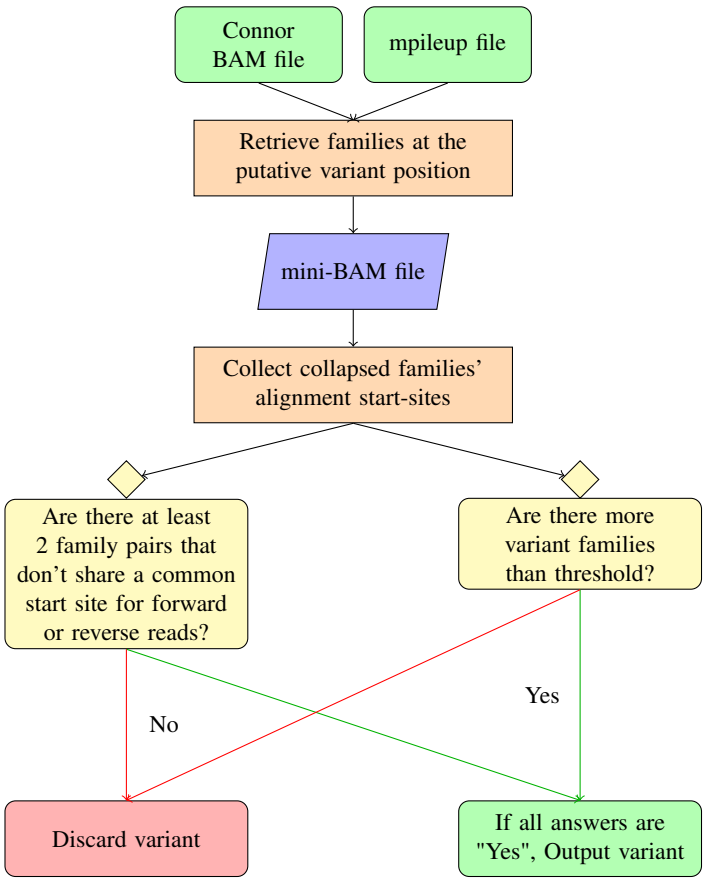


Figure 3.2: The logic which underpins SNAFU v1.0.
A flow chart showing the decisions and processes which SNAFU v1.0 uses to call SNVs.

3.3.3 Investigation of Copy-Number Variation in the HC1 cohort

3.3.3.1 Manual CNV detection

Initial CNV detection was performed manually. The collapsed read families from Section 3.3.2.1 were used to calculate the mean family depth for the targeted positions in the FLCP-1 panel, for each sample. The mean family depth for all positions in the FLCP-1-targeted

regions of *SMARCB1* were also calculated, and divided by the mean family depth for all targeted regions. This gave an internally normalised family depth for *SMARCB1* versus the rest of the panel.

The following describes the detection of CNVs at the exon level. The mean family depth for each exon of *SMARCB1* (e) was calculated, and divided by the mean family depth across the whole FLCP-1 panel, to give the normalised exon depth (e_n). Control samples were selected from tumour types which do not show recurrent CNVs in *SMARCB1*, and the e_n was calculated for each exon in each control sample. Exons 1 and 9 were excluded from Exon-level analysis, as the low and variable e_n in the control samples made the detection of a CNV loss in these exons unreliable. The mean and the standard deviation, for Exons 2-8, of all control samples' e_n was calculated. The e_n value in each exon, of each tumour sample with a mean family depth over 100, was compared to the mean e_n of the controls.

3.3.4 Droplet Digital PCR verification of CNV results

The exon level CNV results yielded a putative loss of *SMARCB1* Exon 5. In order to verify this, a custom ddPCR assay was designed for this region, the design parameters for which are available in Table 3.3. The primers and probe were manufactured by Integrated DNA Technologies, and dissolved in nuclease-free water. An assay mixture was created, containing each primer at 18µM, and the probe at 5µM.

Table 3.3: The design details of a custom ddPCR CNV assay for *SMARCB1* Exon 5

Property	Value
Gene	*SMARCB1*
Amplicon Length	72
Forward primer sequence	AGCTGTGATCCATGAGAACG
Reverse primer sequence	CCATCGATCTCCATGTCCAG
Probe sequence	ATCTCAGCCCGAGGTGCT
Probe modifications	5'-FAM, 3'-BHQ1
Temperature	60°C

Two assay pairs were used for the verification of CNVs. Assay Pair 1 was a combination of the assay in Table 3.3, and a CNV normalisation assay targeted at the RPP30 gene, made by Bio-Rad (Assay ID: dHsaCP2500350). Assay Pair 2 was made up of the

same RPP30 assay, and a Bio-Rad CNV assay targeted at *SMARCB1* Exon 4 (Assay ID: dHsaCP1000520).

The control material for this assay was Bioline Human Genomic DNA, sheared using the parameters in Table 2.2. The control ddPCR reactions were run in quadruplicate, and each assay pair was run on sheared tumour DNA in duplicate.

The ddPCR reactions were set up according to Table 3.4, and were then placed in a Bio-Rad Automated Droplet Generator, in which each reaction was partitioned into the droplets of a water-in-oil emulsion. The emulsions were foil-sealed in a 96 well plate, and Polymerase Chain Reaction (PCR) cycled under the conditions outlined in the Supermix manual.[255] After PCR, the droplets from each emulsion were read by a Bio-Rad QX200 Droplet Reader, and the Copy-Number Variation status was analysed automatically by the QuantaSoft analysis software.

Table 3.4: The generalised reaction mixture for the Droplet Digital PCR testing of sheared Cerebrospinal Fluid samples for Copy-Number Variation

Reagent	Volume (µl)
2X ddPCR Supermix for probes	10
20X FAM assay premix	1
20X HEX assay premix	1
Sample	8
Water	0

3.4 Results

3.4.1 The preliminary and HC1 cohorts

The cohorts and samples used in the preliminary and HC1 studies were previously described in detail, in Section 2.3.1. Briefly, the preliminary cohort was made up of eight samples: two each of the four Horizon Discovery cfDNA Reference Standards, which had defined variants at 0%, 5%, 1% and 0.1% respectively. The HC1 cohort was a retrospective group of eighteen ATRT, Diffuse Intrinsic Pontine Glioma (DIPG), Medulloblastoma, and ACP samples, sixteen of which were sequenced. These samples were a mix of cerebrospinal fluid (CSF), plasma, and cystic fluid samples.

3.4.2 Assessing the suitability of existing analysis methods for Tag-seq data

The most effective strategy to achieve the aims of the project was to use as many existing technologies as possible, provided that these technologies were suitable for the wet-lab methodology, and to focus development time on the novel aspects of the data analysis. To this end, a baseline non-barcoded analysis was performed, and the Duplex Sequencing pipeline and the Curio platform's 'beta' release were trialled.

3.4.2.1 Analysis of preliminary Tag-seq data without barcodes

To create a baseline to compare further analysis against, analysis of the Tag-seq data was performed without barcode information. Barcodes were removed from the reads in the preliminary FASTQ files, and the reads were aligned to the GRCh38 reference genome. Duplicate reads were removed, variants were called within the regions covered by FLCP-1, and variant alleles were annotated.

There were large numbers of variant alleles called which were not advertised. These VAFs varied between 4.9% and 74%, and each variant allele was not stable between the Reference Standards. Communication by lab members with Horizon Discovery revealed that these alleles were real and were a result of the mixing of cell line DNA during production. As a result, only the six variants advertised by Horizon Discovery were considered, and the Reference Standards were used for sensitivity measurement only. Figure 3.3 shows that in samples 3 and 4, with an advertised 5% VAF, eleven of twelve possible variants were detected using this method. When the detection threshold was dropped to 1%, the total number of variants below 30% increased from a mean of 85 $\pm$6 across all samples to 138 $\pm$25. The decision was made to keep the detection threshold at 4.7%, since this was in line with traditional Illumina detection thresholds, and many of the extra variants at the

lower threshold were likely to be false-positives.[246]

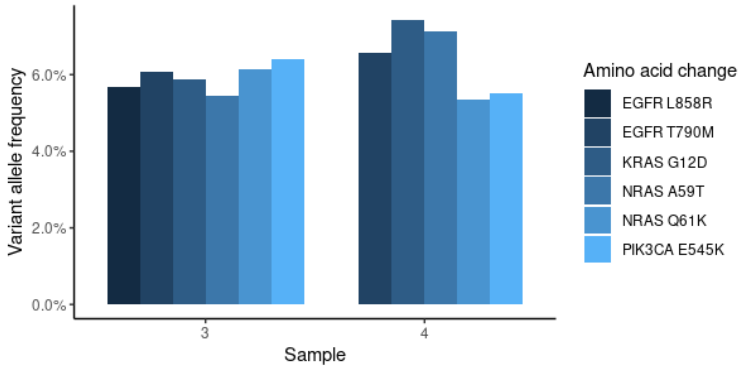


Figure 3.3: Detected variant allele frequencies from variant calling without barcoding information.

Samples 3 and 4 were from the Horizon Discovery 5% cfDNA control in the preliminary cohort.

3.4.2.2 Testing of the Duplex sequencing pipeline with preliminary Tag-seq data

NextSeq data came in the form of 4 pairs of gzipped FASTQ files for each library, one pair for each lane on the sequencer. A Perl script was written and this successfully converted the 4 pairs of files into a single pair of fastq files. These inputs were used in a version of the Duplex sequencing pipeline which was designed to run on UCL's Legion cluster.

The output files generated by the single-stranded consensus maker consisted of strings of 151 N's for approximately one third of their families. An inspection of the Duplex pipeline's Python scripts revealed that the two consensus makers would need to be rewritten to handle the header information from Tag-seq reads, and that the current version of the Duplex sequencing pipeline was not compatible with Tag-seq data.

A complete rewrite of the Duplex Sequencing pipeline had been performed by Kennedy and Kohrn, which moved the construction of the consensus to a point before alignment.[131, 256] The barcodes for Duplex Sequencing were 12nt $\times$ 2 long, so there were theoretically 2.8×10^{14} possible barcode combinations, whilst Tag-seq's 6nt $\times$ 2 barcodes resulted in only 1.7×10^{7} combinations. It was feasible to collapse Duplex data without utilising the positional information inherent to aligned data. This was not, however, possible for Tag-seq data, as too many separate families would have the same barcode pairs, so separating families by their genomic position was necessary for accurate consensus generation. The revised pipeline was not considered for the trials for this reason.

During the assessment of the initial Duplex Sequencing pipeline, the Curio platform

had been developed for the analysis of Tag-seq data and the platform was released for beta testing, so analysis was moved over to this system.

3.4.2.3 Analysis of preliminary Tag-seq data using the Curio platform

The Curio platform, a cloud-based system for the analysis of barcoded sequencing data, was released into 'beta'; this platform was designed to handle Tag-seq data.[257] The concatenated FASTQ files from samples 5 to 8, prepared from the 1% and 0.1% cfDNA controls, were uploaded to the Curio platform, and analysed using the tools available at the time. Due to space constraints on the platform, the plan was to upload the data in two batches, with samples 5 to 8 in batch 1, and samples 1 to 4 in batch 2. This was because the ability to detect variants at 1% and 0.1% were the focus of the project. The results in the following section demonstrated that development time was better used to develop an in-house pipeline, and batch 2 was not analysed on Curio. Original data for the following graphs is available in Appendix A. The six advertised substitutions per sample were used to probe the limits of sensitivity of the overall workflow, in a similar fashion to the barcode-free analysis presented previously. Figure 3.4a shows that five of the six substitutions were, on average, covered to sufficient depth for barcoding to be of use, but *PIK3CA* was not likely to contain sufficient families for rare allele detection in these samples.

Broadly, when barcode processing was not applied to the data (Figures 3.4b, 3.4d and 3.4f), a low level of noise pervaded fourteen of the fifteen sites where reads were present, as represented by the blue bars. This was almost entirely removed by barcode processing, leaving one false-positive family in all three samples. This coincided with an absence of true variant families at two of the original fifteen sites (Figures 3.4c, 3.4e and 3.4g).

Reads from samples 5 and 6 (1% cfDNA Reference Standard) clearly showed that sensitivity at 1% was not an issue, even in the absence of barcoding information (Figures 3.4b and 3.4d). After the barcode processing, the low level noise from the other possible minor alleles (blue bars) was almost completely eradicated (Figures 3.4c and 3.4e).

At a 0.1% VAF, sensitivity was promising, but showed that there were challenges to be overcome (Figures 3.4f-3.4g). Low numbers of variant reads hampered the formation of sufficient families for reliable detection of variant alleles at some of the advertised positions. Sample 7 contained a maximum of 4 reads for any given variant allele, and no families of more than 3 members were formed, so data is not shown. This showed that a higher sequencing depth is necessary before the sensitivity is at an acceptable level for detection at

a 0.1% VAF.

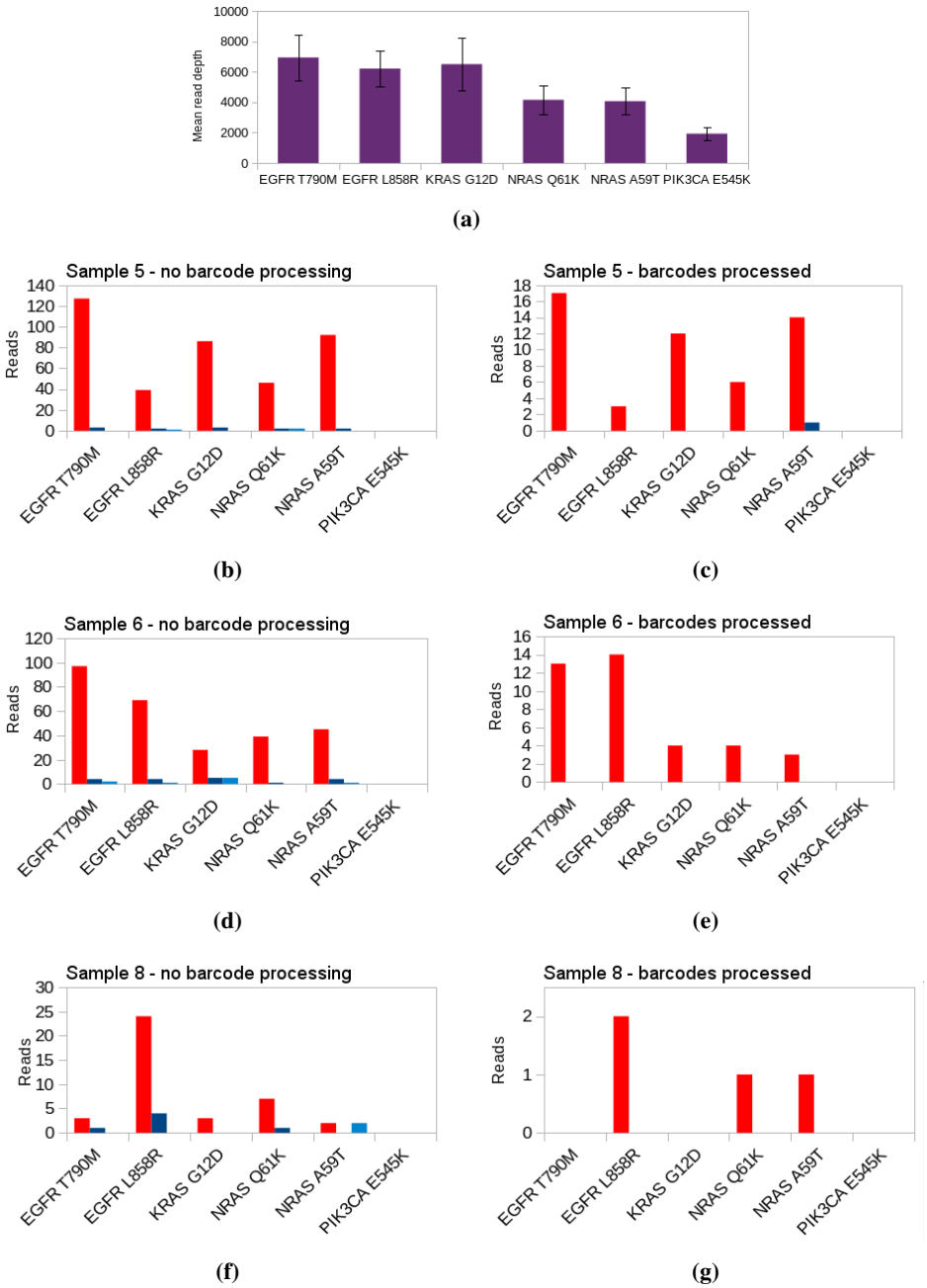


Figure 3.4: Variant allele detection in Horizon Discovery cfDNA controls

3.4a) Mean read depth at each Horizon Discovery cfDNA advertised substitution site for samples 5, 6, 7 and 8. **3.4b to 3.4g)** Variant allele read/read family depths at each advertised substitution site for samples 5, 6 and 8. The red bars represent the advertised minor allele, and the two blue bars represent the two other possible minor alleles.

3.4.3 Development of a prototype pipeline for SNV detection, and its application to the HC1 cohort

The overall results from the Curio platform were promising, especially given the preliminary study's use of its beta release. The pace of development and the customisability of the parameters required for this project, however, did not line up with the schedule that Curio hoped to achieve within the time-frame of the project. As a result, the decision was made to move to an open-source pipeline for the HC1 study.

With the knowledge that a 37% on-target rate for the hybrid capture limited the number of samples which could be sequenced per run on the HiSeq, a small cohort of CSF, cystic fluid, and plasma samples was selected. This cohort, named HC1, would provide the training data for the development of the in-house pipeline. Libraries were prepared from these samples, and sequenced.

3.4.3.1 Selecting an appropriate Hamming Distance for the Connor package

Once the sequencing data was concatenated, pre-processed, and aligned to the genome, the first novel step in optimisation was to determine the optimal Hamming Distance (HD) (described in 3.3.2.1) for Connor's collapsing of reads into families. A high HD would account for PCR and sequencing errors in the barcode strings, as reads with a sequencing error in the barcode would be grouped with reads whose barcodes were sequenced correctly. An excessively high HD would, however, decrease the number of families in the data, by collapsing reads from different original molecules into the same family. This would have a negative effect on the pipeline's ability to detect rare variants, as any real rare variant family was very likely to be collapsed with a family with no variant. This would in turn reduce the likelihood of the variant base appearing in the family's consensus sequence, reducing sensitivity.

Before the HD value for further analysis was set, an early version of Figure 2.12 showed that some samples had unusually high family depths for their inputs. At the time, the source of these inflated family depths was unknown, and this may have had an effect on the results from which a hamming distance was chosen. Samples with inflated mean family depths were therefore excluded from the optimisation of the Connor software's HD parameter.

Connor was set to HDs of 0, 1 and 2, and the mean family depths for each sample are displayed in Figure 3.5a. Between HDs of 0 and 1, there was a 0.07% ±0.16% loss of mean family depth. Between HDs of 1 and 2, there was a correlation between mean family depth and the mean family depth lost between a HDs of 1 and 2 (Figure 3.5b). This correlation suggested that, with an HD of 2, the collapse of multiple molecules' families into a single family outweighed the collapse of artefactual families into single families. Connor's HD parameter was set to 1 for all future analysis, which produced the mean family depths displayed in Figure 3.5c.

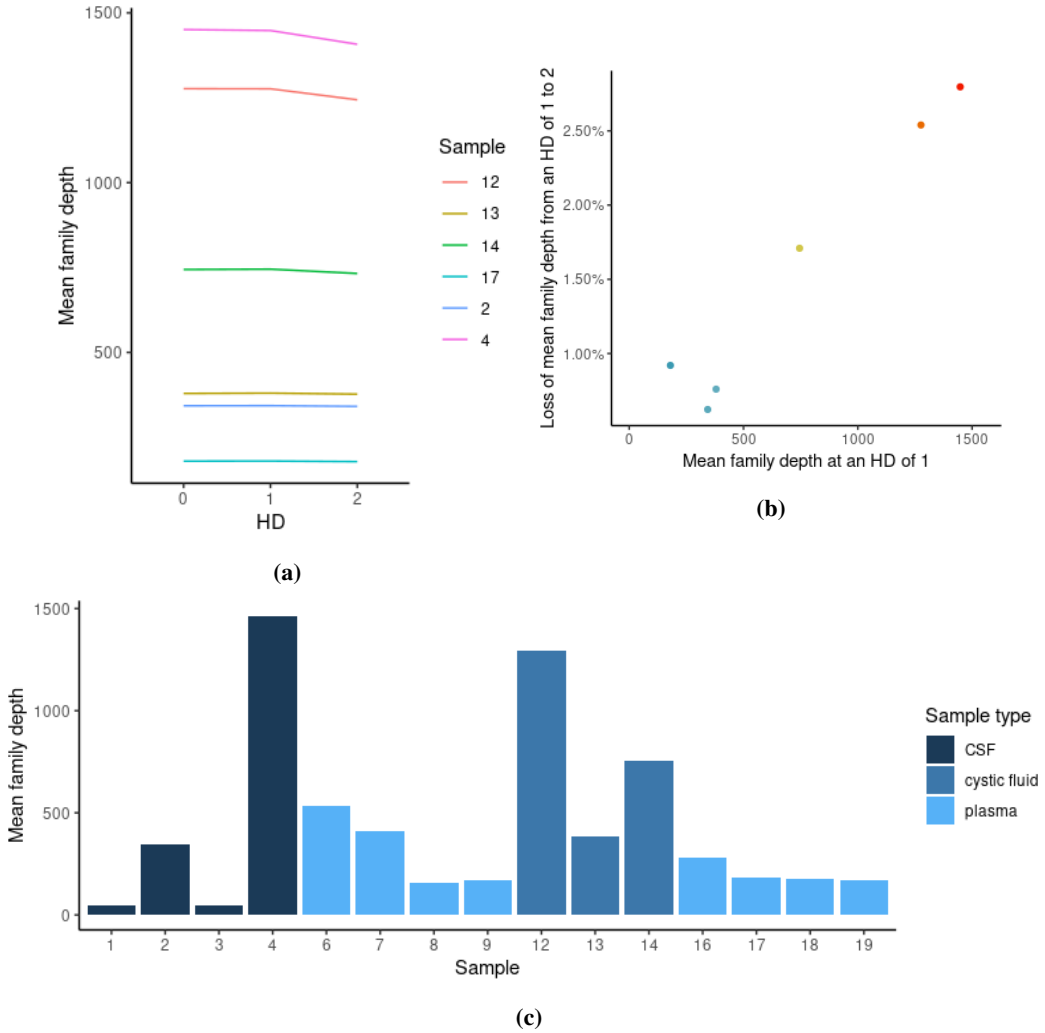
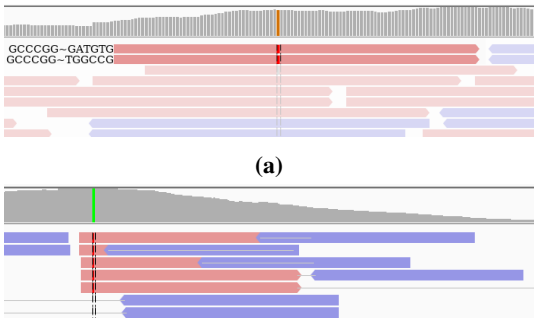


Figure 3.5: The effects of changing Connor's Hamming Distance parameter on the numbers of families produced by the software package, and the resulting mean family depths.

3.5a) The mean family depth, across the FLCP-1 panel, at different HD values. **3.5b)** The correlation between mean family depth and the loss of mean family depth between HD values of 1 and 2. **3.5c)** The resulting mean family depths for each sample of the HC1 cohort.

3.4.3.2 The development and initial trials of the SNAFU variant caller

Once the HC1 data had been collapsed into families using an appropriate HD, the collapsed data was viewed in IGV. This showed that the molecular barcoding was able to suppress the noise greatly, but that there were artefacts within the data. Examples of these artefacts, extracted from the Integrated Genomics Viewer, are displayed in Figure 3.6.



(a)

(b)

Figure 3.6: Artefacts in the sequencing data produced by the Connor deduplicator

3.6a) Two families with the same start site, and the same barcode, had different barcodes on their mates. **3.6b)** All variant families had different barcode pairs and different mate start sites, but all families had the same start site.

Single Nucleotide Alteration Filtering Utility (SNAFU) was created to filter these artefacts out of the data whilst calling SNVs. Due to low family depths in the plasma samples and most of the CSF samples, no variants passed filter for these samples. The results of variant calling on the cystic fluid samples was more positive, showing that large amounts of variant DNA were present in this fluid (Table 3.5). Additionally, all of the SNVs detected in cystic fluid were well known variants associated with ACP, commonly mutated in multiple cancers.[177, 178, 258]

Table 3.5: *CTNNB1* Exon 3 SNVs which were called by the early version of SNAFU.

Sample	GRCh38 position	Ref.	SNV	Family depth	VAF	Transcript and protein change (NM_001904)
HC1-12	chr3:41224606	G	A	611	7.9%	c.G94A:p.D32N
HC1-13	chr3:41224622	C	T	224	17%	C110T:p.S37F
HC1-14	chr3:41224613	G	A	376	9.0%	G101A:p.G34E

3.4.4 Initial detection of Copy-Number aberrations in Cerebrospinal Fluid samples

At the time of the work in this section, there were no published methods for the detection of CNVs in barcoded targeted sequencing data. The work in this section was performed manually, to assess the possibility of CNV detection within these samples.

The samples under scrutiny in this section were the ATRT samples and the Medulloblastoma sample from the HC1 cohort. This is because the vast majority of ATRTs exclusively show aberrations in a single gene: *SMARCB1*; Medulloblastomas can have alterations in the SWI/SNF complex, of which the *SMARCB1* protein is a member.[259] Control samples for CNV calling were made up of a mixture of the cystic fluid samples and DIPG plasma samples from the HC1 cohort, since *SMARCB1* alterations are not found recurrently in either of these tumour types, and ACP does not have any recurrent CNVs.[178, 198, 258]

For each sample, the mean family depth at targeted positions within *SMARCB1* was calculated, and normalised against the mean family depth across the whole FLCP-1 panel, to give a relative copy-number. Figure 3.7 shows that there were possible CNVs in samples HC1-10 and HC1-4. The low mean family depths in samples HC1-10, HC1-1 and HC1-3, however, meant that the CNV calls for these samples were of lower confidence than the other samples.

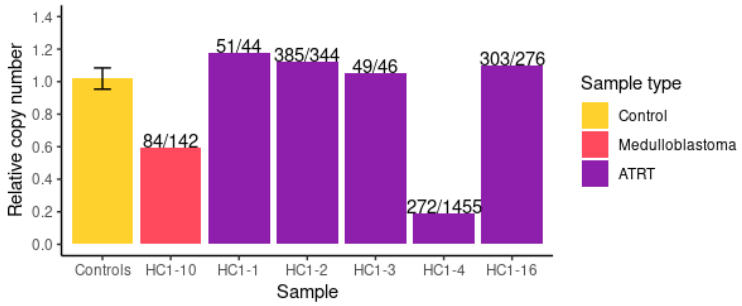


Figure 3.7: The gene-level results of manual Copy-Number Variation detection in Cerebrospinal Fluid samples

Two samples show large deviations from the mean of the control material: one Medulloblastoma, and one ATRT. The numbers above each non-control bar are the mean family depth across the *SMARCB1* targeted regions, and the mean family depth across all FLCP-1 targeted regions. The error bar on the control column represents the standard deviation of the control samples' relative copy-numbers.

The process was repeated, treating each exon of *SMARCB1* separately. The control samples had low relative copy-numbers for Exons 1 and 9, of 0.30 and 0.24 respectively. The possibility of artefactually low values in the controls led to these exons being excluded from the analysis. The results for Exons 2-8, available in Figure 3.8, show that the gene-level results from Samples HC1-10 and HC1-4 were also present at the exon level. Additionally, Sample HC1-2 showed a marked drop in relative copy-number in Exon 5 alone, suggesting a single exon deletion was present here.

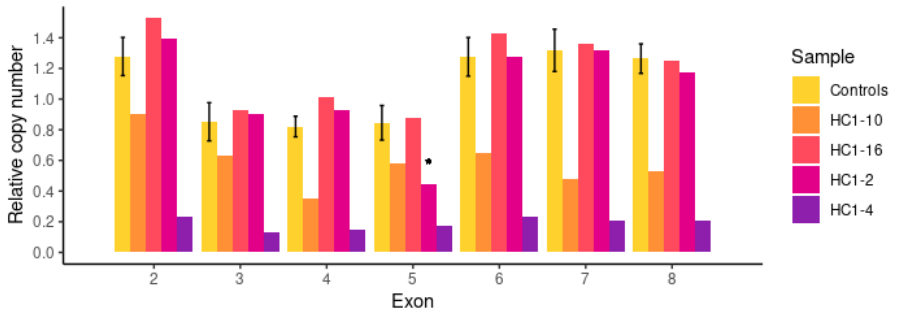


Figure 3.8: The exon level relative haploid copy-number values for the ATRT and Medulloblastoma samples in the HC1 cohort.

The error bars on the control columns represent the standard deviation of the control samples represented by these columns. The asterisk highlights a low Exon 5 relative copy-number for Sample HC1-2.

Based on these two analyses, three putative *SMARCB1* CNV calls were made: a low confidence single copy deletion in HC1-10, a biallelic whole gene deletion in HC1-4, and an Exon 5 deletion in HC1-2. The deletion in Sample HC1-2 was manually viewed in IGV, and is displayed in Figure 3.9. These variants were of great interest, however, they required validation before the continuation of the project. This was especially important, as bounds separating high confidence and low confidence CNV calls could not be established by such a small number of samples.

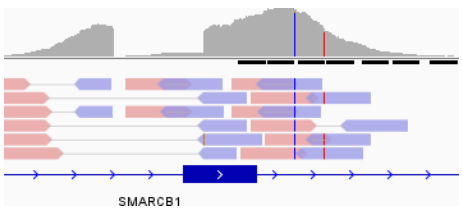


Figure 3.9: Manual viewing of the *SMARCB1* Exon 5 deletion in Sample HC1-2.

This view revealed family pairs which aligned on either side of the deletion, and a sharp drop in read coverage in a 156 base region: chr22:23803175-23803331.

3.4.5 Orthogonal validation of Copy-Number Variants using Droplet Digital PCR

The early results above showed some putative CNVs, and it was necessary to confirm these using an orthogonal method such as ddPCR. Bio-Rad offered a ddPCR assay against Exon 4 of *SMARCB1*, which was geared towards CNV detection. Another assay, against Exon 5 of *SMARCB1* was designed. These 2 assays would be able to confirm the putative whole gene deletions in Samples HC1-10 and HC1-4, by showing a low relative copy-numbers in both assays. The putative Exon 5 deletion in Sample HC1-2 would be confirmed if the Exon 4 assay showed normal relative copy-numbers, whilst Exon 5 showed a lower relative copy-number than the controls. Bioline Human Genomic DNA was used as the control for this verification. The results of the ddPCR (shown in Figure 3.10) confirmed the single exon deletion in Sample HC1-2, the multi-exon deletion in Sample HC1-4, and the lack of a deletion in Exons 4 or 5 in Sample HC1-16. The results from Sample HC1-10 did not, however, show a significant loss in copy-number in either of the exons assayed. This was likely to be an artefactual call in the Next-Generation Sequencing (NGS) data, resulting from the low family depth.

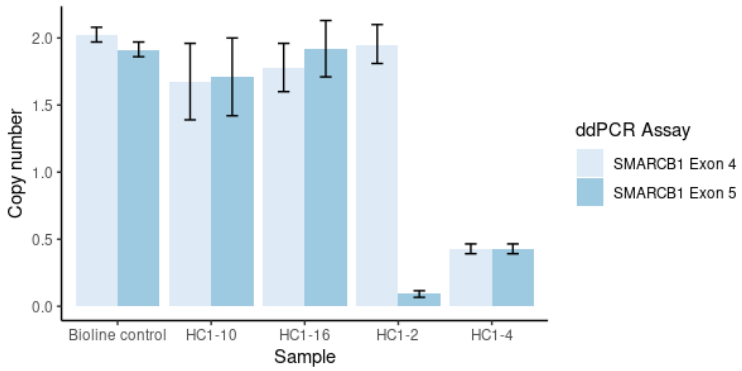


Figure 3.10: CNV results from ddPCR assays on Exons 4 and 5 of *SMARCB1*.

The copy-number values were calculated relative to the assumed diploid RPP30 gene. The error bars represent the 68% Poisson confidence interval, as calculated by QuantaSoft.

3.5 Discussion

The basis of the first half of this chapter was to trial various methods of utilisation of the data, from the recently released barcoded library preparation kit. As more methods were released by other parties, the project moved to keep abreast of the developments, and to include these in the trials. The aims of the preliminary and HC1 studies reflected the shifting nature of the early stages of the work. As methods were settled upon, more development time was applied to them, such as the development of the pipeline centred around Connor. Finally, development time was spent on the creation of software which was able to produce the preliminary results discussed in the preceding Results section.

3.5.1 Investigation of the suitability of pre-existing pipelines and software packages for SNV calling of Tag-seq data

Since there was no software which was explicitly compatible with the Tag-seq kit, at the start of the project, software developed against other barcoding schemes were trialled against Tag-seq data.

A traditional analysis was performed on the data from the preliminary study, to ascertain how much information was available using older techniques. The construction of unbarcoded datasets from the barcoded datasets meant that any technical biases within one dataset were in the other, making the results more comparable. This was a fairer comparison between the two methods than a trial on two different datasets, and represented real-world data better than simulated data. The traditional analysis resulted in a poor 4.7% detection threshold. The inclusion of a true-negative control to this test may have resulted in the selection of a better detection threshold, the result was broadly in line with what was achievable using traditional Illumina sequencing.[246]

The Curio platform was trialled once it was released, and this trial was more focused on the six advertised loci of the Horizon Discovery HD780 samples' SNVs. When focusing on these single sites, there was a low level of noise which pervaded all of the loci, and this was almost completely removed in all samples, by barcode processing. Although the numbers of read families was low in the 0.1% VAF sample (Sample 8), the lack of noise across all samples was promising. The small number of families in Sample 8 exemplified the need for increased read depth to maximise family formation, and the importance of optimising the wet-lab workflow efficiency. These results were the deciding factor in the move to the Illumina HiSeq sequencer at the end of the preliminary study. A higher sequencing

depth than was possible on the NextSeq 500, as well as the higher base quality of the four colour chemistry, promised to increase the number of families, and to reduce the number of sequencing errors in the HC1 study.[237, 260] These factors could increase the sensitivity of the pipeline, and allow for stricter analytical parameters to increase specificity of the assay.

At the time of the beginning of the HC1 study, the Curio platform had not implemented an output to VCF files, and it was difficult to download the BAM files from the platform. The platform had yet to implement a variant annotation system, and the lack of VCF file output meant that variant annotation away from the platform was challenging. The inability of a researcher to view and search through the BAM files meant that the identification of any artefacts was also difficult. Although the platform showed promise, the pace of development and customisability required meant that Curio, in its beta release, was not suitable for this project.

Overall, this set of trials of various software and platforms on Tag-seq data showed that there were no off-the-shelf solutions to handling the data from preprocessing to variant annotation, at the time. The decision was made to shift the focus onto an open-source pipeline, utilising the Connor deduplicator, for maximum customisability.

3.5.2 Development of a prototype pipeline for the detection of SNVs

The development of this pipeline was one of the central aims of the HC1 study. To provide useful training data for the creation of this pipeline, the HC1 cohort was chosen from tumour types which were covered by FLCP-1.

During the optimisation of the HD parameter, samples with low inputs were discarded since they displayed unusually high family depths in an early run of the draft pipeline. This was possibly because the family depths of even high depth samples were low: under 1500 out of a theoretical 1.7×10^7 combinations. The proportionality of the family depth decrease and the family depth, when increasing the HD from 1 to 2, suggested that Connor was collapsing distinct families into single families.

If a rare variant family were to be collapsed into another family, the probability would favour collapsing it with a family which did not harbour a variant. Resulting from such a collapse, the three possible consensus bases at the variant position would be the variant base, an 'N', or the reference base, depending on the collapser parameters and the relative numbers of reads in the two families. The probability of losing variant families when the HD parameter is too high exemplifies its importance, and the optimisation of this contributed to

the construction of a sensitive pipeline.

The development of SNAFU was a direct result of the artefacts remaining in the collapsed data produced by Connor. Unfortunately, the low family depths in many of the samples hampered the calling of variants with low VAF, so development of this variant caller was continued in the HC2 study. In the ACP cystic fluid samples, where the depth was sufficient, *CTNNB1* high confidence variants were found.

This prototype pipeline was developed from the ground up using open-source tools and packages, and was optimised to handle Tag-seq data. Where tools were unavailable, software was developed to overcome the existence of artefacts, and it was implemented in such a way that it was able to work on HPC platforms. This pipeline was an early version, and it was developed further into the Cerberus pipeline, as described in Chapter 4.

3.5.3 Investigation of CNVs, and wet-lab verification

The central principle of molecular barcoding is its ability to remove PCR duplicates in sequencing data, leaving families which are a close approximation to the original molecules used to construct the libraries. As in Figure 3.6a, Connor was not a perfect system for removing PCR duplicates, but since it was able to produce low noise data for SNV calling, there was a high probability of good CNV results. For expediency, the CNV calls were done manually to assess the possibility of using targeted, barcoded data for CNV calling. The putative variants detected during this analysis were all verified by ddPCR with the exception of HC1-10, which was likely to be an artefact of low family depths in this sample. A notable positive result was the verified Exon 5 deletion in Sample HC1-2, which demonstrates the granularity with which these CNV calls can be made in barcoded data. The deletion of both copies of *SMARCB1* in Sample HC1-4 also explained the lack of *SMARCB1* SNV calls in this sample.

The ability to assay both CNVs and SNVs in the same sample made this an attractive assay. Many current commercial assays, such as Oncomine and InVision are amplicon-based, which limits the number of regions that the assay is capable of targeting. [261, 262] An exception to this was the Guardant360 assay.[47] The Guardant360 assay was promising, but as of 2017, when the HC2 study began, there was no independent verification of these results by an unconnected group. Additionally, the proprietary nature of the Digital Sequencing platform upon which the Guardant360 assay was built was proprietary. The latter factor meant that although the assay was compelling, it did not enable academic re-

search to be based on the work of Lanman *et al.*, nor did it allow for custom panels to be created using Digital Sequencing technology. As a result of this, work continued on the improvement of the technologies in this project in an open manner.

Chapter 4

The creation of a pipeline for the detection of multiple variant types in liquid biopsies

4.1 Introduction

Following on from the work in the HC1 study, development was switched to creation of a new pipeline with three streams. This new pipeline, called Cerberus, was designed to take the data from a single library, and to call SNVs, Insertions/Deletionss (InDels), and CNVs on it. The suppression of the artefacts identified during the HC1 study was improved upon, with the creation of SNAFU v1.2.

The HC2 cohort selected as a testing set for the development of this new pipeline was more heavily weighted to tumours with variants in the *SMARCB1* and *CTNNB1* genes, which was the basis for the selection of ATRT, MRT and *WNT*-Activated Medulloblastoma (WAM) samples. This work culminated in the detection of low VAF SNVs, InDels, and CNVs, including combinations of variant types within the same sample.

The detection of variant families in ACP samples during the HC1 study led to the creation of a cohort of ACP samples, designated HC2C. The samples were serial cystic fluid samples from two patients who underwent experimental Interferon-α (IFN-α) treatment at Great Ormond Street Hospital (GOSH). These were used to test the Cerberus pipeline and SNAFU v1.2 on its monitoring of these patients throughout treatment. The clinical hypothesis for this section of the study was that good treatment response of ACP to IFN-α would result in an increase in cell death. The cystic fluid, not being connected to a wider fluid system, would accumulate the cellular debris, causing a higher concentration of variant DNA to be present within the cystic fluid.

As was true in the HC1 study, the samples in both the HC2 and HC2C cohorts were

supplied retrospectively. As a result, the data produced here, whilst successful, constituted a pilot study and proof of concept for the use of targeted liquid biopsy DNA sequencing in the detection of tumour DNA.

4.1.1 Aims and objectives

The first aim was to create a pipeline framework, which was able to handle data for SNVs, InDels and CNVs calling. Secondly, the study aimed to improve SNAFU, to minimise the effects of artefacts in collapsed data, and to test its specificity on the cohorts. The third aim was to implement a more robust method for CNV calling, and to test this method on both HC1 datasets where ground truth was known, as well as samples of the HC2 cohort. Another aim was to implement small InDel calling within the pipeline. This study also sought to use SNAFU and the HC2C cohort to produce preliminary data to answer the hypothesis that treatment success correlated with higher variant DNA within the cystic fluid. The final aim was to use serial samples from the same individual to compare the tracking of patients using routine cytology with the Version 2 workflow and the Cerberus pipeline.

4.2 Materials and methods

4.2.1 Cerberus - a bioinformatics pipeline for detection of SNVs, CNVs and InDels

At the start of this project, the overall pipeline for the calling of SNVs was set, based on the prototype pipeline developed for the HC1 study. The major changes to this pipeline would be the upgrade to the SNAFU v1.2 variant caller. The framework for data processing before CNV or InDel calling were not yet developed. This section describes the overall pipeline scheme, and how each component fits into the whole. The calling of each variant type is described in subsequent sections.

The first three stages of the pipeline, as depicted in Figure 4.1, were similar to the HC1 SNV pipeline. The sequencing data for the HC2 cohort was downloaded in fastq.gz format from Illumina Basespace onto UCL's Myriad High Performance Computing Cluster. The data for each sample was in two pairs of files, one from each lane on the HiSeq, and these were combined into a single pair of files using a custom script. Each pair of files was QCed using FastQC. The ends of each read in the files were trimmed to remove adaptors and UMI stem sequences using Trim Galore. Alignment of the reads to human genome version GRCh38 was performed using BWA mem, and the resulting BAM files were sorted, filtered to remove any unaligned/poorly aligned reads, and indexed by Samtools.

To accommodate the introduction of InDel calling to the pipeline, GATK InDel realignment was added to the pipeline.[263] This was due to a number of studies' reports of small increases in InDel calling sensitivity after realignment.[264, 265] Additionally, Connor was run with more promiscuous parameters, as the InDel error rate of both PCR and Illumina HiSeq sequencing was low compared to their SNV error rates.[113, 266] This enabled the retrieval of families which may rarely include substitution errors, and would thus be unsuitable for rare SNV calling, but were unlikely to have InDel errors. Connor was run such that output families were made up of at least two reads, and any base within a family with less than 80% consensus was changed to an 'N'.

The addition of CoNVaDING to the pipeline was achieved simply by using the BAM files prepared for SNV calling as inputs for the CoNVaDING package. For SNV and CNV detection, more stringent parameters were used for Connor than for InDel calling. A hamming distance of 1, a minimum family size of 3, and a minimum consensus threshold of 95% were used, and the output was fed into each variant caller.

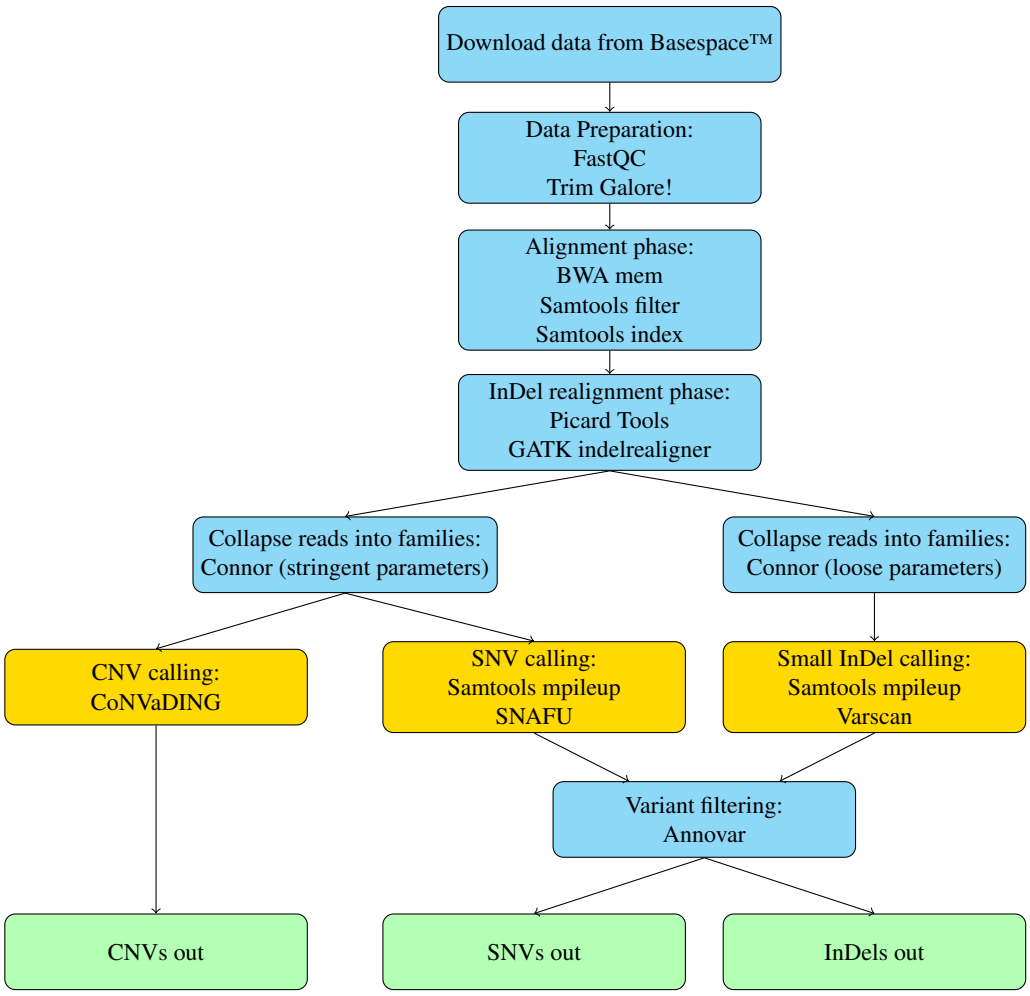


Figure 4.1: The Cerberus pipeline.
The main phases of the pipeline, and the principle software packages and commands used at each stage. The gold boxes highlight the three variant calling phases.

4.2.2 Implementing InDel calling in the Cerberus pipeline

The collapsed families were converted into mpileup format using Samtools, and Varscan was used on the result, for variant calling.[244] Variants were called at a position if at least five families overlapped the position, at least four families supported the variant, and the p-value for the variant was less than 0.05. The resulting variants were filtered and annotated using Annovar, with the parameters listed in Table 4.1.[220, 245, 247–249, 251, 267, 268]

Table 4.1: Parameters used for Annovar filter-based annotation of SNVs and InDels in the Cerberus pipeline

Parameter	Value
Genome build version	GRCh38
Refseq Gene variant exonic function [220]	Exclude synonymous variants
ClinVar clinical signature [268]	Exclude variants marked as non-pathogenic
ESP 6500siv2 database allele frequency [250]	Exclude variants in database at above 0.1%
ExAC 03 database overall allele frequency [251]	Exclude variants in database at above 0.1%
Kaviar 23 September 2015 database allele frequency [249]	Exclude variants in database at above 0.1%
HRC R1 database allele frequency [247]	Exclude variants in database at above 0.1%

4.2.3 Single Nucleotide Variant detection using SNAFU v1.2

In preparation for variant calling, the collapsed families were converted into mpileup format using Samtools. The output was fed, along with the collapsed BAM file, into an upgraded version of SNAFU.

Based on the results from the HC1 study, there were some artefacts which remained unsuppressed by the original version of SNAFU. With a maximum input of 29,000 haploid copies (100ng) into library preparation, and a maximum efficiency at this input of 15.9%, as shown in Figure 2.13, the maximum expected mean family depth was 4600. This was four orders of magnitude below the 16777216 possible barcodes in a 2×6nt scheme. All variant families deriving from separate original template molecules, overlapping a given point on the genome, could therefore be expected to have different barcodes. The main change from SNAFU v1.0 to v1.2 was this assumption.

During early analysis of the sequencing data, it was noticed that collapsed data contained potential artefacts of the following types:

1. One family's consensus read pair had different start sites from another family, but both families had the same barcode pair.

2. Two reads, in separate families, shared an alignment start site, and a barcode. This was observed in files with low family depths where its occurrence was highly unlikely.

3. Two families had consensus reads where the forward start sites and reverse start sites, respectively, were the same. The barcodes were different. This was also observed in very low depth files.

These groups of families were discovered because they had shared variants within them, which were not present in any other families. The presence of a variant where all variant families had the same start sites were strongly indicative of an artefact of the workflow. SNAFU removed the effects of these variants by collapsing variant read families into new families, based purely on barcodes. Within this re-collapsed data, a pair of variant families was sought where:

1. All forward consensus reads from new family (a) were a minimum distance from forward consensus reads from new family (b)

2. All reverse consensus reads from new family (a) were a minimum distance from reverse consensus reads from new family (b)

3. The forward barcodes from the two families were different

4. The reverse barcodes from the two families were different

After re-collapsing variants by barcode, variants which passed the above test, and had variant frequencies and family depths above threshold, were outputted in a VCF file. A schematic for SNAFU's logic is presented in Figure 4.2, and the thresholds used are presented in Table 4.2. A minimum family depth of 100 was chosen as a threshold, to minimise the effects of any remaining barcode family artefacts, as the artefacts found thus far were made up of up to ten members. These variants were then filtered and annotated using Annovar and the parameters in Table 4.1.

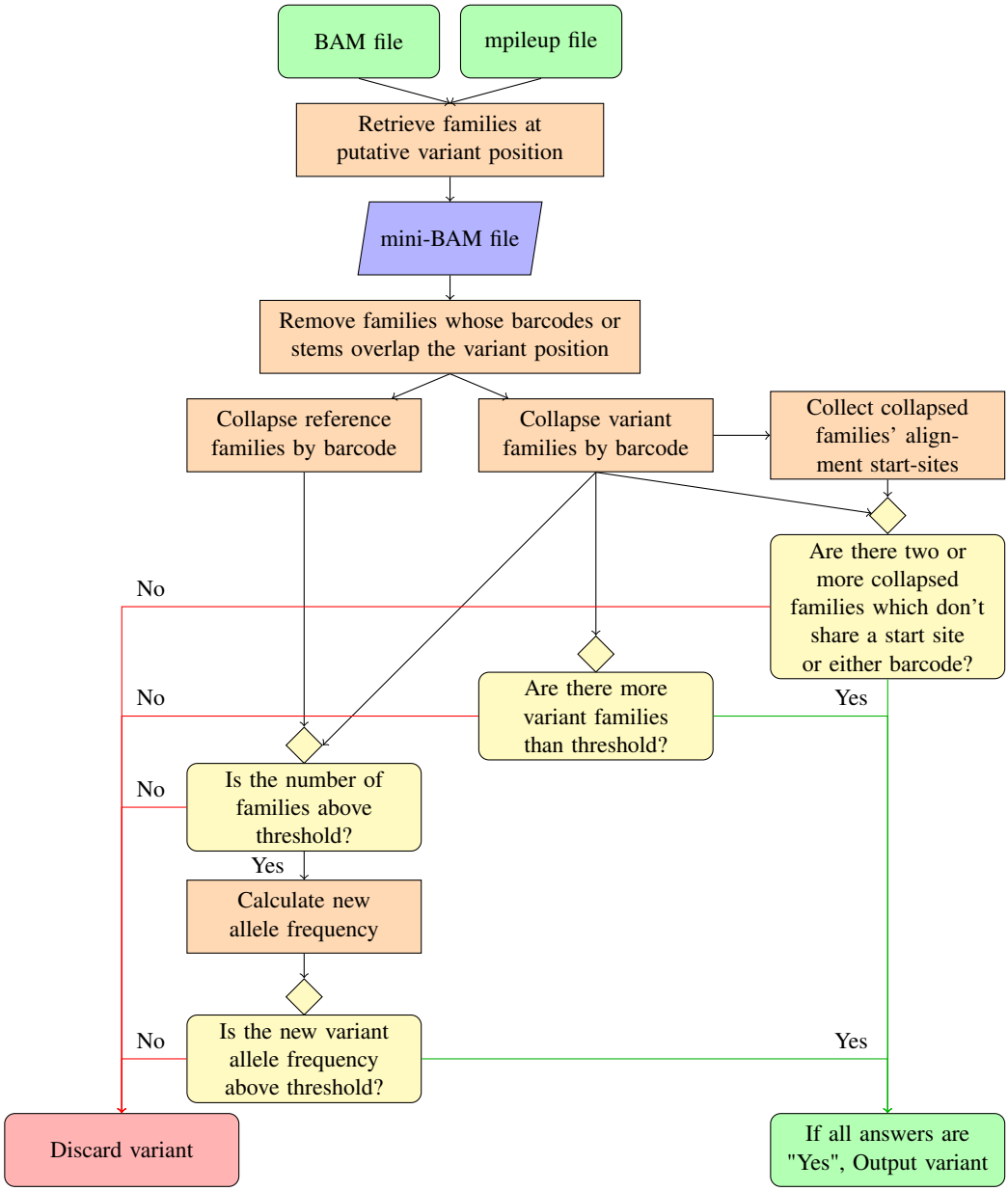


Figure 4.2: The logic which underpins SNAFU v1.2.
A flow chart showing the decisions and processes which SNAFU used to call SNVs. These decisions centred around the removal of artefactual groups of read families, which had filterable features such as barcode similarities, as discussed in 3.4.3.2. The traditional variant calling thresholds of VAF, variant family number and family depth were also implemented in this variant caller.

4.2.4 Using CoNVaDING for the detection of CNVs

CoNVaDING is a package which takes an aligned tumour targeted NGS dataset, and compares it to a pool of similar, CNV-neutral control datasets. Some regions of FLCP-1 were

Table 4.2: The thresholds used for the SNAFU v1.2 variant caller during the HC2 study

Parameter	Value
Minimum variant frequency	0.4%
Minimum variant families	3
Minimum overall family depth	100
Minimum distance between alignment start-sites	3

prone to alignment artefacts, so the first step in implementation was to remove these regions from consideration by CoNVaDING, to make both the control and the tumour data as consistent as possible. Due to the low number of CNV-neutral control samples, it was postulated that CoNVaDING would have difficulty in comparing sex chromosome regions between tumour and normal samples. As a result, regions on sex-chromosomes were removed from analysis.

Alignment artefacts were discovered in the data, which presented as a single base at which the family depth dropped by over 25%, creating a cliff-shape in a read depth graph, as presented in the Results section (Figure 4.4). A Perl script was written to locate alignment artefacts of the type. Across all covered regions of the FLCP-1 panel, the script looked for adjacent positions where the family depth of one position was less than 75% of the depth at the adjacent position. Bedtools was used to find which FLCP-1 region a problem position belonged to, and any exons which appeared five or more times in the data (Figure 4.6) were removed from FLCP-1, for the purposes of CNV analysis.[231]

Ideally, the CNV-neutral control samples are of the same fluid type as the tumour samples, and they are processed in the same way as the tumour samples. The author was unable to obtain ethical approval for control CSF samples during this study, so tumour samples which were likely to be CNV-neutral were sought. ATRT patients which were positive for two of either SNVs, InDels, or a combination of the two, were unlikely to have an additional CNV. All samples from patients where two variants were previously found were used as CNV-neutral control samples for the remainder. A total of fourteen ATRT samples from two individuals were used as CNV-neutral controls, along with thirteen ACP cystic fluid samples and two ACP plasma samples.

During the first step in the analysis, CoNVaDING internally normalised the read depth of each targeted region to the mean depth across all targeted regions of the panel, for each sample. The normalised read depth profile across all regions for a tumour sample was

compared to the profiles of all CNV-neutral control samples.

Each control sample was given an Average Best Match Score (ABMS), which is a measure of how dissimilar the CNV-neutral control's profile is to the tumour sample. The CNV-neutral controls which are most similar to the tumour sample were selected, and a mean ABMS was calculated. The default number of CNV-neutral controls chosen by CoNVaDING was 30 from a much larger pool of controls. The Connor-collapsed data used in this study was largely free from PCR duplicates, and the number of CNV-neutral controls available in this study was small. It was for these reasons that CoNVaDING was set to select the top 6 CNV-neutral controls for each tumour sample. The Mean ABMS for the controls associated with a given tumour sample was calculated, and used as a quality metric for the reliability of variant calling. A mean ABMS threshold of 0.095 was used here, and any samples above this threshold were discarded from further analysis.

CoNVaDING called CNVs by comparing the internally normalised family depth at an exon of the tumour sample's dataset against the mean of the normalised family depths of the controls. Default, strict, parameters were used during calling.

4.3 Results

4.3.1 The cohorts of the HC2 study

The HC2 cohort was primarily made up of ATRT and Malignant Rhabdoid Tumour (MRT) samples. The rationale for this selection was that both tumour-types were driven by the biallelic inactivation of the *SMARCB1* gene, with no known alterations in the rest of the genome.[160–164] This lack of expected variants would be useful for testing the specificity of the pipeline, in the absence of true normal liquid biopsy samples. The samples are detailed in Appendix D.

The HC2C cohort was made up of five ACP cystic fluid samples each from two patients treated with IFN-α. These samples were collected over the course of a 22-25 day IFN-α treatment programme, and at some time points, there were multiple samples. The purpose of this cohort was to test the SNV calling portion of the pipeline on its ability to track patients through the course of treatment. This was used to answer the hypothesis that a successful response to treatment would cause an increase in the VAF of cystic fluid DNA. The cystic fluid fraction of each sample (i.e.: supernatant, cellular fraction or unspun), and their library preparation inputs and outputs, were previously detailed in Table 2.6. The presence of cellular and supernatant samples at the same timepoint also allowed for the assessment of whether cellular fractions of cystic fluid yielded differing amounts of variant DNA from their cell-free counterparts.

4.3.1.1 Termination of analysis for low-quality samples from the HC2 cohort

During the pre-library preparation sample handling stage for the HC2 cohort, quantification of the DNA was performed using ddPCR, as presented in Section 2.3.1.3. The samples from patients with MRT, Pilocytic Astrocytoma (PA) or WAM did not produce data which was likely to yield variants. Table 4.3 shows the very low family depths, and the theoretical minimum detectable VAF based on the requirement of two families to support a variant and no minimum family depth. The final parameters for SNV calling in Cerberus required a minimum family depth of 100 before calling, and these read depths did not pass CoNVaDING's QC for CNV calling. As a result, no variants of these types were able to be called for any of these samples. The raw InDel data for these samples is available in Appendix B, where there are some artefactual InDels in *PTEN*, and some intronic *BRAF* InDels which cannot be verified with existing data. As a result of the factors listed above, these tumours were excluded from further analysis.

The three DIPG plasma samples (HC2-53, HC2-54 and HC2-55) which were run through this pipeline yielded better mean family depths (861, 278, and 385 respectively). Their yields were not, however, enough to detect variants above the noise floor of this system, so analysis was terminated in these samples.

Sample HC2-37 failed to produce any quantifiable yield after library preparation. The theoretical maximum sensitivity for a sample was calculated based on the requirement of at least two variant families to support calling of a variant.

Table 4.3: Samples with low inputs, resulting family depths and their theoretical minimum detectable VAF.

Sample	Tumour type	Patient	Original input (ng)	Mean family depth	Theoretical minimum detectable VAF
HC2-32	MRT	B7	0	5.5	36.6%
HC2-33	MRT	B7	0	6.9	29.1%
HC2-35	MRT	B8	0	2.6	78.1%
HC2-36	MRT	B8	1	2.9	68.7%
HC2-37	MRT	B8	0	<Failed>	
HC2-38	MRT	B9	0.9	21.0	9.5%
HC2-39	WAM	B10	0.3	6.1	33.1%
HC2-40	WAM	B11	0.4	11.3	17.7%
HC2-41	WAM	B11	0.2	6.3	31.6%
HC2-42	PA	B12	0	2.1	96.2%
HC2-43	DIPG	B13	0	3.9	50.9%

4.3.2 Improving the SNAFU variant caller

The method used to filter artefactual SNVs implemented in SNAFU v1.0 needed improvement. By grouping variant family pairs by their barcodes and their alignment start sites, it was hypothesised that it was possible to remove many of the artefacts from the data. To test this, the remaining thirty ATRT datasets were run on the Cerberus pipeline. Each sample's reads were collapsed into families, SNVs were called by SNAFU, and the variants were filtered by Annovar. The SNVs found within *SMARCB1* are presented in Table 4.4.

Table 4.4: *SMARCB1* SNVs called by SNAFU, which passed Annovar filter-based annotation.

Patient	Sample	GRCh38 Position	Ref.	SNV	family depth	VAF	Transcript and protein change (NM_003073)
B1	HC2-11	chr22-23791894	G	T	134	2.2%	c.G232T:p.D78Y
B5	HC2-29	chr22-23816830	C	A	310	1.0%	c.C689A:p.P230Q
B15	HC2-45	chr22-23800967	G	T	439	0.70%	c.G386T:p.S129I
B16	HC2-48	chr22-23793558	G	C	164	51%	Exon 3 acceptor splice site

4.3.2.1 Testing the specificity of SNAFU v1.2 on ATRT and ACP samples

For specificity calculation, the HC2 ATRT samples and HC2C ACP samples were initially treated separately, then combined into a final result with a weighted mean.

Since ACPs are driven by activating substitution mutations in the *CTNNB1* Exon 3 phosphorylation motif, and are mutationally quiet otherwise, any SNVs with a VAF below 30% and not in *CTNNB1* could be assumed to be false-positives for ACP samples.[176, 177] All variants above 30% VAF were considered germline, and only *CTNNB1* Exon 3 variants were considered true somatic. This gave a mean false positive rate of 38.6 per sample, and an overall per base specificity of 99.956%.

Many of the ATRT samples had low mean family depths, and inclusion of these into specificity calculations would artificially decrease the number of false-positives, and give an inflated specificity value. For this reason, only the eight samples with a mean family depth above 100 were included. ATRTs are driven by biallelic *SMARCB1* inactivation, with no other somatic alterations (apart from rare *SMARCA4*-driven tumours).[16, 156, 157] Any variants outside of *SMARCB1*, with a VAF of 30% or above, were treated as likely germline variants, and any with a VAF below 30% were treated as false-positives. Only variants within *SMARCB1* were treated as true-positives. Within these samples, the mean false-positive rate was 11.63, and the per base specificity was 99.987%.

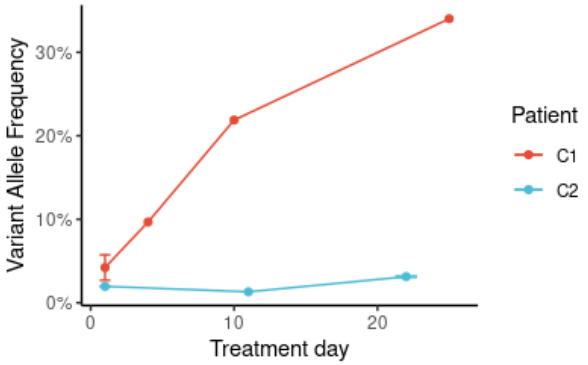
Combining the two with a weighted mean resulted in an overall per base specificity of 99.970%.

4.3.3 Using SNAFU v1.2 to correlate variant allele frequency with treatment success in ACP

Following the success in the detection of variant DNA in the cystic fluid samples of the HC1 cohort, the HC2C cohort, was retrospectively procured. This cohort was run through the Version 2 wet-lab workflow, and through the Cerberus pipeline. SNV calling yielded *CTNNB1* variants from all ten samples in HC2C, as presented in tabular and graphical form, in Figure 4.3. Patient C1, who responded well to treatment, had increasing levels of variant DNA in the cystic fluid supernatant. Contrastingly, Patient C2, who did not respond well to treatment, had steadily low relative amounts of variant DNA in their cystic fluid throughout the course of treatment. The day 1 timepoint for Patient C1 had a standard deviation (σ) of 1.5%, whilst the day 1 and day 22 timepoints for Patient C2 had σ of 0.01% and 0.07% respectively. These low σ values were achieved when combining sequencing data from different sample types (Table 2.6).

Patient	Sample	Treatment day	Position	Ref.	SNV	Family depth	Frequency
C1	HC2C-1	1	chr3:41224634	C	T	1050	3.14%
C1	HC2C-2	1	chr3:41224634	C	T	1529	5.30%
C1	HC2C-3	4	chr3:41224634	C	T	1437	9.67%
C1	HC2C-4	10	chr3:41224634	C	T	1408	21.9%
C1	HC2C-5	25	chr3:41224634	C	T	1308	34.0%
C2	HC2C-6	1	chr3:41224622	C	T	1435	1.95%
C2	HC2C-7	1	chr3:41224622	C	T	1269	1.97%
C2	HC2C-8	11	chr3:41224622	C	T	458	1.31%
C2	HC2C-9	22	chr3:41224622	C	T	1793	3.18%
C2	HC2C-10	22	chr3:41224622	C	T	1790	3.07%

(a)



(b)

Figure 4.3: The difference in relative cystic fluid variant DNA levels between two patients over the course of IFN-α treatment.

Error bars represent the standard deviation in VAF between samples at the same timepoint, where available.

4.3.4 Adding InDel calling to the Cerberus pipeline

The addition of InDel calling to the pipeline leveraged the relative rates of different types of error in both PCR and Illumina HiSeq sequencing. In both of these, insertion/deletion errors are far rarer than substitution errors.[113, 266] This allowed for the use of Connor on more promiscuous parameters to minimise the number of molecules lost during the collapsing of reads into families. The resulting families would give rise to more false-positive SNVs, but any InDels within the data would likely be real. The more complete data was used to maximise the family depth for the large number of HC2 samples which had mean family depths of below 100. For initial variant calling, four variant families were required for a positive call.

The resulting InDel calls within *SMARCB1* are presented in Table 4.5. The minimum family depth requirement of five for variant calling, and the inclusion of ATRT samples with very low mean family depths, many of the VAFs in this table were unreliable.

Table 4.5: InDels detected by Varscan, following Annovar-based filtering and annotation. All transcript changes were described based on the NM_003073 transcript

Patient	Sample	GRCh38 Position	Ref.	InDel	family depth	VAF	Transcript change
B1	HC2-11	chr22:23834166-23834166	-	C	92	65%	c.1145dupC p.A382fs
B1	HC2-12	chr22:23834166-23834166	-	C	16	60%	c.1145dupC p.A382fs
B1	HC2-16	chr22:23834166-23834166	-	C	96	7.3%	c.1145dupC p.A382fs
B1	HC2-19	chr22:23834166-23834166	-	C	2901	2.90%	c.1145dupC p.A382fs
B2	HC2-21	chr22:23834165-23834165	G	-	5	80%	c.1143delG p.T381fs
B2	HC2-24	chr22:23834165-23834165	G	-	10	40%	c.1143delG p.T381fs
B2	HC2-24	chr22:23834167-23834167	C	-	7	80%	c.1145delC p.A382fs

The specificity of the InDel calling was investigated next. For this, the HC2 ATRT samples and HC2C ACP samples were initially treated separately, then combined into a final result, in a similar manner to the SNAFU specificity calculation. Since ACPs are driven by activating substitution mutations in the *CTNNB1* Exon 3 phosphorylation motif,

and are mutationally quiet otherwise, any InDels with a VAF below 30% across FLCP-1 could be assumed to be false-positives for ACP samples.[176, 177] As discussed above, any InDels outside of *SMARCB1* with a VAF below 30% were also assumed to be false-positives for ATRT samples. In both cases, only samples with a mean family depth above 100 were considered for specificity measurement, resulting in eight ATRT samples and all ten ACP samples. The ATRT samples had a mean false-positive rate of 11.25, and a mean per base specificity of 99.987%. The ACP samples had a mean false-positive rate of 17.5, and a mean per base specificity of 99.98%. The final per base specificity, a weighted mean of the two values, was 99.983%.

4.3.5 Implementation of CNV calling to the Cerberus pipeline

4.3.5.1 Identification of exons prone to possible alignment artefacts

The first step toward the addition of CNV calling was the removal of problematic regions, which could impact the accuracy of the CNV calling. Manual viewing of the sequencing data from this cohort in IGV revealed possible alignment artefacts which could interfere with calls.[254] A script was written to identify exons which were prone to these artefacts, an example of which is in Figure 4.4. The script was run on the FLCP-1 covered regions of both the tumour and the CNV-neutral controls used in this study. Any regions in which an artefact of this type was found five or more times were excluded from use as input for the CoNVaDING CNV calling package (Table 4.6).

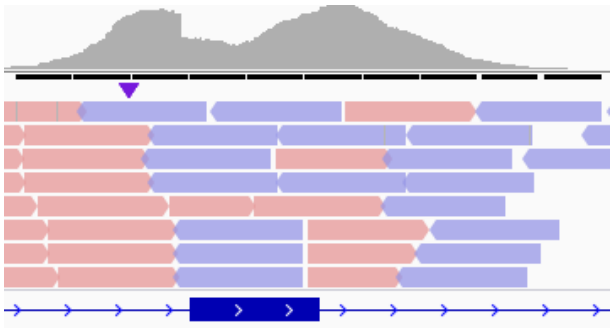


Figure 4.4: An exon with a suspected alignment artefact.

A position where the alignment of a significant proportion of reads' starts indicates an artefact which would skew family depth-based CNV detection. This artefact was found in *SUFU* Exon 8.

Table 4.6: The exons in which suspected alignment artefacts appeared at least 5 times, in the data which was to be used for CNV calling.

Chromosome	Start	End	Gene	No. suspected alignment artefacts
chr10	87960890	87961129	*PTEN*	11
chr10	102599369	102599608	*SUFU*	10
chr17	31232051	31232290	*NF1*	9
chr17	7675983	7676282	*TP53*	6
chr17	7673467	7673903	*TP53*	5
chr19	11021711	11021980	*SMARCA4*	5
chr5	68273331	68273540	*PIK3R1*	5
chr6	135190102	135190371	*MYB*	13
chr6	26031630	26032079	*HIST1H3B*	11
chr7	140782237	140782566	*BRAF*	22
chr7	116769621	116769890	*MET*	14
chr7	140787313	140787882	*BRAF*	13
chr7	140784933	140787002	*BRAF*	11
chr7	140783565	140784644	*BRAF*	8
chr9	95457988	95458317	*PTCH1*	7

4.3.5.2 Using CoNVaDING to call CNVs in the HC2 cohort

During read depth or family depth-based CNV calling, tumour samples are compared to CNV-neutral control samples. In the absence of available true control CSF samples in this study, ATRT samples where two previous variants had been found in the patient were used as CNV-neutral controls. This was done because ATRTs are driven by biallelic inactivation of *SMARCB1*, with the rare exception of *SMARCA4*-driven tumours, and none have any other somatic variant.[16, 156, 157] This means that where two variants had been found, there were unlikely to be any more variants. Additionally, samples of the HC2C cohort, and ACP samples of the HC1 cohort were used as CNV-neutral controls. This is because ACP is similarly driven by variants in a single region: *CTNNB1* Exon 3, with no known CNV involvement.[174, 178, 258] This formed a pool of fourteen ATRT samples, thirteen ACP cystic fluid samples, and two ACP plasma samples. Sixteen ATRT CSF samples were analysed, along with the five sequenced MRT samples, and one DIPG CSF sample. Three CSF samples from the HC1 study were added to the analysis as controls, all of which had their copy-number status in Exons 4 and 5 confirmed by ddPCR (Figure 3.10). To avoid confusion between these three samples and the CNV-neutral control samples, these samples are referred to as "testers".

During the first step of analysis, CoNVaDING internally mean normalised the family depth of each covered region within a tumour sample. The profile of the normalised family depths of the CNV-neutral control samples were compared to the tumour sample, and the six controls with the lowest ABMS were chosen. Using a mean ABMS threshold of 0.095, Figure 4.5 shows that the family depth was related to the Mean ABMS, and a higher family depth was correlated with a lower Mean ABMS.

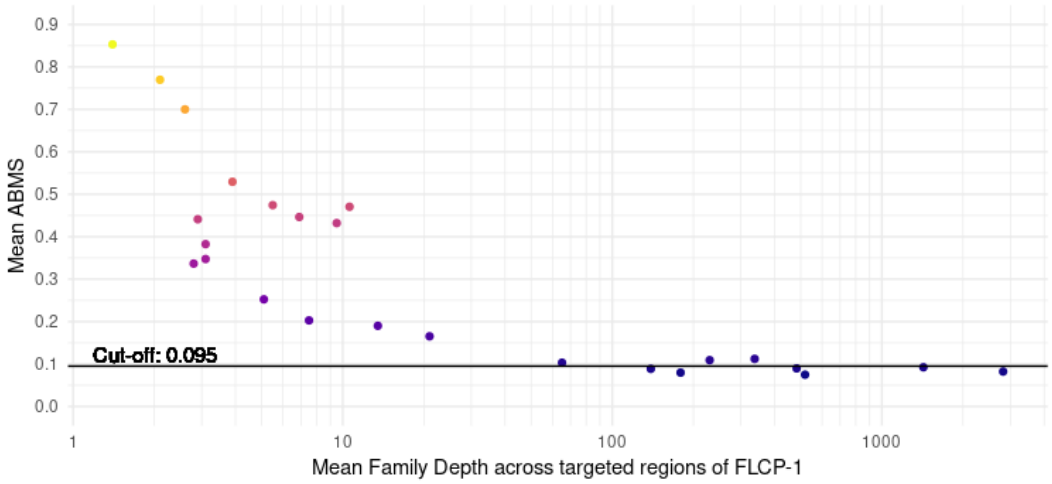


Figure 4.5: The Mean Average Best Matchscore for tumour samples and their relation to family depth.
The Mean ABMS for each sample was automatically calculated by CoNVaDING based on the six CNV-neutral controls that best matched the sample.

Six samples passed this QC: four ATRT samples from the current cohort, and two testers. Of the testers, Sample HC1-4 had a verified Exon 4 and 5 deletion, whilst no deletion had been detected in HC1-16. As mentioned in 4.3.1.1, no MRT samples passed this QC, which was likely due to low family depths across the panel causing increased variation in the mean normalised depth profiles. The samples which passed the quality check were subjected to CNV calling, and the results are in Table 4.7. No CNV call was made in *SMARCB1* for HC2-49 because the CNV-neutral controls for this sample showed significant normalised family depth variability in all nine exons of *SMARCB1*. CoNVaDING marked all nine exons as poor quality, and analysis was terminated at this stage.

For each exon, the ratio of the normalised copy-number of a tumour sample to the mean of the CNV-neutral controls was calculated. These results (Table 4.6) expand on the

Table 4.7: CoNVaDING CNV calling results of samples from HC1 and HC2

Sample	CNV result
HC2-45	*SMARCB1* deletion
HC2-46	No CNV detected
HC2-48	NF1 duplication
HC2-49	No call in *SMARCB1*
HC1-4 (deletion-positive)	*SMARCB1* deletion
HC1-16 (deletion negative)	No CNV detected

ddPCR results of Figure 3.10, showing that the deletion of HC1-4 was a full gene deletion. Sample HC2-45 displays bimodal behaviour at the exon level. In Exons 1 to 5, the copy-number ratio was 0.23-0.25, consistent with the total loss exemplified by HC1-4. Exons 6 to 9, on the other hand, had ratios of 0.59-0.61, which was higher than values associated with a total loss, but below the values in HC1-16. This data suggests that there was a loss of both copies of *SMARCB1* between Exons 1 and 5, and a single copy of Exons 5 to 9 remained in the tumour.

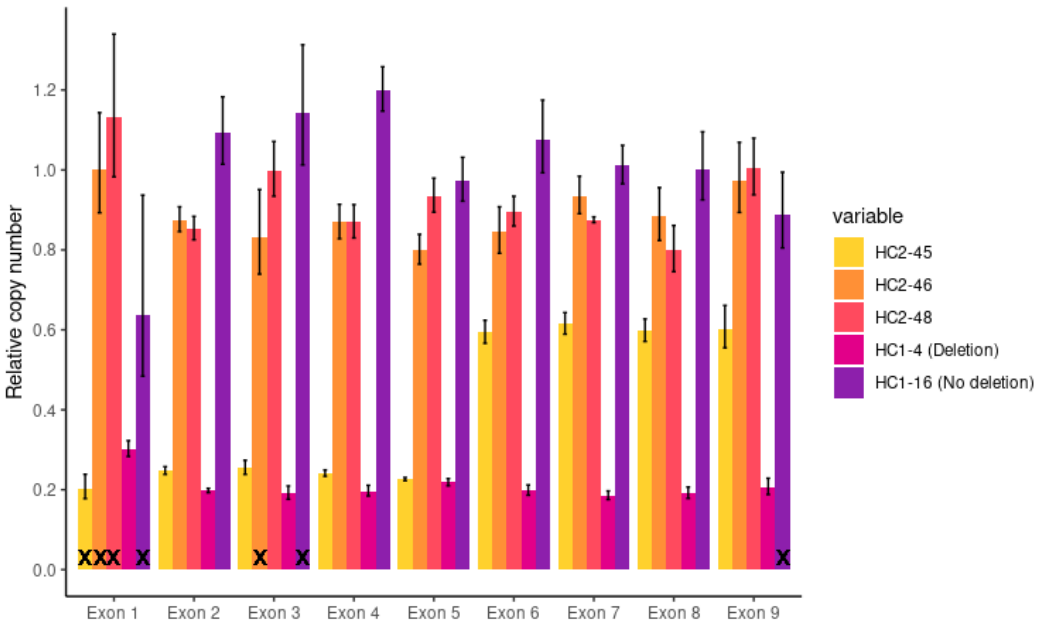


Figure 4.6: The normalised copy-number values for samples of interest relative to their CoNVaDING-chosen controls, at exon resolution.

X's indicate low-quality exons where the controls chosen for the sample by CoNVaDING showed a coefficient of variation of larger than 0.1.

4.3.6 Tracking of genetic variants in ATRT patients throughout treatment

Of the thirty ATRT CSF samples, nine were derived from Patient B1 and five samples were derived from Patient B2. The detection of variants in any sample from a patient allowed for the tracking of circulating tumour DNA (ctDNA) in other samples from the same individual. This is possible irrespective of the fact that many variants did not pass the stringent detection thresholds used for primary variant calling in all samples, as ATRT is genomically stable.[269]

Table 4.8 illustrates that whilst the cytological results for this patient were inconclusive throughout the course of treatment, ctDNA was observed at six out of nine time points. Where ctDNA was not observed, the family depth was below 30 - a family depth which would make rare variant detection problematic, as explored previously in 4.3.1.1.

Table 4.9 shows that cfDNA detection for samples from Patient B2 was more successful than in Patient B1. Any position in which a variant had been detected in at least one sample harboured variant families in all other samples where the depth was above zero. The high VAFs observed for these variants are discussed in 4.4.

Sample	Timing of sample in treatment schedule	Cytological detection of tumour cells in CSF	chr22:23834166-23834166 c.1145insC p.A382fs	chr22:23791894-23791894 c.G232T p.D78Y
HC2-19	Post diagnosis-two weeks into treatment	Inflammation	2.9% (2901) Detected	0.12% (5032) Below threshold
HC2-17	Six weeks into treatment	Persistent inflammation, but no definite cells seen	10.0% (17) Below threshold	Not observed (14)
HC2-16	Eight weeks into treatment	Negative	7.3% (96) Detected	Not observed (116)
HC2-14	Twelve weeks into treatment	Atypical cells	Not observed (22)	Not observed (25)
HC2-13	Fourteen weeks into treatment	Some atypical cells? Reactive	30% (6) Below threshold	Not observed (2)
HC2-18	After eighteen weeks induction treatment	Negative	Not observed (2)	Not observed (26)
HC2-15	Post high dose - three months after starting maintenance treatment	Negative	Not observed (6)	Not observed (13)
HC2-12	One week after disease progression in spine	Negative	63% (16) Detected	Not observed (12)
HC2-11	Four months post progression, two months post radiotherapy	Positive	65% (92) Detected	2.3% (130) Detected

Table 4.8: The tracking of variants throughout the treatment course of Patient B1. Each column headed by a variant shows VAF, and overall family depth in parentheses.

Sample	Timing of sample in treatment schedule	Cytological detection of tumour cells in CSF	chr22:23834165-23834165 c.1143delG p.T381fs	chr22:23834167- 23834167 c.1145delC p.A382fs
HC2-24	Post diagnosis - seven weeks into treatment	Positive	40% (10) Detected	50% (12) Detected
HC2-23	Ten weeks into treatment	Positive	50% (6) Below threshold	50% (6) Below threshold
HC2-22	Twelve weeks into treatment	Negative	50% (2) Below threshold	50% (2) Below threshold
HC2-21	Twelve weeks into treatment	Single tumour cell seen	80% (5) Detected	20% (5) Below threshold
HC2-20	Thirteen weeks into treatment	Acute inflammation and apoptosis, no viable tumour cells	No call (0)	No call (0)

Table 4.9: The tracking of variants throughout the treatment course of Patient B2. Each column headed by a variant shows VAF, and overall family depth in parentheses.

4.4 Discussion

This phase of the project involved the creation of a system which was able to call SNVs, CNVs, and InDels from datasets generated from the same sample, in liquid biopsies. The Version 2 wet-lab workflow was used to generate raw data, which was used as training data for the creation of the Cerberus pipeline. The overall pipeline relied on the principle of simplicity, minimising the number of steps needed to generate the results of the three variant calling and annotation phases. The ability to run the same annotation software and scripts on both the InDel and SNV calls made both the development and the collation of data at the end of a run simple. The implementation of Cerberus on Myriad took advantage of the cluster's scheduling, allowing for all samples in a run to be processed simultaneously. This implementation of the pipeline was successful in achieving the aim of producing a pipeline framework for the calling of SNVs, InDels and CNVs.

4.4.1 Rewriting the SNAFU variant caller, for minimisation of artefacts in the collapsed data

The original version of SNAFU focused on the start sites of the forward and reverse read families. This version was able to suppress many of the artefacts in the output of Connor, but any changes in the alignment start site due to imperfect alignment formed a new family, regardless of the barcode pair. The rewritten SNAFU v1.2 overcame this by assuming that the highest family depth in the best quality sample was below 5000. Since 5000 was four orders of magnitude lower than the 16777216 possible barcode combinations in a 2×6nt barcoding scheme, the chances of clashes were rare. SNAFU required at least two variant families to call a variant, so the probability of an extra reference family clashing with one of these was $2 \div 16777216 = 1.19 \times 10^{-7}$. Using the binomial probability calculator at `https://stattrek.com/online-calculator/binomial.aspx`, the probability of one or more of five thousand randomly selected barcodes intersecting with these two barcodes was 6.0×10^{-4}. Increasing the number of variant families, and therefore the VAF, increased the probability of a barcode clash, but also increased the number of clashes necessary for a call not to be made, resulting in a lower probability of clashes causing a false-negative call. Decreasing the number of reference families also decreased the probability of clashes causing a false-negative. The theoretical maximum probability of a false-negative due to SNAFU's assumption was represented by the scenario outlined above, and this was deemed to be acceptable for SNAFU to be designed in this manner.

The results of the annotated and filtered SNAFU outputs were promising, especially the detection of a variant at 0.70% VAF, with a specificity of 99.970%. Despite this, the data must be considered as preliminary, as the size of the cohorts were small, the samples were of variable quality, and collection methodology was not controlled.

4.4.2 Implementation of InDel calling in the Cerberus pipeline

The implementation of InDel calling was done using relatively well established methods, with one main novel process: the acceptance of smaller families when collapsing using Connor. This was mainly to counteract the effects of a lack of DNA in the original CSF samples of the HC2 cohort, as samples with more input DNA would result in higher family depths, making InDel calling easier.

The lack of commercial control material containing known InDels in the HC1 or HC2 cohorts meant that the system's InDel calling sensitivity was unknown whilst developing the pipeline. As a result, the minimum variant family parameter was set to four to minimise the number of false-positives, at the risk of lowering sensitivity. The aim of this study was to implement InDel calling in the pipeline, and not to optimise the individual parameters, as the cohort was smaller than was ideal for optimisation. Despite this, the overall specificity in the data of 99.983% shows that the noise suppression of the molecular barcoding performed well.

Future work to improve the variant calling parameters for a specific sample type would use a large prospective cohort of samples, with a controlled volume. The cohort would also include negative control samples collected using the same methodology, and positive samples with known alterations at known VAF. These positive controls could be created by diluting InDel-harbouring cell line DNA into reference DNA, in a similar manner to the production of the Horizon Discovery cfDNA controls.

Although the study was primarily technologically based, some preliminary data with possible clinical implications was produced. The InDels found within this study were unusually clustered. Using the NM_003073 transcript as a base, a span of three bases (c.1143-1145 p.381-382) was host to a total of one insertion and two deletions in two individuals. These were unlikely to be alignment artefacts, as no single read harbouring InDels at these positions was found in any other sample of any of the cohorts of the project. All of these variants caused frameshifts at the 3' end of the gene, and similar variants have been found in patients in other studies.[17, 218, 269] These results point to a potential hotspot for

frameshifts in the SMARCB1 gene.

4.4.3 Using CoNVaDING for CNV calling in targeted sequencing

At the time of this study, CoNVaDING was a relatively new CNV calling package.[270] Other packages for CNV calling existed, but CoNVaDING had advantages over these, such as not requiring whole exome or Whole Genome Sequencing (WGS) data, and not requiring a specific version of the genome.[271–273] Some packages, such as CNVkit, counted reads in regions between the targeted regions of the panel. This is useful when using non-molecularly barcoded data, as the read counts are higher than collapsed family counts, making the so called "anti-target" read counts stable. Collapsing the data by barcode resulted in zero or small family counts in these so-called 'anti-target' regions, and CNVkit's normalisation against these numbers gave highly variable results.[274] CoNVaDING was chosen partially due to implementation reasons, and mainly because its suite of QC metrics allowed for the separation of high quality CNV calls from low quality ones.[275]

The suitability of the different sample types' datasets for use as normalisers in CoNVaDING was a concern during development, as mentioned in both the Methods and Results sections. Different sample-types can have different DNA fragment length distributions, owing to the proportions of cellular versus cell-free DNA and the fragmentation they underwent prior to extraction.[70, 71, 276] Longer fragments which overlap a targeted region would lead to larger mean family depths at the region than the same number of shorter fragments, an effect which Plagnol *et al.* sought to remove.[271] Ideally, the control samples would need to be from the same fluid type as the tumour samples, or a fragment length agnostic method could be implemented to reduce the effects of different fluids.

One of the main advantages of using molecular barcoding in reducing the effects of fragment-length on the analysis was the fact that the collapsing reduced the number of PCR duplicates present in the data. This not only reduced the effects of stochasticity on the family depths, but also reduced the effects of fragment length on PCR efficiency. This lead to the positive results in the study, including the detection of the bimodal loss of *SMARCB1* at the exon level.

More broadly, one must compare barcoded NGS in CNV calling to other modern techniques. Fluorescence *in situ* Hybridisation (FISH), the process of applying fluorescent probes to metaphase spreads of chromosomes, is dependent on viewing such CNVs through a microscope, which leads to low achieved resolutions. Array-CGH, relying on

the concept of 'comparative genomic hybridization' between normal and tumour DNA to normal metaphase chromosomes, has improved upon traditional FISH and has a reported resolution of 5-10Mb.[277] These are still far lower than the resolutions achieved within this study. Theoretically, it is possible to fluorescently label short, liquid biopsy DNA for use in array-CGH, but current protocols require inputs of the order of 100ng, which is not tractable given the amounts of DNA within the liquid biopsies used in this study.[278, 279] Multiplex Ligation-Dependent Probe Amplification (MLPA) has been used as a basis for CNV calling since 2002.[280] It relies on the annealing of two probes, which have sequences complementary to adjacent regions on the genome, to sample DNA. Ligation of the probes creates a contiguous template which is able to be PCR amplified and detected. Since then, digitalMLPA has incorporated molecular barcodes into the probes, and the amplified products are sequenced on an NGS platform.[281] The sequenced data can be deduplicated, and the collapsed counts for each probe pair can be analysed for CNV detection. This approach relies on many of the same principles as the work in this study, and the initial input amounts of 40ng in the initial study were low enough to be within the range of liquid biopsies.[123, 281] The annealing of probes to the sample DNA means that the probes must be designed in regions of the genome with no Single Nucleotide Polymorphisms (SNPs). Using such a technique on short fragments of DNA as opposed to full length tumour DNA, such as those commonly found in cfDNA, may reduce the number of input molecules able to be assayed. This is because a short fragment may contain the complementary sequence for one probe, but an incomplete sequence which is insufficient to bind the other probe. Probe lengths must be minimised to reduce the effects of this, or extra DNA above the stated minimum for genomic DNA must be used for such tests. Overall, digitalMLPA is a compelling new technique, and may be amenable to the field of cfDNA with some modifications. One caveat is the ability of this technique to assay for CNVs alone, where the work of this project was aimed at multiple variant types.

4.4.4 Preliminary evidence for the correlation between treatment success and cystic fluid DNA in ACP

The ACP samples used in this retrospective cohort were all of high quality, and resulted in ideal family depths, making them suitable for tracking of the levels of DNA which harboured variants. This resulted in the stark difference between the results of the two patients, with Patient C1, who responded well to treatment, having elevated VAFs throughout treat-

ment after Day 1. The low variability of the samples taken at the same timepoint, from the same patient, but from different fractions of the sample, demonstrate the reliability of the method. They also provided data which showed that separating the cells/debris and supernatant was not necessary, as combinations of cellular fractions, supernatants, and unspun samples showed similar VAFs.

It must be noted that these levels of variant-harbouring cystic fluid DNA were relative, not absolute. The IFN-α treatment was supposedly targeted at the cell type from which ACP originates, but the molecular rationale which underpins IFN-α's effectiveness in ACP is not well understood.[183, 184, 189] As a result, the effect on the non-neoplastic tissue in and around the cyst is not known. The VAF is by its definition a relative term, and the amount of variant DNA was recorded as a percentage of the whole. One can however surmise from data from Patient C1 that IFN-α, especially the supernatant samples, that there was increased cell death in the mutant cells, and that this predominated over death of wild-type cells.

The variability in the response to treatment of such a genetically quiet tumour, particularly the low VAFs in the samples from Patient C2, shows that genetics may not be the main factor which drives treatment resistance. This data agrees with previous assessments of gross tumour size or presence, over the course of treatment, which show similar variability.[182, 183, 189] Probing the transcriptome of ACP cells using an RNA-sequencing experiment could provide answers to why some tumours respond to this therapy, and some do not. Clustering bisulphite sequencing data or methylation array data between responders and non-responders could also identify methylation states in genes which are important for treatment response.

Overall, this experiment was successful in answering the hypothesis that treatment success correlated with higher variant DNA within the cystic fluid, with the caveat that the quantification was relative to the wild-type DNA levels in the sample.

4.4.5 Tracking of variants throughout treatment, and comparison to routine cytology

Whilst tracking the InDels in Patient B2, one of the HC2 cohort patients with ATRT, it was observed that VAFs were high for circulating samples, and that for each sample the VAFs added up to close to 100%. The high VAFs observed for these variants could have been caused by the aggressive tumour releasing large quantities of ctDNA into the CSF, but these

could also be artefacts arising from the low family depths in these samples.

Mosaicism was also a possibility. The total loss of SMARCB1 in embryos before the blastocyst stage is lethal, which makes germline mutation an impossibility.[282–284] It is possible that one InDel was acquired later during embryonic development, leading to a high degree of prevalence in cells in the brain. More samples from other tissues, and from cells derived from other germ layers, would be needed to confirm this hypothesis.

The Version 2 wet-lab workflow, combined with the Cerberus pipeline, was able to track ctDNA throughout the course of treatment with an accuracy that may be better than current cytological methods, for both Patients B1 and B2. This is particularly promising, because the samples used for this analysis were both smaller in volume than the 3ml routine cytological samples used at GOSH, and because the samples did not include the cellular fraction, both of which could improve sensitivity. More data from samples comparable to those used in cytology is needed to confirm this.

Chapter 5

Conclusions and future work

To reiterate, the main hypotheses of the project were as follows:

1. It was possible to create a versatile workflow, which was capable of taking multiple liquid biopsy types and processing them into barcoded, targeted libraries, suitable for Illumina sequencing.

2. It was possible to apply molecular barcoding to liquid biopsies, to improve on cytology: the current gold standard of monitoring in Paediatric Brain Tumours (PBTs) with leptomeningeal involvement.

To explore these hypotheses, the framework for a wet-lab workflow needed to be developed, added to, and optimised. The initial version of this workflow was created during the preliminary study, taking advantage of as many off-the-shelf components as possible, such that development efforts could be concentrated on novel components. One part of this was the creation of FLCP-1, a 118kb panel which was based on the Agilent XT chemistry. This platform allowed for the creation of new panels using the Agilent SureDesign software, adding to the assay-agnosticism of the workflow. This workflow was built upon with the introduction of clinical CSF, plasma and cystic fluid samples, during the HC1 study. Optimising the shearing of DNA previously sheared by QIAamp column processing, whilst keeping short DNA intact, allowed for the workflow to process genomic DNA as well as the short DNA common in liquid biopsies.[68, 116]

Throughout the project, the techniques used in the wet-lab workflow were refined. This lead to the development of a custom magnetic rack for the quick separation of magnetic beads from liquids, allowing for better temperature control during the hybrid capture of libraries. The on-target percentage steadily increased to an overall median above 60%.

Concurrently, the library normalisation and pooling procedures were improved. The control of pooling was such that the predicted fraction of a run's sequencing reads deriving from a sample versus its actual fraction correlated with a Pearson's coefficient of 0.96 (Figure 2.11).

Overall, the results of Chapter 2 demonstrate the possibility of versatile workflow creation, and give a positive answer to the first hypothesis.

This project began at a time when many molecular barcoding technologies existed, but the development of data analysis pipelines to handle the data produced by these technologies was embryonic.[10, 68, 127, 132, 285–287] To answer the second hypothesis, off-the-shelf software was trialled as an initial approach. When existing software was found to be suboptimal, development of an in-house pipeline was started. During the HC1 study, this resulted in the discovery of artefacts in the output of Connor, and the subsequent development of SNAFU v1.0 in an attempt to counteract these. This study also demonstrated the ability of the pipeline to detect CNVs within the targeted sequencing data, at the exon level; results which were validated using ddPCR. This development continued in the HC2 study, with the creation of the Cerberus pipeline. Cerberus combined CNV calling, InDel calling and SNV calling in a single pipeline, along with the improved SNAFU v1.2. Overall specificities of 99.970% for SNV calling, and 99.983% for InDel calling were promising, and it was unfortunate that the samples in the HC2 and HC2C cohorts did not allow for an assessment of specificity. Despite this, the pipeline was able to track patients throughout their treatment, detecting residual disease when cytology was negative or inconclusive. The tracking ability of the pipeline also provided evidence to answer the hypothesis that treatment success lead to an increase in variant-harbouring cystic fluid DNA in ACP cysts. This section of the project represents the first ever sequencing of DNA derived from cystic fluid of ACP patients.

Overall, although the cohorts for this project were retrospective and not ideal, the combination of the workflow and Cerberus were able to provide preliminary evidence that barcoded, targeted sequencing was better than the CSF cytology used at GOSH.

5.0.1 Comparisons to other technologies

In assessing the work performed in this project, one must compare it to other techniques and processes. Mouliere *et al.* were able to detect CNVs in CSF using low-pass WGS.[71] This technology is certainly strong, as it is not only capable of detecting CNVs over large sections

of the genome, but can potentially be used to look for structural rearrangements.[288] The main problem with low-pass sequencing is its lack of sensitivity for rare InDels or SNVs, caused by the low read depth at variant positions. As mentioned in 4.3.1.1, the sensitivity of SNV and InDel detection depends on the read/family depth at the variant position and the number of variant reads/families, so a low read depth leads to low sensitivity. Increasing the depth of WGS is possible, but sequencing costs quickly become prohibitive as a result. The resolution of this sequencing is also low, with every 500bp region represented by a small number of reads. The low resolution leads to the lack of sensitivity for focal InDels which are larger than the small InDels defined by the read aligner, but which are still much smaller than 500bp.[252] One example of this could be the 156bp Exon 5 deletion in Sample HC1-2 (Figure 3.9). The deep, panel based sequencing used in this project has the advantage of being able to detect all three of these variant types, but with a lowered ability to detect novel genetic alterations outside of the panel, and a higher per-sample cost.

Work by Miller *et al.*, published in 2019, highlights the speed of progress in this field.[123] One advantage of their work was the 3.5ml of CSF or plasma per sample, as well as the access to tumour and germline DNA. This allowed for their results to be assessed against the ground truth, and for sensitivity and specificity to be assessed accurately. Whilst the per-base specificity of this system is excellent, there are still some false-positives in the variant list outputs. Work to eliminate these remaining artefacts, such as the recently described Illumina index hopping, is crucial in making this test reliable for minimal residual disease detection.[289, 290] The Memorial Sloan Kettering IMPACT system is likely to be compatible with the work in this project.[291] A hybrid system, which merges methods from this project and from the IMPACT pipeline, could deliver vastly superior results to any produced by either method alone. In particular, the addition of molecular barcodes, as opposed to the Illumina indexes which were referred to as 'barcodes' in their paper, could increase the sensitivity and specificity within their system.[123, 291]

The Oncomine assay family, based on amplicon sequencing technology, have achieved highly multiplex PCRs with molecular barcodes on the primers.[292] These assays are advertised to have detection limits of 0.1% to 0.01%, which are impressive when combined with panels which sequence genomic regions in the megabase range.[293] Assays of this type are highly efficient when converting template DNA into sequenced libraries, as they do not rely on an inefficient ligation step.[294, 295] One downside is that these assays are

prone to first round PCR artefacts, as discussed in 1.2.1. Another downside to amplicon-based assays in general, including the Oncomine family, are the inability to detect structural variants where the partner is unknown. Where both breakpoints of a structural variant occur at known locations, primers can be designed to bridge the breakpoint, and the PCR products can be sequenced. In more complex genomic landscapes, such as that of the Ph-like B-cell precursor acute lymphoblastic leukemia, breakpoints are diverse, and primer design becomes complex.[19] Primer binding sites may also be lost in cases of large InDels, which would reduce the ability of amplicon sequencing to resolve the ends of such InDels. As a technology, amplicon-based barcoded sequencing is less versatile than capture-based sequencing, as the latter does not require priming for targeting, but the annealing of RNA baits. The versatility of capture-based sequencing is also ahead of assays such as the Oncomine family in its ability to add to a panel. Technologies such as the Agilent XT capture kit require the checking that additional baits do not anneal to existing baits in the panel, before addition. This is not so with amplicon sequencing, as any new primers must have their priming conditions validated, and the possibility of priming the amplification of an off-target region of the genome increases as the number of primers in the panel increases. Overall, there are pros and cons to these two main classes of technology, and the choice of capture-based sequencing at the start of this project was made to ensure maximum versatility in the overall workflow developed.

The recently released Agilent XT HS2 kits did not advertise the details of their barcoding scheme, but work by Yamaguchi *et al.* on their short-lived predecessor suggested that their scheme was similar to Duplex Sequencing.[296, 297] This system featured dual 3nt barcodes, giving up to 4096 possible combinations. This system, integrated with the newest version of Agilent's capture technology, was a step towards a workflow with the versatility of custom panels, and a robust barcoding scheme. The closed-source nature of the library preparation kit, combined with the lack of data on the library conversion efficiency, meant that a direct comparison with this kit and the ThruPLEX Tag-seq kit was not possible. It remains to be seen how these two technologies compare to each other.

5.0.2 Future improvements to the wet-lab workflow

Despite the successes in CSF and cystic fluid, plasma still proved problematic for the overall system. More work is needed to increase the amount of DNA which survives between extraction and sequencing, and an increase in specificity would help to increase confidence

in variant calls. One of the ways in which the workflow efficiency could be increased is to optimise the library conversion efficiency. Lanman *et al.* asserted that their Digital Sequencing platform had a library conversion efficiency of over 80%, but neglected to publish a method for how this was achieved.[47] Assuming that negligible numbers of molecules are lost from the already amplified libraries following the use of the ThruPLEX Tag-seq kit, the kit had an approximate efficiency of between 4.5% and 17%. This leaves much room for improvement, and the movement of the barcoding to library preparation kits with higher ligation efficiencies, or the research into a highly efficient library preparation kit represents an avenue for future research.

One of the artefacts from the Connor output was the presence of read pairs with the same read start sites for both paired end reads, but different barcode pairs. One possible explanation for this was that the use of 8-11nt stem sequences 3' of the barcode allowed unligated adaptors to prime amplification of libraries with different barcodes but the same template sequence. Agilent's XT HS2 barcoded library preparation scheme incidentally uses 1-2 "dark bases" in place of these stem sequences. The optimisation of the library preparation technology by reducing the length of the stems, by adding a wash step to remove unligated adaptors, or by adding blocking oligos which bind the stem sequences, could remove these artefacts. Integrated DNA Technologies filed a patent in 2018 for reverse complement adaptors, or blocking oligos, to mitigate this.[298] Using a closed-source system such as the ThruPLEX Tag-seq kit hampered such optimisations in this project, but future work could improve upon this system.

5.0.3 Future improvements to the Cerberus pipeline

Currently, the SNAFU variant caller is designed to remove artefacts from the collapse of barcodes by Connor. These artefacts are not adequately removed by Connor, and it may be more computationally efficient to perform barcode-based collapsing simultaneously with variant calling. Some of the improvements to collapsing performed by SNAFU v1.2 can be integrated into a discrete collapser, such as the collapsing of nearby reads with the same barcodes into a single family. Other improvements, including the disregarding of variants where all variant families start at similar locations, would need to be done at the variant calling level. Future work to increase the efficiency of SNAFU, or a similar variant caller, should focus here. The SNAFU caller performed with a high specificity on the data in this cohort, but there were still artefacts in the data which were not suppressed. One of these

artefact types could be due to Illumina Index switching.[289, 290] A simple solution to this would be to search all other aligned and collapsed BAM files in a run for variant families which have the same barcodes as a given variant. The indexing of BAM files and the use of SAMtools, used in a similar way to the methods used in SNAFU already, would make this a computationally simple process. Further work in the identification of the source of these artefacts, both in the wet-lab workflow and in the data processing, would be needed for progress.

The Cerberus pipeline was able to simultaneously scan for SNVs, InDels and CNVs, but was not designed to detect Copy-Number Neutral Loss of Heterozygosity, or structural variants. Further work on a future iteration of this pipeline, which allows for the detection of more types genetic alteration would greatly improve the power of this system.

5.0.4 Other future work

There is data on the dynamics of cell-free DNA (cfDNA) in plasma, but the turnover rate of cfDNA in CSF and cystic fluid, both of which were used in this project, are not yet known. The cystic fluid was assumed to have no turnover other than that which was due to nucleic acid degradation within the cyst, but this has not been demonstrated. Ethical approval for a temporal study on the levels of ctDNA in CSF following the excision of a tumour from a human patient may be difficult to obtain owing to the invasiveness of obtaining CSF. Serial samples of cystic fluid may be possible, however, since some patients have intracystic catheters, and the aspiration of cystic fluid is an existing part of the care pathway. Mammalian models with similar physiologies may provide a suitable alternative to human studies. This would provide a much needed foundation for future studies on both novel detection methods and treatment efficacy, which would have a major impact on those with PBTs. Studies have been attempted using human xenografts in mice, in the context of plasma DNA as a marker for tumour burden.[299–301] One problem encountered was the ability to extract sufficient quantities of plasma from the animal to accurately perform an assay upon the sample. This is exacerbated by the smaller volume of CSF than blood in a mammal. One potential solution to this, albeit one which requires specialist facilities, is to use a larger mammalian model, such as a domestic pig. Pigs have been used as model organisms for cancer studies, particularly because their physiology and size makes drug dosing similar to that of humans.[302–305]

In any future studies which build on this work or trial similar methods, one major

improvement would be to use samples which compare more closely to those already used in routine care. As exemplified by the HC2 cohort, the lack of DNA in the low volume CSF samples severely hampered the project as a whole. Samples of 3ml including cellular fractions, as used in routine cytology, would give a more fair comparison between any novel methods and the current gold standard. The fact that positive results gained by this project with such small samples were testament to the efficiency of the system as a whole.

5.1 Final remarks

This project supports previous work showing that variants can be detected in CSF, and provides the techniques to facilitate its application to the analysis of cystic fluid in research.[55, 114, 205] Further development of this versatile system has significant potential for a direct impact on patient care, with regards to diagnosis, stratification and tracking.

Appendix A

Table A.1: The original read depths for each allele at each Horizon Discovery cfDNA advertised mutational site for samples 5 to 8. These samples are displayed before and after barcode processing.

TF-5	No barcode processing					
Base	*EGFR* T790M	EGFR L858R	*KRAS* G12D	*NRAS* Q61K	NRAS A59T	*PIK3CA* E545K
A	3	1	3	2	2	0
C	7951	2	7677	2	4454	0
G	0	39	0	4606	0	2108
T	127	6812	86	46	92	0
N	1054	1214	872	720	553	96
TF-5	Barcodes processed					
Base	*EGFR* T790M	EGFR L858R	*KRAS* G12D	*NRAS* Q61K	NRAS A59T	*PIK3CA* E545K
A	0	0	0	0	1	0
C	1093	0	1036	0	573	0
G	0	3	0	613	0	325
T	17	957	12	6	14	0
N	12	7	1	8	2	0
TF-6	No barcode processing					
Base	*EGFR* T790M	EGFR L858R	*KRAS* G12D	*NRAS* Q61K	NRAS A59T	*PIK3CA* E545K
A	4	1	5	0	4	0
C	7470	4	7093	1	4537	0
G	2	69	5	4626	1	2088

T	97	7015	28	39	45	0
N	984	1296	835	771	544	103

TF-6 Barcodes processed

Base	*EGFR* T790M	EGFR L858R	*KRAS* G12D	*NRAS* Q61K	NRAS A59T	*PIK3CA* E545K
A	0	0	0	0	0	0
C	977	0	944	0	587	0
G	0	14	0	621	0	316
T	13	971	4	4	3	0
N	5	8	7	3	4	0

TF-7 No barcode processing

Base	*EGFR* T790M	EGFR L858R	*KRAS* G12D	*NRAS* Q61K	NRAS A59T	*PIK3CA* E545K
A	3	4	0	0	4	0
C	7426	0	7218	2	4405	0
G	1	1	1	4550	0	2222
T	3	6444	0	1	4	0
N	956	1191	823	728	583	112

TF-7 Barcodes processed

Base	*EGFR* T790M	EGFR L858R	*KRAS* G12D	*NRAS* Q61K	NRAS A59T	*PIK3CA* E545K
A	0	0	0	0	0	0
C	965	0	984	0	570	0
G	0	0	0	599	0	325
T	0	987	0	0	0	0
N	6	7	2	3	3	0

TF-8 No barcode processing

Base	*EGFR* T790M	EGFR L858R	*KRAS* G12D	*NRAS* Q61K	NRAS A59T	*PIK3CA* E545K
A	1	4	3	0	2	0
C	4734	0	3941	1	2749	0
G	0	24	0	2756	0	1310

T	3	4467	0	7	2	0
N	606	849	493	499	383	62

TF-8 Barcodes processed

Base	*EGFR* T790M	EGFR L858R	*KRAS* G12D	*NRAS* Q61K	NRAS A59T	*PIK3CA* E545K
A	0	0	0	0	0	0
C	634	0	571	0	368	0
G	0	2	0	378	0	187
T	0	666	0	1	1	0
N	2	5	3	4	3	0

Sample	Position	Ref	Alt	Function	Gene	Depth	Reference depth	Alt. depth	Frequency
HS3-40	chr7:140784650-140784650	-	T	intronic	*BRAF*	11	3	5	45.45%
HS3-40	chr10:87864104-87864104	T	-	splicing	*PTEN*	18	6	12	66.67%
HS3-32	chr7:140782463-140782463	-	A	intronic	*BRAF*	32	15	6	22.22%
HS3-32	chr10:87864104-87864104	T	-	splicing	*PTEN*	18	0	18	100.00%
HS3-39	chr7:140786185-140786185	-	A	intronic	*BRAF*	26	18	6	24.00%
HS3-33	chr7:140789219-140789219	-	A	intronic	*BRAF*	49	38	5	10.20%
HS3-33	chr7:140793160-140793160	C	-	intronic	*BRAF*	22	13	9	40.91%
HS3-33	chr10:102597263-102597263	G	-	exonic	*SUFU*	34	29	5	14.71%
HS3-33	chr17:31258455-31258455	A	-	exonic	*NF1*	5	1	4	80.00%
HS3-43	chr10:87864104-87864104	T	-	splicing	*PTEN*	18	9	9	50.00%
HS3-36	chr10:87864104-87864104	T	-	splicing	*PTEN*	8	1	7	87.50%
HS3-38	chr10:87864104-87864104	T	-	splicing	*PTEN*	27	0	27	100.00%

Figure B.1: The raw post-Annovar InDel calling results for WAM, MRT, DIPG (CSF) and PA samples in Chapter 4. These results show a likely artefactual *PTEN* deletion, numerous intronic *BRAF* InDels which are unable to be verified, and 2 likely false-positives from Sample HS3-33.

Appendix C

Parameter	Value
Incident power (W)	18
Duty factor	20%
Cycles per burst	50
Treatment time (s)	300
Temperature	6-9°C
Water level	10
Sample volume	15µl
E220 – Intensifier	Yes
Container	MicroTUBE-15

(a)

Parameter	Value
Incident power (W)	18
Duty factor	20%
Cycles per burst	50
Treatment time (s)	120
Temperature	6-9°C
Water level	10
Sample volume	15µl
E220 – Intensifier	Yes
Container	MicroTUBE-15

(b)

Parameter	Value
Incident power (W)	18
Duty factor	20%
Cycles per burst	50
Treatment time (s)	80
Temperature	6-9°C
Water level	10
Sample volume	15µl
E220 – Intensifier	Yes
Container	MicroTUBE-15

(c)

Parameter	Value
Incident power (W)	18
Duty factor	20%
Cycles per burst	50
Treatment time (s)	64
Temperature	6-9°C
Water level	10
Sample volume	15µl
E220 – Intensifier	Yes
Container	MicroTUBE-15

(d)

Figure C.1: A comprehensive list of the shearing conditions tested, for the optimal shearing of lDNA and preservation of sDNA.

(This figure continues on the following pages.) A total of 39 conditions were trialled on a Covaris E220 Evolution sonicator. The main variables changed during testing were the incident power, the treatment time, and the duty factor.

Parameter	Value
Incident power (W)	18
Duty factor	20%
Cycles per burst	50
Treatment time (s)	48
Temperature	6-9°C
Water level	10
Sample volume	15µl
E220 – Intensifier	Yes
Container	MicroTUBE-15

(e)

Parameter	Value
Incident power (W)	18
Duty factor	20%
Cycles per burst	50
Treatment time (s)	31
Temperature	6-9°C
Water level	10
Sample volume	15µl
E220 – Intensifier	Yes
Container	MicroTUBE-15

(f)

Parameter	Value
Incident power (W)	18
Duty factor	20%
Cycles per burst	50
Treatment time (s)	15
Temperature	6-9°C
Water level	10
Sample volume	15µl
E220 – Intensifier	Yes
Container	MicroTUBE-15

(g)

Parameter	Value
Incident power (W)	75
Duty factor	20%
Cycles per burst	1000
Treatment time (s)	95
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(h)

Parameter	Value
Incident power (W)	75
Duty factor	20%
Cycles per burst	1000
Treatment time (s)	62
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(i)

Parameter	Value
Incident power (W)	75
Duty factor	20%
Cycles per burst	1000
Treatment time (s)	40
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(j)

Parameter	Value
Incident power (W)	75
Duty factor	10%
Cycles per burst	1000
Treatment time (s)	50
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(k)

Parameter	Value
Incident power (W)	75
Duty factor	10%
Cycles per burst	1000
Treatment time (s)	70
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(l)

Parameter	Value
Incident power (W)	75
Duty factor	20%
Cycles per burst	1000
Treatment time (s)	70
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(m)

Parameter	Value
Incident power (W)	75
Duty factor	20%
Cycles per burst	1000
Treatment time (s)	63
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(n)

Parameter	Value
Incident power (W)	75
Duty factor	20%
Cycles per burst	1000
Treatment time (s)	57
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(o)

Parameter	Value
Incident power (W)	75
Duty factor	20%
Cycles per burst	1000
Treatment time (s)	50
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(p)

Parameter	Value
Incident power (W)	50
Duty factor	20%
Cycles per burst	1000
Treatment time (s)	100
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(q)

Parameter	Value
Incident power (W)	50
Duty factor	20%
Cycles per burst	1000
Treatment time (s)	50
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(r)

Parameter	Value
Incident power (W)	37
Duty factor	20%
Cycles per burst	1000
Treatment time (s)	150
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(s)

Parameter	Value
Incident power (W)	37
Duty factor	20%
Cycles per burst	1000
Treatment time (s)	100
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(t)

Parameter	Value
Incident power (W)	37
Duty factor	20%
Cycles per burst	1000
Treatment time (s)	50
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(u)

Parameter	Value
Incident power (W)	37
Duty factor	20%
Cycles per burst	500
Treatment time (s)	50
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(v)

Parameter	Value
Incident power (W)	37
Duty factor	20%
Cycles per burst	200
Treatment time (s)	100
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(w)

Parameter	Value
Incident power (W)	25
Duty factor	20%
Cycles per burst	500
Treatment time (s)	150
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(x)

Parameter	Value
Incident power (W)	25
Duty factor	20%
Cycles per burst	200
Treatment time (s)	200
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(y)

Parameter	Value
Incident power (W)	20
Duty factor	20%
Cycles per burst	200
Treatment time (s)	200
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(z)

Parameter	Value
Incident power (W)	20
Duty factor	15%
Cycles per burst	200
Treatment time (s)	250
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(aa)

Parameter	Value
Incident power (W)	15
Duty factor	15%
Cycles per burst	200
Treatment time (s)	300
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(ab)

Parameter	Value
Incident power (W)	20
Duty factor	10%
Cycles per burst	200
Treatment time (s)	250
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(ac)

Parameter	Value
Incident power (W)	15
Duty factor	10%
Cycles per burst	200
Treatment time (s)	350
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(ad)

Parameter	Value
Incident power (W)	15
Duty factor	15%
Cycles per burst	200
Treatment time (s)	600
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(ae)

Parameter	Value
Incident power (W)	15
Duty factor	15%
Cycles per burst	200
Treatment time (s)	775
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(af)

Parameter	Value
Incident power (W)	15
Duty factor	15%
Cycles per burst	200
Treatment time (s)	950
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(ag)

Parameter	Value
Incident power (W)	15
Duty factor	15%
Cycles per burst	200
Treatment time (s)	1125
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(ah)

Parameter	Value
Incident power (W)	15
Duty factor	15%
Cycles per burst	200
Treatment time (s)	1300
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(ai)

Parameter	Value
Incident power (W)	15
Duty factor	15%
Cycles per burst	200
Treatment time (s)	300
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(aj)

Parameter	Value
Incident power (W)	15
Duty factor	15%
Cycles per burst	200
Treatment time (s)	433
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(ak)

Parameter	Value
Incident power (W)	15
Duty factor	15%
Cycles per burst	200
Treatment time (s)	567
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(al)

Parameter	Value
Incident power (W)	15
Duty factor	15%
Cycles per burst	200
Treatment time (s)	700
Temperature	6-9°C
Water level	6
Sample volume	50µl
E220 – Intensifier	Yes
Container	MicroTUBE-50

(am)

Appendix D

Table D.1: Details of the samples in the HC2 cohort, and the mean family depth sequenced from those samples.

Diagnosis	Patient	Sample	Sample type	Status
ATRT	B1	HC2-11	CSF supernatant	Sequenced
ATRT	B1	HC2-12	CSF supernatant	Sequenced
ATRT	B1	HC2-13	CSF supernatant	Sequenced
ATRT	B1	HC2-14	CSF supernatant	Sequenced
ATRT	B1	HC2-15	CSF supernatant	Sequenced
ATRT	B1	HC2-16	CSF supernatant	Sequenced
ATRT	B1	HC2-17	CSF supernatant	Sequenced
ATRT	B1	HC2-18	CSF supernatant	Sequenced
ATRT	B1	HC2-19	CSF supernatant	Sequenced
ATRT	B2	HC2-20	CSF supernatant	Sequenced
ATRT	B2	HC2-21	CSF supernatant	Sequenced
ATRT	B2	HC2-22	CSF supernatant	Sequenced
ATRT	B2	HC2-23	CSF supernatant	Sequenced
ATRT	B2	HC2-24	CSF supernatant	Sequenced
ATRT	B3	HC2-25	CSF supernatant	Sequenced
ATRT	B4	HC2-26	CSF supernatant	Sequenced
ATRT	B5	HC2-27	CSF supernatant	Sequenced
ATRT	B5	HC2-28	CSF supernatant	Sequenced
ATRT	B5	HC2-29	CSF supernatant	Sequenced
ATRT	B6	HC2-30	CSF supernatant	Sequenced
ATRT	B6	HC2-31	CSF supernatant	Sequenced
MRT	B7	HC2-32	CSF supernatant	Sequenced
MRT	B7	HC2-33	CSF supernatant	Sequenced
MRT	B7	HC2-34	CSF supernatant	No DNA
MRT	B8	HC2-35	CSF supernatant	Sequenced
MRT	B8	HC2-36	CSF supernatant	Sequenced
MRT	B8	HC2-37	CSF supernatant	Failed library prep
MRT	B9	HC2-38	CSF supernatant	Sequenced
WAM	B10	HC2-39	CSF supernatant	Sequenced
WAM	B11	HC2-40	CSF supernatant	Sequenced
WAM	B11	HC2-41	CSF supernatant	Sequenced
PA	B12	HC2-42	CSF supernatant	Sequenced
DIPG	B13	HC2-43	CSF supernatant	Sequenced
ATRT	B14	HC2-44	CSF supernatant	Sequenced

ATRT	B15	HC2-45	CSF supernatant	Sequenced
ATRT	B15	HC2-46	CSF supernatant	Sequenced
ATRT	B15	HC2-47	CSF supernatant	Sequenced
ATRT	B16	HC2-48	CSF supernatant	Sequenced
ATRT	B6	HC2-49	CSF supernatant	Sequenced
ATRT	A3	HC2-50	CSF supernatant	Sequenced
ATRT	B4	HC2-51	CSF supernatant	Sequenced
ATRT	A5	HC2-52	CSF supernatant	Sequenced
DIPG	A6	HC2-53	Plasma	Sequenced
DIPG	A7	HC2-54	Plasma	Sequenced
DIPG	A8	HC2-55	Plasma	Sequenced

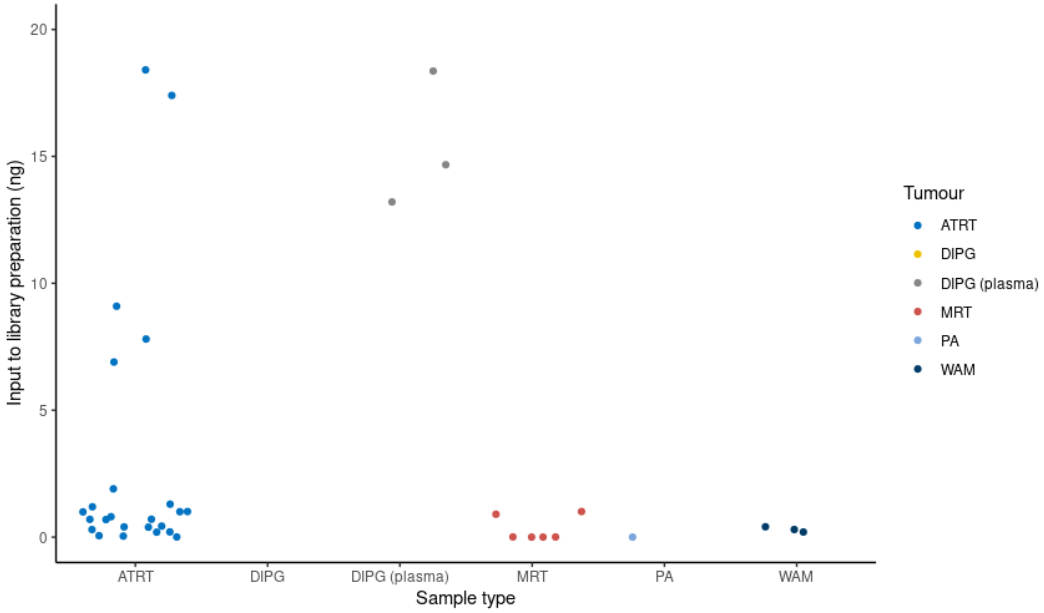


Figure D.1: DNA inputs for library preparation of the HC2 cohort.

Input amounts were calculated ddPCR results from 1µl of each sample, and the four CSF samples prepped using Bioanalyzer quantification are omitted from this figure. Unless otherwise indicated, all sample types were CSF. This figure is a zoomed version of Figure 2.6, which does not include two 100ng ATRT samples, to highlight the differences between the low input samples.

Appendix E

Table E.1: The raw read depths of the HC1 cohort's sequencing run

Sample	Raw reads overlapping FLCP-1	Mean bases per FLCP-1 position
HC1-10	260,668,195,493	301,335,146
HC1-12	7,297,118,076,487	8,435,544,416
HC1-13	1,007,907,461,013	1,165,151,511
HC1-14	3,139,966,742,920	3,629,834,223
HC1-16	880,312,910,861	1,017,650,884
HC1-17	973,141,226,847	1,124,961,383
HC1-18	1,237,798,284,892	1,430,907,696
HC1-19	478,462,903,736	553,108,095
HC1-1	52,024,291,625	60,140,622
HC1-2	1,384,955,116,200	1,601,022,524
HC1-3	46,341,586,031	53,571,356
HC1-4	7,321,597,006,139	8,463,842,314
HC1-6	1,049,952,963,202	1,213,756,549
HC1-7	824,505,697,425	953,137,164
HC1-8	338,905,852,974	391,778,692
HC1-9	259,316,562,154	299,772,644

Table E.2: The raw read depths of the HC2 cohort's sequencing run

Sample	Raw read depth	Mean read depth per base of FLCP-1
HC2-11	734,169,800	6,240
HC2-12	139,100,000	1,182
HC2-13	139,100,000	1,182
HC2-14	139,100,000	1,182
HC2-15	139,100,000	1,182
HC2-16	2,016,950,000	17,144
HC2-17	139,100,000	1,182
HC2-18	1,351,356,500	11,487
HC2-19	8,067,800,000	68,577
HC2-20	139,100,000	1,182
HC2-21	139,100,000	1,182
HC2-22	139,100,000	1,182
HC2-23	139,100,000	1,182
HC2-24	139,100,000	1,182
HC2-25	139,100,000	1,182
HC2-26	139,100,000	1,182
HC2-27	139,100,000	1,182
HC2-28	2,492,950,200	21,190
HC2-29	1,403,797,200	11,932
HC2-30	139,100,000	1,182
HC2-31	153,288,200	1,303
HC2-32	139,100,000	1,182
HC2-33	139,100,000	1,182
HC2-35	139,100,000	1,182
HC2-36	139,100,000	1,182
HC2-38	139,100,000	1,182
HC2-39	139,100,000	1,182
HC2-40	139,100,000	1,182
HC2-41	139,100,000	1,182
HC2-42	139,100,000	1,182
HC2-43	139,100,000	1,182
HC2-44	139,100,000	1,182
HC2-45	8,067,800,000	68,577
HC2-46	1,484,475,200	12,618
HC2-47	139,100,000	1,182
HC2-48	629,288,400	5,349
HC2-49	556,678,200	4,732
HC2-50	139,100,000	1,182
HC2-51	139,100,000	1,182
HC2-52	139,100,000	1,182
HC2-53	493,749,360	4,197
HC2-54	354,983,200	3,017
HC2-55	394,515,420	3,353

Acronyms

EGFR Epidermal Growth Factor Receptor

σ standard deviation

ABMS Average Best Match Score

ACP Adamantinomatous Craniopharyngioma

ATRT Atypical Teratoid/Rhabdoid Tumour

B-ALL B-cell Precursor Acute Lymphoblastic Leukemia

C19MC Chromosome 19 microRNA Cluster

cfDNA cell-free DNA

CNS Central Nervous System

CNV Copy-Number Variation

CSF cerebrospinal fluid

CT Computerized Tomography X-ray scan

CTC Circulating Tumour Cell

ctDNA circulating tumour DNA

ddPCR droplet digital Polymerase Chain Reaction

DIPG Diffuse Intrinsic Pontine Glioma

FDA Food and Drug Administration

FISH Fluorescence *in situ* Hybridisation

FLCP-1 Forshew Lab Capture Panel 1

GOSH Great Ormond Street Hospital

HC1 HiSeq Capture 1

HC2 HiSeq Capture 2

HCC Hepatocellular Carcinoma

HD Hamming Distance

HPC High Performance Computing

ICH Institute of Child Health

IFN-α Interferon-α

IGV Integrated Genomics Viewer

IHC Immunohistochemistry

InDel Insertions/Deletions

LOH Loss of Heterozygosity

MIP Molecular Inversion Probe

MLPA Multiplex Ligation-Dependent Probe Amplification

MRI Magnetic Resonance Imaging

MRT Malignant Rhabdoid Tumour

NGS Next-Generation Sequencing

NHS National Health Service

PA Pilocytic Astrocytoma

PBT Paediatric Brain Tumour

PCR Polymerase Chain Reaction

QC Quality Check/Control

qPCR quantitative Polymerase Chain Reaction

RTPS Rhabdoid Tumor Predisposition Syndrome

SNAFU Single Nucleotide Alteration Filtering Utility

SNP Single Nucleotide Polymorphism

SNV Single Nucleotide Variant

VAF Variant Allele Frequency

WAM *WNT*-Activated Medulloblastoma

WGS Whole Genome Sequencing

Bibliography

[1] L. A. Diaz and A. Bardelli. Liquid biopsies: genotyping circulating tumor DNA. *J Clin Oncol*, 32(6):579–586, Feb 2014.

[2] E Crowley, Nicolantonio F Di, F Loupakis, and A Bardelli. Liquid biopsy: monitoring cancer-genetics in the blood. *Nat Rev Clin Oncol*, 10(8):472-84., 2013.

[3] E Heitzer, P Ulz, and JB Geigl. Circulating tumor DNA as a liquid biopsy for cancer. *Clin Chem*, 61(1):112-23., 2015.

[4] M Offin, JJ Chabon, P Razavi, JM Isbell, CM Rudin, M Diehn, and BT Li. Capturing Genomic Evolution of Lung Cancers through Liquid Biopsy for Circulating Tumor DNA. *J Oncol*, 2017:4517834., 2017.

[5] A Ono, A Fujimoto, Y Yamamoto, S Akamatsu, N Hiraga, M Imamura, T Kawaoka, M Tsuge, H Abe, CN Hayes, D Miki, M Furuta, T Tsunoda, S Miyano, M Kubo, H Aikata, H Ochi, YI Kawakami, K Arihiro, H Ohdan, H Nakagawa, and K Chayama. Circulating Tumor DNA Analysis for Liver Cancers and Its Usefulness as a Liquid Biopsy. *Cell Mol Gastroenterol Hepatol*, 1(5):516-534., 2015.

[6] J Wang and C Bettegowda. Applications of DNA-Based Liquid Biopsy for Central Nervous System Neoplasms. *J Mol Diagn*, 19(1):24-34., 2017.

[7] MR Voisin, C Ovenden, DS Tsang, AA Gupta, A Huang, AF Gao, P Diamandis, JP Almeida, and F Gentili. Atypical Teratoid/Rhabdoid Sellar Tumor in an Adult with a Familial History of a Germline SMARCB1 Mutation: Case Report and Review of the Literature. *World Neurosurg*, 127:336-345., 2019.

[8] A Gajjar, M Fouladi, AW Walter, SJ Thompson, DA Reardon, TE Merchant, JJ Jenkins, A Liu, JM Boyett, LE Kun, and RL Heideman. Comparison of lumbar and shunt cerebrospinal fluid specimens for cytologic detection of leptomeningeal disease in pediatric patients with brain tumors. *J Clin Oncol*, 17(6):1825-8., 1999.

[9] SYY Low, CM Wei, KTE Chang, CY Huak, NL Ping, SW Tew, and DCY Low. Intra-operative cerebrospinal fluid sampling versus post-operative lumbar puncture for detection of leptomeningeal disease in malignant paediatric brain tumours. *PLoS One*, 13(5):e0196696., 2018.

[10] SR Kennedy, MW Schmitt, EJ Fox, BF Kohrn, JJ Salk, EH Ahn, MJ Prindle, KJ Kuong, JC Shen, RA Risques, and LA Loeb. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc*, 9(11):2586-606., 2014.

[11] D. B. Sloan, A. K. Broz, J. Sharbrough, and Z. Wu. Detecting Rare Mutations and DNA Damage with Sequencing-Based Methods. *Trends Biotechnol*, 36(7):729–740, 07 2018.

[12] F Inzani, G Petrone, and G Rindi. The New World Health Organization Classification for Pancreatic Neuroendocrine Neoplasia. *Endocrinol Metab Clin North Am*, 47(3):463-470., 2018.

[13] CDM Fletcher, KK Unni, and F Mertens. *Pathology and genetics of tumours of soft tissue and bone*. IARC Press, Lyon, 2002. 14–16 pp.

[14] LA Doyle. Sarcoma classification: an update based on the 2013 World Health Organization Classification of Tumors of Soft Tissue and Bone. *Cancer*, 120(12):1763-74., 2014.

[15] DN Louis, H Ohgaki, OD Wiestler, WK Cavenee, PC Burger, A Jouvet, BW Scheithauer, and P Kleihues. The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol*, 114(2):97-109., 2007.

[16] DN Louis, A Perry, G Reifenberger, Deimling A von, D Figarella-Branger, WK Cavenee, H Ohgaki, OD Wiestler, P Kleihues, and DW Ellison. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol*, 131(6):803-20., 2016.

[17] C Dardis, J Yeo, K Milton, LS Ashby, KA Smith, S Mehta, E Youssef, J Eschbacher, K Tucker, L Dawes, N Lambie, E Algar, and E Hovey. Atypical Teratoid Rhabdoid Tumor: Two Case Reports and an Analysis of Adult Cases with Implications for Pathophysiology and Treatment. *Front Neurol*, 8:247., 2017.

[18] C. A. Stiller, M. E. Kroll, P. J. Boyle, and Z. Feng. Population mixing, socioeconomic status and incidence of childhood acute lymphoblastic leukaemia in England and Wales: analysis by census ward. *Br J Cancer*, 98(5):1006–1011, Mar 2008.

[19] C. Schwab and C. J. Harrison. Advances in B-cell Precursor Acute Lymphoblastic Leukemia Genomics. *Hemasphere*, 2(4):e53, 08 2018.

[20] A. C. Winters and K. M. Bernt. MLL-Rearranged Leukemias-An Update on Science and Clinical Approaches. *Front Pediatr*, 5:4, 2017.

[21] ES Kim, RS Herbst, II Wistuba, JJ Lee, GR Jr Blumenschein, A Tsao, DJ Stewart, ME Hicks, J Jr Erasmus, S Gupta, CM Alden, S Liu, X Tang, FR Khuri, HT Tran, BE Johnson, JV Heymach, L Mao, F Fossella, MS Kies, V Papadimitrakopoulou, SE Davis, SM Lippman, and WK Hong. The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discov*, 1(1):44-53., 2011.

[22] JJ Tao, AM Schram, and DM Hyman. Basket Studies: Redefining Clinical Trials in the Era of Genome-Driven Oncology. *Annu Rev Med*, 69:319-331., 2018.

[23] J. J. H. Park, G. Hsu, E. G. Siden, K. Thorlund, and E. J. Mills. An overview of precision oncology basket and umbrella trials for clinicians. *CA Cancer J Clin*, 70 (2):125–137, 03 2020.

[24] J. J. H. Park, E. Siden, M. J. Zoratti, L. Dron, O. Harari, J. Singer, R. T. Lester, K. Thorlund, and E. J. Mills. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. *Trials*, 20(1):572, Sep 2019.

[25] M. W. Redman and C. J. Allegra. The Master Protocol Concept. *Semin Oncol*, 42 (5):724–730, Oct 2015.

[26] US National Library of Medicine. A Study of SNDX-5613 in R/R Leukemias Including Those With an MLLr/KMT2A Gene Rearrangement or NPM1 Mutation (AUGMENT-101), 2019. `https://clinicaltrials.gov/ct2/show/NCT04065399` [accessed: 2021-06-19].

[27] M. G. Kris, R. B. Natale, R. S. Herbst, T. J. Lynch, D. Prager, C. P. Belani, J. H. Schiller, K. Kelly, H. Spiridonidis, A. Sandler, K. S. Albain, D. Cella, M. K. Wolf, S. D. Averbuch, J. J. Ochs, and A. C. Kay. Efficacy of gefitinib, an inhibitor of the

epidermal growth factor receptor tyrosine kinase, in symptomatic patients with non-small cell lung cancer: a randomized trial. *JAMA*, 290(16):2149–2158, Oct 2003.

[28] J. J. H. Park, O. Harari, L. Dron, R. T. Lester, K. Thorlund, and E. J. Mills. An overview of platform trials with a checklist for clinical readers. *J Clin Epidemiol*, 125:1–8, 09 2020.

[29] B. R. Saville and S. M. Berry. Efficiencies of platform clinical trials: A vision of the future. *Clin Trials*, 13(3):358–366, 06 2016.

[30] H. Kimura, T. Ohira, O. Uchida, J. Matsubayashi, S. Shimizu, T. Nagao, N. Ikeda, and K. Nishio. Analytical performance of the cobas EGFR mutation assay for Japanese non-small-cell lung cancer. *Lung Cancer*, 83(3):329–333, Mar 2014.

[31] L. R. Rowe, B. G. Bentz, and J. S. Bentz. Detection of BRAF V600E activating mutation in papillary thyroid carcinoma using PCR with allele-specific fluorescent probe melting curve analysis. *J Clin Pathol*, 60(11):1211–1215, Nov 2007.

[32] M Nagahashi, Y Shimada, H Ichikawa, H Kameyama, K Takabe, S Okuda, and T Wakai. Next generation sequencing-based gene panel tests for the management of solid tumors. *Cancer Sci*, 110(1):6-15., 2019.

[33] A Gajjar, DC Bowers, MA Karajannis, S Leary, H Witt, and NG Gottardo. Pediatric Brain Tumors: Innovative Genomic Information Is Transforming the Diagnostic and Clinical Landscape. *J Clin Oncol*, 33(27):2986-98., 2015.

[34] H. Davies, G. R. Bignell, C. Cox, P. Stephens, S. Edkins, S. Clegg, J. Teague, H. Woffendin, M. J. Garnett, W. Bottomley, N. Davis, E. Dicks, R. Ewing, Y. Floyd, K. Gray, S. Hall, R. Hawes, J. Hughes, V. Kosmidou, A. Menzies, C. Mould, A. Parker, C. Stevens, S. Watt, S. Hooper, R. Wilson, H. Jayatilake, B. A. Gusterson, C. Cooper, J. Shipley, D. Hargrave, K. Pritchard-Jones, N. Maitland, G. Chenevix-Trench, G. J. Riggins, D. D. Bigner, G. Palmieri, A. Cossu, A. Flanagan, A. Nicholson, J. W. Ho, S. Y. Leung, S. T. Yuen, B. L. Weber, H. F. Seigler, T. L. Darrow, H. Paterson, R. Marais, C. J. Marshall, R. Wooster, M. R. Stratton, and P. A. Futreal. Mutations of the BRAF gene in human cancer. *Nature*, 417(6892):949–954, Jun 2002.

[35] R. Kamps, R. D. Brandão, B. J. Bosch, A. D. Paulussen, S. Xanthoulea, M. J. Blok, and A. Romano. Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification. *Int J Mol Sci*, 18(2), Jan 2017.

[36] Food and Drug Administration. FDA announces approval, CMS proposes coverage of first breakthrough-designated test to detect extensive number of cancer biomarkers, 2017. `https://www.fda.gov/news-events/press-announcements/fda-announces-approval-cms-proposes-coverage-first-breakthrough-designated-test-detect-extensive` [accessed: 2020-04-23].

[37] Food and Drug Administration. Nucleic Acid Based Tests, 2020. `https://www.fda.gov/medical-devices/vitro-diagnostics/nucleic-acid-based-tests` [accessed: 2020-04-27].

[38] Food and Drug Administration. Oncomine Dx Target Test Approval Order, 2020. `http://www.accessdata.fda.gov/cdrh_docs/pdf16/P160045S019A.pdf` [accessed: 2021-06-18].

[39] NHS England. National Genomic Test Directory, 2020. `https://www.england.nhs.uk/publication/national-genomic-test-directories/` [accessed: 2021-06-18].

[40] RA Burrell, N McGranahan, J Bartek, and C Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338-45., 2013.

[41] MS Lawrence, P Stojanov, P Polak, GV Kryukov, K Cibulskis, A Sivachenko, SL Carter, C Stewart, CH Mermel, SA Roberts, A Kiezun, PS Hammerman, A McKenna, Y Drier, L Zou, AH Ramos, TJ Pugh, N Stransky, E Helman, J Kim, C Sougnez, L Ambrogio, E Nickerson, E Shefler, ML Cortés, D Auclair, G Saksena, D Voet, M Noble, D DiCara, P Lin, L Lichtenstein, DI Heiman, T Fennell, M Imielinski, B Hernandez, E Hodis, S Baca, AM Dulak, J Lohr, DA Landau, CJ Wu, J Melendez-Zajgla, A Hidalgo-Miranda, A Koren, SA McCarroll, J Mora, B Crompton, R Onofrio, M Parkin, W Winckler, K Ardlie, SB Gabriel, CWM Roberts, JA Biegel, K Stegmaier, AJ Bass, LA Garraway, M Meyerson, TR Golub,

DA Gordenin, S Sunyaev, ES Lander, and G Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214-218., 2013.

[42] MC King, E Levy-Lahad, and A Lahad. Population-based screening for BRCA1 and BRCA2: 2014 Lasker Award. *JAMA*, 312(11):1091-2., 2014.

[43] JG Paez, PA Jänne, JC Lee, S Tracy, H Greulich, S Gabriel, P Herman, FJ Kaye, N Lindeman, TJ Boggon, K Naoki, H Sasaki, Y Fujii, MJ Eck, WR Sellers, BE Johnson, and M Meyerson. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, 304(5676):1497-500., 2004.

[44] F Barlesi, J Mazieres, JP Merlio, D Debieuvre, J Mosser, H Lena, L Ouafik, B Besse, I Rouquette, V Westeel, F Escande, I Monnet, A Lemoine, R Veillon, H Blons, C Audigier-Valette, PP Bringuier, R Lamy, M Beau-Faller, JL Pujol, JC Sabourin, F Penault-Llorca, MG Denis, S Lantuejoul, F Morin, Q Tran, P Missy, A Langlais, B Milleron, J Cadranel, JC Soria, and G Zalcman. Routine molecular profiling of patients with advanced non-small-cell lung cancer: results of a 1-year nationwide programme of the French Cooperative Thoracic Intergroup (IFCT). *Lancet*, 387(10026):1415-26., 2016.

[45] AH Tsang, KH Cheng, AS Wong, SS Ng, BB Ma, CM Chan, NB Tsui, LW Chan, BY Yung, and SC Wong. Current and future molecular diagnostics in colorectal cancer and colorectal adenoma. *World J Gastroenterol*, 20(14):3847-57., 2014.

[46] R Govindan. Cancer. Attack of the clones. *Science*, 346(6206):169-70., 2014.

[47] RB Lanman, SA Mortimer, OA Zill, D Sebisanovic, R Lopez, S Blau, EA Collisson, SG Divers, DS Hoon, ES Kopetz, J Lee, PG Nikolinakos, AM Baca, BG Kermani, H Eltoukhy, and A Talasaz. Analytical and Clinical Validation of a Digital Sequencing Panel for Quantitative, Highly Accurate Evaluation of Cell-Free Circulating Tumor DNA. *PLoS One*, 10(10):e0140712., 2015.

[48] HP Ellis, M Greenslade, B Powell, I Spiteri, A Sottoriva, and KM Kurian. Current Challenges in Glioblastoma: Intratumour Heterogeneity, Residual Disease, and Models to Predict Disease Recurrence. *Front Oncol*, 5:251., 2015.

[49] E Ziv, JC Durack, and SB Solomon. The Importance of Biopsy in the Era of Molecular Medicine. *Cancer J*, 22(6):418-422., 2016.

[50] R. M. Seliem, M. W. Assaad, S. J. Gorombey, L. A. Moral, J. R. Kirkwood, and C. N. Otis. Fine-needle aspiration biopsy of the central nervous system performed freehand under computed tomography guidance without stereotactic instrumentation. *Cancer*, 99(5):277–284, Oct 2003.

[51] E. Crowley, F. Di Nicolantonio, F. Loupakis, and A. Bardelli. Liquid biopsy: monitoring cancer-genetics in the blood. *Nat Rev Clin Oncol*, 10(8):472–484, Aug 2013.

[52] Hui L., Maron J.L., and Gahan P.B. Other body fluids as non-invasive sources of cell-free dna/rna. In P.B. Gahan, editor, *Circulating Nucleic Acids in Early Diagnosis, Prognosis and Treatment Monitoring.*, chapter 11, pages 295–323. Springer, Dordrecht, 2015.

[53] P. B. Larrabee, K. L. Johnson, E. Pestova, M. Lucas, K. Wilber, E. S. LeShane, U. Tantravahi, J. M. Cowan, and D. W. Bianchi. Microarray analysis of cell-free fetal DNA in amniotic fluid: a prenatal molecular karyotype. *Am J Hum Genet*, 75 (3):485–491, Sep 2004.

[54] Y. H. Su, M. Wang, T. M. Block, O. Landt, I. Botezatu, O. Serdyuk, A. Lichtenstein, H. Melkonyan, L. D. Tomei, and S. Umansky. Transrenal DNA as a diagnostic tool: important technical notes. *Ann N Y Acad Sci*, 1022:81–89, Jun 2004.

[55] L De Mattos-Arruda, R Mayor, CK Ng, B Weigelt, F Martínez-Ricarte, D Torrejon, M Oliveira, A Arias, C Raventos, J Tang, E Guerini-Rocco, E Martínez-Sáez, S Lois, O Marín, X de la Cruz, S Piscuoglio, R Towers, A Vivancos, V Peg, S Ramon y Cajal, J Carles, J Rodon, M González-Cao, J Tabernero, E Felip, J Sahuquillo, MF Berger, J Cortes, JS Reis-Filho, and J Seoane. Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. *Nat Commun*, 6:8839., 2015.

[56] S. Jahr, H. Hentze, S. Englisch, D. Hardt, F. O. Fackelmayer, R. D. Hesch, and R. Knippers. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res*, 61(4): 1659–1665, Feb 2001.

[57] F Diehl, M Li, D Dressman, Y He, D Shen, S Szabo, LA Jr Diaz, SN Goodman, KA David, H Juhl, KW Kinzler, and B Vogelstein. Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc Natl Acad Sci U S A*, 102(45):16368-73., 2005.

[58] M Ivanov, A Baranova, T Butler, P Spellman, and V Mileyko. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics*, 13:S1., 2015.

[59] F Mouliere, B Robert, Peyrotte E Arnau, Rio M Del, M Ychou, F Molina, C Gongora, and AR Thierry. High fragmentation characterizes tumour-derived circulating DNA. *PLoS One*, 6(9):e23418., 2011.

[60] Y. Y. Lui, K. W. Chik, R. W. Chiu, C. Y. Ho, C. W. Lam, and Y. M. Lo. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clin Chem*, 48(3):421–427, Mar 2002.

[61] S. A. Leon, B. Shapiro, D. M. Sklaroff, and M. J. Yaros. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res*, 37(3):646–650, Mar 1977.

[62] A. Kustanovich, R. Schwartz, T. Peretz, and A. Grinshpun. Life and death of circulating cell-free DNA. *Cancer Biol Ther*, 20(8):1057–1067, 2019.

[63] M. Fleischhacker and B. Schmidt. Circulating nucleic acids (CNAs) and cancer–a survey. *Biochim Biophys Acta*, 1775(1):181–232, Jan 2007.

[64] D Chandrananda, NP Thorne, and M Bahlo. High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA. *BMC Med Genomics*, 8:29., 2015.

[65] IH Wong, YM Lo, J Zhang, CT Liew, MH Ng, N Wong, PB Lai, WY Lau, NM Hjelm, and PJ Johnson. Detection of aberrant p16 methylation in the plasma and serum of liver cancer patients. *Cancer Res*, 59(1):71-3., 1999.

[66] K. L. Reckamp, V. O. Melnikova, C. Karlovich, L. V. Sequist, D. R. Camidge, H. Wakelee, M. Perol, G. R. Oxnard, K. Kosco, P. Croucher, E. Samuelsz, C. R. Vibat, S. Guerrero, J. Geis, D. Berz, E. Mann, S. Matheny, L. Rolfe, M. Raponi, M. G. Erlander, and S. Gadgeel. A Highly Sensitive and Quantitative Test Platform

for Detection of NSCLC EGFR Mutations in Urine and Plasma. *J Thorac Oncol*, 11 (10):1690–1700, Oct 2016.

[67] Y Wang, S Springer, M Zhang, KW McMahon, I Kinde, L Dobbyn, J Ptak, H Brem, K Chaichana, GL Gallia, ZL Gokaslan, ML Groves, GI Jallo, M Lim, A Olivi, A Quinones-Hinojosa, D Rigamonti, GJ Riggins, DM Sciubba, JD Weingart, JP Wolinsky, X Ye, SM Oba-Shinjo, SK Marie, M Holdhoff, N Agrawal, LA Jr Diaz, N Papadopoulos, KW Kinzler, B Vogelstein, and C Bettegowda. Detection of tumor-derived DNA in cerebrospinal fluid of patients with primary tumors of the brain and spinal cord. *Proc Natl Acad Sci U S A*, 112(31):9704-9., 2015.

[68] T Forshew, M Murtaza, C Parkinson, D Gale, DW Tsui, F Kaper, SJ Dawson, AM Piskorz, M Jimenez-Linan, D Bentley, J Hadfield, AP May, C Caldas, JD Brenton, and N Rosenfeld. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci Transl Med*, 4(136):136ra68., 2012.

[69] F Mouliere and N Rosenfeld. Circulating tumor-derived DNA is shorter than somatic DNA in plasma. *Proc Natl Acad Sci USA*, 112(11):3178-9., 2015.

[70] HR Underhill, JO Kitzman, S Hellwig, NC Welker, R Daza, DN Baker, KM Gligorich, RC Rostomily, MP Bronner, and J Shendure. Fragment Length of Circulating Tumor DNA. *PLoS Genet*, 12(7):e1006162., 2016.

[71] F Mouliere, R Mair, D Chandrananda, F Marass, CG Smith, J Su, J Morris, C Watts, KM Brindle, and N Rosenfeld. Detection of cell-free DNA fragmentation and copy number alterations in cerebrospinal fluid from glioma patients. *EMBO Mol Med*, 10(12):e9323., 2018.

[72] G. Siravegna, S. Marsoni, S. Siena, and A. Bardelli. Integrating liquid biopsies into the management of cancer. *Nat Rev Clin Oncol*, 14(9):531–548, Sep 2017.

[73] G. Poulet, J. Massias, and V. Taly. Liquid Biopsy: General Concepts. *Acta Cytol*, 63 (6):449–455, 2019.

[74] A. Di Meo, J. Bartlett, Y. Cheng, M. D. Pasic, and G. M. Yousef. Liquid biopsy: a step forward towards precision medicine in urologic malignancies. *Mol Cancer*, 16 (1):80, 04 2017.

[75] R. Malani, M. Fleisher, P. Kumthekar, X. Lin, A. Omuro, M. D. Groves, N. U. Lin, M. Melisko, A. B. Lassman, S. Jeyapalan, A. Seidman, A. Skakodub, A. Boire, L. M. DeAngelis, M. Rosenblum, J. Raizer, and E. Pentsova. Cerebrospinal fluid circulating tumor cells as a quantifiable measurement of leptomeningeal metastases in patients with HER2 positive cancer. *J Neurooncol*, 148(3):599–606, Jul 2020.

[76] C. Alix-Panabières and K. Pantel. Challenges in circulating tumour cell research. *Nat Rev Cancer*, 14(9):623–631, 09 2014.

[77] L. M. Millner, M. W. Linder, and R. Valdes. Circulating tumor cells: a review of present methods and the need to identify heterogeneous phenotypes. *Ann Clin Lab Sci*, 43(3):295–304, 2013.

[78] W. J. Allard, J. Matera, M. C. Miller, M. Repollet, M. C. Connelly, C. Rao, A. G. Tibbe, J. W. Uhr, and L. W. Terstappen. Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. *Clin Cancer Res*, 10(20):6897–6904, Oct 2004.

[79] L. Dizdar, G. Fluegen, G. van Dalum, E. Honisch, R. P. Neves, D. Niederacher, H. Neubauer, T. Fehm, A. Rehders, A. Krieg, W. T. Knoefel, and N. H. Stoecklein. Detection of circulating tumor cells in colorectal cancer patients using the GILUPI CellCollector: results from a prospective, single-center study. *Mol Oncol*, 13(7): 1548–1558, 07 2019.

[80] F. C. Bidard, C. Proudhon, and J. Y. Pierga. Circulating tumor cells in breast cancer. *Mol Oncol*, 10(3):418–430, Mar 2016.

[81] V. Maly, O. Maly, K. Kolostova, and V. Bobek. Circulating Tumor Cells in Diagnosis and Treatment of Lung Cancer. *In Vivo*, 33(4):1027–1037, 2019.

[82] M. Torre, E. Q. Lee, U. N. Chukwueke, L. Nayak, E. S. Cibas, and A. C. Lowe. Integration of rare cell capture technology into cytologic evaluation of cerebrospinal fluid specimens from patients with solid tumors and suspected leptomeningeal metastasis. *J Am Soc Cytopathol*, 9(1):45–54, 2020.

[83] M. T. J. van Bussel, D. Pluim, B. Milojkovic Kerklaan, M. Bol, K. Sikorska, D. T. C. Linders, D. van den Broek, J. H. Beijnen, J. H. M. Schellens, and D. Brandsma. Cir-

culating epithelial tumor cell analysis in CSF in patients with leptomeningeal metastases. *Neurology*, 94(5):e521–e528, 02 2020.

[84] S. Yanagisawa, I. Kadouchi, K. Yokomori, M. Hirose, M. Hakozaki, H. Hojo, K. Maeda, E. Kobayashi, and T. Murakami. Identification and metastatic potential of tumor-initiating cells in malignant rhabdoid tumor of the kidney. *Clin Cancer Res*, 15(9):3014–3022, May 2009.

[85] Immunicon Corporation. Immunicon Corporation Announces FDA Clearance of the CellSearch Circulating Tumor Cell Kit for Monitoring Patients with Metastatic Prostate Cancer , 2008. `https://www.sec.gov/Archives/edgar/data/1083132/000119312508041224/dex991.htm` [accessed: 2021-06-19].

[86] Food and Drug Administration. 510(K) SUMMARY, 2009. `https://www.accessdata.fda.gov/cdrh_docs/pdf7/k073338.pdf` [accessed: 2021-06-19].

[87] Food and Drug Administration. The therascreen PIK3CA RGQ PCR Kit - P190001 and P190004, 2019. `https://www.fda.gov/medical-devices/recently-approved-devices/therascreen-pik3ca-rgq-pcr-kit-p190001-and-p190004` [accessed: 2021-06-18].

[88] Food and Drug Administration. cobas EGFR Mutation Test v2, 2016. `https://www.fda.gov/drugs/resources-information-approved-drugs/cobas-egfr-mutation-test-v2` [accessed: 2021-06-18].

[89] S. Heeke, J. Benzaquen, V. Hofman, M. Ilié, M. Allegra, E. Long-Mira, S. Lassalle, V. Tanga, C. Salacroup, C. Bonnetaud, J. Fayada, L. Gazoppi, L. Ribeyre, O. Castelnau, G. Garnier, F. Cattet, I. Nanni, F. de Fraipont, C. Cohen, J. P. Berthet, S. Leroy, M. Poudenx, C. H. Marquette, M. G. Denis, F. Barlesi, and P. Hofman. Critical Assessment in Routine Clinical Practice of Liquid Biopsy for EGFR Status Testing in Non-Small-Cell Lung Cancer: A Single-Laboratory Experience (LPCE, Nice, France). *Clin Lung Cancer*, 21(1):56–65, 01 2020.

[90] A. G. Sacher, C. Paweletz, S. E. Dahlberg, R. S. Alden, A. O'Connell, N. Feeney, S. L. Mach, P. A. Jänne, and G. R. Oxnard. Prospective Validation of Rapid Plasma

Genotyping for the Detection of EGFR and KRAS Mutations in Advanced Lung Cancer. *JAMA Oncol*, 2(8):1014–1022, Aug 2016.

[91] Bio-Rad Laboratories Inc. Bio-Rad Releases First FDA-Cleared Digital PCR System and Test for Monitoring Chronic Myeloid Leukemia Treatment Response, 2019. https://www.bio-rad.com/en-uk/life-science-research/news/bio-rad-releases-first-fda-cleared-digital-pcr-system-test-for-monitoring-chronic-myeloid-leukemia-treatment-response?vertical=LSR&ID=Bio-Rad-Releases-Fir_1550257994 [accessed: 2021-06-18].

[92] Food and Drug Administration. FDA approves liquid biopsy NGS companion diagnostic test for multiple cancers and biomarkers, 2020. https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-liquid-biopsy-ngs-companion-diagnostic-test-multiple-cancers-and-biomarkers [accessed: 2021-06-18].

[93] I Garcia-Murillas, G Schiavon, B Weigelt, C Ng, S Hrebien, RJ Cutts, M Cheang, P Osin, A Nerurkar, I Kozarewa, JA Garrido, M Dowsett, JS Reis-Filho, IE Smith, and NC Turner. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci Transl Med*, 7(302):302ra133., 2015.

[94] V Rimelen, G Ahle, E Pencreach, N Zinniger, A Debliquis, L Zalmaï, I Harzallah, R Hurstel, I Alamome, F Lamy, J Voirin, and B Drénou. Tumor cell-free DNA detection in CSF for primary CNS lymphoma diagnosis. *Acta Neuropathol Commun*, 7(1):43., 2019.

[95] R. Gatenby and J. Brown. The Evolution and Ecology of Resistance in Cancer Therapy. *Cold Spring Harb Perspect Med*, 8(3), 03 2018.

[96] F. Michor, M. A. Nowak, and Y. Iwasa. Evolution of resistance to cancer therapy. *Curr Pharm Des*, 12(3):261–271, 2006.

[97] R. A. Gatenby and J. S. Brown. Integrating evolutionary dynamics into cancer therapy. *Nat Rev Clin Oncol*, 17(11):675–686, 11 2020.

[98] P Fitzmorris and AK Singal. Surveillance and Diagnosis of Hepatocellular Carcinoma. *Gastroenterol Hepatol (N Y)*, 11(1):38-46., 2015.

[99] S Mittal and HB El-Serag. Epidemiology of hepatocellular carcinoma: consider the population. *J Clin Gastroenterol*, 47Suppl:S2-6., 2013.

[100] CL Kao, SH Chiou, DM Ho, YJ Chen, RS Liu, CW Lo, FT Tsai, CH Lin, HH Ku, SM Yu, and TT Wong. Elevation of plasma and cerebrospinal fluid osteopontin levels in patients with atypical teratoid/rhabdoid tumor. *Am J Clin Pathol*, 123(2):297-304., 2005.

[101] G. Mor, I. Visintin, Y. Lai, H. Zhao, P. Schwartz, T. Rutherford, L. Yue, P. Bray-Ward, and D. C. Ward. Serum protein markers for early detection of ovarian cancer. *Proc Natl Acad Sci U S A*, 102(21):7677–7682, May 2005.

[102] Y. M. Lo, K. C. Chan, H. Sun, E. Z. Chen, P. Jiang, F. M. Lun, Y. W. Zheng, T. Y. Leung, T. K. Lau, C. R. Cantor, and R. W. Chiu. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med*, 2(61):61ra91, Dec 2010.

[103] S. Mader and K. Pantel. Liquid Biopsy: Current Status and Future Perspectives. *Oncol Res Treat*, 40(7-8):404–408, 2017.

[104] O. A. Zill, K. C. Banks, S. R. Fairclough, S. A. Mortimer, J. V. Vowles, R. Mokhtari, D. R. Gandara, P. C. Mack, J. I. Odegaard, R. J. Nagy, A. M. Baca, H. Eltoukhy, D. I. Chudova, R. B. Lanman, and A. Talasaz. The Landscape of Actionable Genomic Alterations in Cell-Free Circulating Tumor DNA from 21,807 Advanced Cancer Patients. *Clin Cancer Res*, 24(15):3528–3538, 08 2018.

[105] JCM Wan, C Massie, J Garcia-Corbacho, F Mouliere, JD Brenton, C Caldas, S Pacey, R Baird, and N Rosenfeld. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer*, 17(4):223-238., 2017.

[106] R. B. Shah, J. Bentley, Z. Jeffery, and A. M. DeMarzo. Heterogeneity of PTEN and ERG expression in prostate cancer on core needle biopsies: implications for cancer risk stratification and biomarker sampling. *Hum Pathol*, 46(5):698–706, May 2015.

[107] M Jamal-Hanjani, GA Wilson, S Horswell, R Mitter, O Sakarya, T Constantin, R Salari, E Kirkizlar, S Sigurjonsson, R Pelham, S Kareht, B Zimmermann, and C Swanton. Detection of ubiquitous and heterogeneous mutations in cell-free DNA

from patients with early-stage non-small-cell lung cancer. *Ann Oncol*, 27(5):862-7., 2016.

[108] P Adamo, CM Cowley, CP Neal, V Mistry, K Page, AR Dennison, J Isherwood, R Hastings, J Luo, DA Moore, PJ Howard, ML Miguel, C Pritchard, M Manson, and JA Shaw. Profiling tumour heterogeneity through circulating tumour DNA in patients with pancreatic cancer. *Oncotarget*, 8(50):87221-87233., 2017.

[109] C Bettegowda, M Sausen, RJ Leary, I Kinde, Y Wang, N Agrawal, BR Bartlett, H Wang, B Luber, RM Alani, ES Antonarakis, NS Azad, A Bardelli, H Brem, JL Cameron, CC Lee, LA Fecher, GL Gallia, P Gibbs, D Le, RL Giuntoli, M Goggins, MD Hogarty, M Holdhoff, SM Hong, Y Jiao, HH Juhl, JJ Kim, G Siravegna, DA Laheru, C Lauricella, M Lim, EJ Lipson, SK Marie, GJ Netto, KS Oliner, A Olivi, L Olsson, GJ Riggins, A Sartore-Bianchi, K Schmidt, lM Shih, SM Oba-Shinjo, S Siena, D Theodorescu, J Tie, TT Harkins, S Veronese, TL Wang, JD Weingart, CL Wolfgang, LD Wood, D Xing, RH Hruban, J Wu, PJ Allen, CM Schmidt, MA Choti, VE Velculescu, KW Kinzler, B Vogelstein, N Papadopoulos, and LA Jr Diaz. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med*, 6(224):224ra24., 2014.

[110] J. Moss, J. Magenheim, D. Neiman, H. Zemmour, N. Loyfer, A. Korach, Y. Samet, M. Maoz, H. Druid, P. Arner, K. Y. Fu, E. Kiss, K. L. Spalding, G. Landesberg, A. Zick, A. Grinshpun, A. M. J. Shapiro, M. Grompe, A. D. Wittenberg, B. Glaser, R. Shemer, T. Kaplan, and Y. Dor. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat Commun*, 9 (1):5068, 11 2018.

[111] F. C. Wong, K. Sun, P. Jiang, Y. K. Cheng, K. C. Chan, T. Y. Leung, R. W. Chiu, and Y. M. Lo. Cell-free DNA in maternal plasma and serum: A comparison of quantity, quality and tissue origin using genomic and epigenomic approaches. *Clin Biochem*, 49(18):1379–1386, Dec 2016.

[112] J Zhou, L Chang, Y Guan, L Yang, X Xia, L Cui, X Yi, and G Lin. Application of Circulating Tumor DNA as a Non-Invasive Tool for Monitoring the Progression of Colorectal Cancer. *PLoS One*, 11(7):e0159708., 2016.

[113] M Schirmer, UZ Ijaz, R D'Amore, N Hall, WT Sloan, and C Quince. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res*, 43(6):e37., 2015.

[114] EI Pentsova, RH Shah, J Tang, A Boire, D You, S Briggs, A Omuro, X Lin, M Fleisher, C Grommes, KS Panageas, F Meng, SD Selcuklu, S Ogilvie, N Distefano, L Shagabayeva, M Rosenblum, LM DeAngelis, A Viale, IK Mellinghoff, and MF Berger. Evaluating Cancer of the Central Nervous System Through Next-Generation Sequencing of Cerebrospinal Fluid. *J Clin Oncol*, 34(20):2404-15., 2016.

[115] CE Teunissen, A Petzold, JL Bennett, FS Berven, L Brundin, M Comabella, D Franciotta, JL Frederiksen, JO Fleming, R Furlan, RQ Hintzen, SG Hughes, MH Johnson, E Krasulova, J Kuhle, MC Magnone, C Rajda, K Rejdak, HK Schmidt, Pesch V van, E Waubant, C Wolf, G Giovannoni, B Hemmer, H Tumani, and F Deisenhammer. A consensus protocol for the standardization of cerebrospinal fluid collection and biobanking. *Neurology*, 73(22):1914-22., 2009.

[116] W Shi, C Lv, J Qi, W Zhao, X Wu, R Jing, X Wu, S Ju, and J Chen. Prognostic value of free DNA quantification in serum and cerebrospinal fluid in glioma patients. *J Mol Neurosci*, 46(3):470-5., 2012.

[117] M. K. Tuck, D. W. Chan, D. Chia, A. K. Godwin, W. E. Grizzle, K. E. Krueger, W. Rom, M. Sanda, L. Sorbara, S. Stass, W. Wang, and D. E. Brenner. Standard operating procedures for serum and plasma collection: early detection research network consensus statement standard operating procedure integration working group. *J Proteome Res*, 8(1):113–117, Jan 2009.

[118] Y. W. Kim, Y. H. Kim, Y. Song, H. S. Kim, H. W. Sim, S. Poojan, B. W. Eom, M. C. Kook, J. Joo, and K. M. Hong. Monitoring circulating tumor DNA by analyzing personalized cancer-specific rearrangements to detect recurrence in gastric cancer. *Exp Mol Med*, 51(8):1–10, 08 2019.

[119] J Ahn, B Hwang, Kim H Young, H Jang, HP Kim, SW Han, TY Kim, Lee J Hyun, and D Bang. Asymmetrical barcode adapter-assisted recovery of duplicate reads and error correction strategy to detect rare mutations in circulating tumor DNA. *Sci Rep*, 7:46678., 2017.

[120] Q. Ma, F. Schlegel, S. B. Bachmann, H. Schneider, Y. Decker, M. Rudin, M. Weller, S. T. Proulx, and M. Detmar. Lymphatic outflow of cerebrospinal fluid is reduced in glioma. *Sci Rep*, 9(1):14815, 10 2019.

[121] S. W. Bothwell, D. Janigro, and A. Patabendige. Cerebrospinal fluid dynamics and intracranial pressure elevation in neurological diseases. *Fluids Barriers CNS*, 16(1): 9, Apr 2019.

[122] M. Poca and J. Sahuquillo. [Intracranial pressure monitoring and CSF dynamics in patients with neurological disorders: indications and practical considerations]. *Neurologia*, 16(7):303–320, 2001.

[123] AM Miller, RH Shah, EI Pentsova, M Pourmaleki, S Briggs, N Distefano, Y Zheng, A Skakodub, SA Mehta, C Campos, WY Hsieh, SD Selcuklu, L Ling, F Meng, X Jing, A Samoila, TA Bale, DWY Tsui, C Grommes, A Viale, MM Souweidane, V Tabar, CW Brennan, AS Reiner, M Rosenblum, KS Panageas, LM DeAngelis, RJ Young, MF Berger, and IK Mellinghoff. Tracking tumour evolution in glioma through liquid biopsies of cerebrospinal fluid. *Nature*, 565(7741):654-658., 2019.

[124] C. Pan, B. H. Diplas, X. Chen, Y. Wu, X. Xiao, L. Jiang, Y. Geng, C. Xu, Y. Sun, P. Zhang, W. Wu, Y. Wang, Z. Wu, J. Zhang, Y. Jiao, H. Yan, and L. Zhang. Molecular profiling of tumors of the brainstem by sequencing of CSF-derived circulating tumor DNA. *Acta Neuropathol*, 137(2):297–306, 02 2019.

[125] C. Grommes, S. S. Tang, J. Wolfe, T. J. Kaley, M. Daras, E. I. Pentsova, A. F. Piotrowski, J. Stone, A. Lin, C. P. Nolan, M. Manne, P. Codega, C. Campos, A. Viale, A. A. Thomas, M. F. Berger, V. Hatzoglou, A. S. Reiner, K. S. Panageas, L. M. DeAngelis, and I. K. Mellinghoff. Phase 1b trial of an ibrutinib-based combination therapy in recurrent/refractory CNS lymphoma. *Blood*, 133(5):436–445, 01 2019.

[126] M. Simonelli, A. Dipasquale, F. Orzan, E. Lorenzi, P. Persico, P. Navarria, F. Pessina, M. C. Nibali, L. Bello, A. Santoro, and C. Boccaccio. Cerebrospinal fluid tumor DNA for liquid biopsy in glioma patients' management: Close to the clinic? *Crit Rev Oncol Hematol*, 146:102879, Feb 2020.

[127] I Kinde, J Wu, N Papadopoulos, KW Kinzler, and B Vogelstein. Detection and

quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A*, 108(23):9530-5., 2011.

[128] J Eboreime, SK Choi, SR Yoon, N Arnheim, and P Calabrese. Estimating Exceptionally Rare Germline and Somatic Mutation Frequencies via Next Generation Sequencing. *PLoS One*, 11(6):e0158340., 2016.

[129] Rubicon Genomics. *ThruPLEX Tag-seq Kit*, 2016. ADX-335-001.

[130] C Gates. *Connor(URL: https://github.com/umich-brcf-bioinf/Connor)*. [(09, 2019) accessed].

[131] SR Kennedy. Duplex Sequencing, 2014. `https://github.com/Kennedy-Lab-UW/Duplex-Sequencing` [accessed: 2021-05-21].

[132] DI Lou, JA Hussmann, RM McBee, A Acevedo, R Andino, WH Press, and SL Sawyer. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc Natl Acad Sci U S A*, 110(49):19872-7., 2013.

[133] JB Hiatt, CC Pritchard, SJ Salipante, BJ O'Roak, and J Shendure. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res*, 23(5):843-54., 2013.

[134] MW Schmitt, EJ Fox, MJ Prindle, KS Reid-Bayliss, LD True, JP Radich, and LA Loeb. Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat Methods*, 12(5):423-5., 2015.

[135] MW Schmitt, SR Kennedy, JJ Salk, EJ Fox, JB Hiatt, and LA Loeb. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A*, 109(36):14508-13., 2012.

[136] D. Fernandez-Garcia, A. Hills, K. Page, R. K. Hastings, B. Toghill, K. S. Goddard, C. Ion, O. Ogle, A. R. Boydell, K. Gleason, M. Rutherford, A. Lim, D. S. Guttery, R. C. Coombes, and J. A. Shaw. Plasma cell-free DNA (cfDNA) as a predictive and prognostic marker in patients with metastatic breast cancer. *Breast Cancer Res*, 21 (1):149, 12 2019.

[137] J Chung, DS Son, HJ Jeon, KM Kim, G Park, GH Ryu, WY Park, and D Park. The minimal amount of starting DNA for Agilent's hybrid capture-based targeted massively parallel sequencing. *Sci Rep*, 6:26732., 2016.

[138] Cancer Research UK. Children's cancers incidence statistics, 2016. `http://www.cancerresearchuk.org/health-professional/cancer-statistics/childrens-cancers/incidence` [accessed: 2016-12-21].

[139] F Bray, J Ferlay, I Soerjomataram, RL Siegel, LA Torre, and A Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, 68(6):394-424., 2018.

[140] E Ward, C DeSantis, A Robbins, B Kohler, and A Jemal. Childhood and adolescent cancer statistics, 2014. *CA Cancer J Clin*, 64(2):83-103., 2014.

[141] Cancer Research UK. Children's cancers mortality statistics, 2016. `http://www.cancerresearchuk.org/health-professional/cancer-statistics/childrens-cancers/mortality` [accessed: 2016-12-21].

[142] KJ Johnson, J Cullen, JS Barnholtz-Sloan, QT Ostrom, CE Langer, MC Turner, R McKean-Cowdin, JL Fisher, PJ Lupo, S Partap, JA Schwartzbaum, and ME Scheurer. Childhood brain tumor epidemiology: a brain tumor epidemiology consortium review. *Cancer Epidemiol Biomarkers Prev*, 23(12):2716-36., 2014.

[143] K. Juraschka and M. D. Taylor. Medulloblastoma in the age of molecular subgroups: a review. *J Neurosurg Pediatr*, 24(4):353–363, 10 2019.

[144] L. Bauchet, V. Rigau, H. Mathieu-Daudé, P. Fabbro-Peray, G. Palenzuela, D. Figarella-Branger, J. Moritz, S. Puget, F. Bauchet, L. Pallusseau, H. Duffau, P. Coubes, B. Trétarre, F. Labrousse, P. Dhellemmes, N. Aghakhani, M. Ali Benali, B. Alliez, D. Amat, A. Amlashi, F. Arbez-Gindre, F. Arbion, R. Assaker, M. H. Aubriot Lorton, J. Auque, A. Autricque, I. Auvigne, G. Averous, P. Baldet, B. Bataille, A. Bazin, J. Beaurain, J. Benezech, A. Bergemer Fouquet, G. Besson, F. Beuvon, C. Billotet, S. Blond, S. Boetto, H. Boissonnet, G. Bonyhay, P. Bouillot, P. Bourgeois, C. Bouvier, G. Brassier, C. Broche, J. Brunon, P. Cabal, V. Cahn, F. Caire, P. Calvet, D. Cazals-Hatem, F. Chapon, J. Chazal, T. Civit, S. Colnat, M. Colombat, J. Comoy, A. Couvelard, A. Czorny, P. Dam Hieu, C. Daumas-Duport,

M. Dautheribes, P. David, B. Debono, M. Delage Corre, M. Delhaye, M. B. Delisle, G. Delsol, J. M. Derlon, C. Desenclos, A. Desplat, B. Devaux, F. Di Rocco, A. Diaz, M. D. Diebold, G. Dorfmuller, G. Dran, T. Dufour, B. Dumas, J. M. Dumollard, L. Durand, R. Duthel, S. Eimer, H. El Fertit, E. Emery, C. Espagno, P. Esposito, M. P. Etchandy, R. P. Eyremandi, T. Faillot, S. Felix, C. Fernandez, J. Fesselet, D. Fontaine, D. Fournier, P. François, S. Froelich, J. M. Fuentes, S. Fuentes, R. Gadan, C. Gaspard, G. Gay, M. Gigaud, S. Gil Robles, J. Godard, M. F. Gontier, J. M. Goujon, F. Gray, Y. Grignon, F. Grisoli, J. Guarnieri, J. Guyotat, P. Hallacq, A. Hamlat, G. Hayek, A. Heitzmann, V. Hennequin, J. C. Huot, B. Irthum, G. Jacquet, M. Jan, F. Jaubert, E. Jouanneau, A. Jouvet, E. Justrabo, M. Kalamarides, P. Kehrli, J. L. Kemeny, Y. Keravel, R. Kerdraon, T. Khalil, K. Khouri, S. Khouri, O. Klein, M. Kujas, C. Lacroix, J. Lagarrigue, O. Langlois, F. Lapierre, A. Laquerriere, M. C. Laurent, F. Le Gall, C. Le Guerinel, M. Le Houcq, E. Lechapt, D. Legars, J. J. Lemaire, G. Lena, J. F. Lepeintre, B. Leriche, J. P. Lescure, P. Levillain, D. Liguoro, E. Lioret, A. Listrat, H. Loiseau, M. Lonjon, M. Lopes, G. Lot, E. Louis, J. Maheut-Lourmière, A. Maillard, F. Maitre, D. Maitrot, E. Majek-Zakine, E. Mandonnet, N. Manzo, J. C. Marchal, B. Marie, C. A. Maurage, P. Menei, P. Mercier, E. Mergey, P. Metellus, S. Michalak, J. F. Michiels, S. Milinkevitch, J. F. Mineo, C. Miquel, E. Mireau, M. Mohr, K. Mokhtari, X. Morandi, S. Morar, J. J. Moreau, S. Moreno, K. L. Mourier, C. Mottolese, F. Nataf, A. Neuville, L. Nogues, R. Noudel, C. Nuti, P. Page, P. Paquis, M. Parent, F. Parker, F. Pasqualini, M. Patey, I. Pelissou-Guyotat, M. Peoc'h, J. C. Peragut, P. Peruzzi, A. Pierre-Kahn, C. Pinelli, M. Polivka, I. Pommepuy, T. Ponnelle, V. Porhiel, F. Proust, I. Quintin-Roue, O. Ragragui, D. Rasendrarijao, P. Raynaud, A. Redondo, L. Renjard, N. Reyns, S. Richard, J. Richaud, T. Riem, L. Riffaud, F. Ringenbach, G. Robert, P. H. Roche, M. A. Rodriguez, T. Roujeau, P. Rousseaux, M. C. Rousselet, F. E. Roux, F. X. Roux, M. M. Ruchoux, J. Sabatier, P. Sabatier, S. Saïkali, J. P. Saint Andre, G. Saint Pierre, C. Saint-Rose, F. San Galli, J. L. Sautreaux, B. Sawan, D. Scavarda, F. Segnarbieux, E. Seigneuret, M. Sindou, R. Sorbara, A. Sorin, B. Stilhart, P. Straub, S. Taha, J. P. Ternier, M. C. Tortel, P. Toussaint, G. Touzet, M. Tremoulet, J. Trouillas, A. Tubiana, E. Uro-Coste, F. Vandenbos, P. Varlet, S. Velut, J. Vidal, G. Viennet, J. M. Vignaud, J. R. Vignes, M. Vinchon, A. Vital, M. Wager, N. Weinbreck, and M. Zerah. Clinical

epidemiology for childhood primary central nervous system tumors. *J Neurooncol*, 92(1):87–98, Mar 2009.

[145] DM Ho, CY Hsu, TT Wong, LT Ting, and H Chiang. Atypical teratoid/rhabdoid tumor of the central nervous system: a comparative study with primitive neuroectodermal tumor/medulloblastoma. *Acta Neuropathol*, 99(5):482-8., 2000.

[146] JA Biegel, JY Zhou, LB Rorke, C Stenstrom, LM Wainwright, and B Fogelgren. Germ-line and acquired mutations of INI1 in atypical teratoid and rhabdoid tumors. *Cancer Res*, 59(1):74-9., 1999.

[147] CL Nesvick, AA Nageswara Rao, A Raghunathan, JA Biegel, and DJ Daniels. Case-based review: atypical teratoid/rhabdoid tumor. *Neurooncol Pract*, 6(3):163-178., 2018.

[148] L Lafay-Cousin, C Hawkins, AS Carret, D Johnston, S Zelcer, B Wilson, N Jabado, K Scheinemann, D Eisenstat, C Fryer, A Fleming, C Mpofu, V Larouche, D Strother, E Bouffet, and A Huang. Central nervous system atypical teratoid rhabdoid tumours: the Canadian Paediatric Brain Tumour Consortium experience. *Eur J Cancer*, 48(3):353-9., 2012.

[149] CS Lau, K Mahendraraj, and RS Chamberlain. Atypical teratoid rhabdoid tumors: a population-based clinical outcomes study involving 174 patients from the Surveillance, Epidemiology, and End Results database (1973-2010). *Cancer Manag Res*, 7:301-9., 2015.

[150] D Schrey, Lechón F Carceller, G Malietzis, L Moreno, C Dufour, S Chi, L Lafay-Cousin, Hoff K von, T Athanasiou, LV Marshall, and S Zacharoulis. Multimodal therapy in children and adolescents with newly diagnosed atypical teratoid rhabdoid tumor: individual pooled data analysis and review of the literature. *J Neurooncol*, 126(1):81-90., 2016.

[151] M Babgi, A Samkari, A Al-Mehdar, and S Abdullah. Atypical Teratoid/Rhabdoid Tumor of the Spinal Cord in a Child: Case Report and Comprehensive Review of the Literature. *Pediatr Neurosurg*, 53(4):254-262., 2018.

[152] KH Chan, Haspani MS Mohammed, YC Tan, and F Kassim. A case report of atypical teratoid/rhabdoid tumour in a 9-year-old girl. *Malays J Med Sci*, 18(3):82-6., 2011.

[153] JM Hilden, S Meerbaum, P Burger, J Finlay, A Janss, BW Scheithauer, AW Walter, LB Rorke, and JA Biegel. Central nervous system atypical teratoid/rhabdoid tumor: results of therapy in children enrolled in a registry. *J Clin Oncol*, 22(14):2877-84., 2004.

[154] IH Lee, SY Yoo, JH Kim, H Eo, OH Kim, IO Kim, JE Cheon, AY Jung, and BJ Yoon. Atypical teratoid/rhabdoid tumors of the central nervous system: imaging and clinical findings in 16 children. *Clin Radiol*, 64(3):256-64., 2009.

[155] D Segal and MA Karajannis. Pediatric Brain Tumors: An Update. *Curr Probl Pediatr Adolesc Health Care*, 46(7):242-250., 2016.

[156] RS Lee, C Stewart, SL Carter, L Ambrogio, K Cibulskis, C Sougnez, MS Lawrence, D Auclair, J Mora, TR Golub, JA Biegel, G Getz, and CW Roberts. A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. *J Clin Invest*, 122(8):2983-8., 2012.

[157] JA Biegel, L Tan, F Zhang, L Wainwright, P Russo, and LB Rorke. Alterations of the hSNF5/INI1 gene in central nervous system atypical teratoid/rhabdoid tumors and renal and extrarenal rhabdoid tumors. *Clin Cancer Res*, 8(11):3461-7., 2002.

[158] C Thomas, F Knerlich-Lukoschus, H Reinhard, PD Johann, D Sturm, F Sahm, S Bens, J Vogt, K Nemes, F Oyen, U Kordes, R Siebert, R Schneppenheim, M Messing-Jünger, T Pietsch, Deimling A von, W Paulus, SM Pfister, M Kool, MC Frühwald, and M Hasselblatt. Two molecularly distinct atypical teratoid/rhabdoid tumors (or tumor components) occurring in an infant with rhabdoid tumor predisposition syndrome 1. *Acta Neuropathol*, 137(5):847-850., 2019.

[159] V. A. Fitzhugh. Rhabdoid Tumor Predisposition Syndrome and Pleuropulmonary Blastoma Syndrome. *J Pediatr Genet*, 5(2):124–128, Jun 2016.

[160] K Kohashi, Y Tanaka, H Kishimoto, H Yamamoto, Y Yamada, T Taguchi, Y Iwamoto, and Y Oda. Reclassification of rhabdoid tumor and pediatric undifferentiated/unclassified sarcoma with complete loss of SMARCB1/INI1 protein expression: three subtypes of rhabdoid tumor according to their histological features. *Mod Pathol*, 29(10):1232-42., 2016.

[161] K Kohashi and Y Oda. Oncogenic roles of SMARCB1/INI1 and its deficient tumors. *Cancer Sci*, 108(4):547-552., 2017.

[162] AC Hoot, P Russo, AR Judkins, EJ Perlman, and JA Biegel. Immunohistochemical analysis of hSNF5/INI1 distinguishes renal and extra-renal malignant rhabdoid tumors from other pediatric soft tissue tumors. *Am J Surg Pathol*, 28(11):1485-91., 2004.

[163] K Kohashi, Y Oda, H Yamamoto, S Tamiya, T Izumi, S Ohta, T Taguchi, S Suita, and M Tsuneyoshi. Highly aggressive behavior of malignant rhabdoid tumor: a special reference to SMARCB1/INI1 gene alterations using molecular genetic analysis including quantitative real-time PCR. *J Cancer Res Clin Oncol*, 133(11):817-24., 2007.

[164] U Kordes, S Gesk, MC Frühwald, N Graf, I Leuschner, M Hasselblatt, A Jeibmann, F Oyen, O Peters, T Pietsch, R Siebert, and R Schneppenheim. Clinical and molecular features in patients with atypical teratoid rhabdoid tumor or malignant rhabdoid tumor. *Genes Chromosomes Cancer*, 49(2):176-81., 2010.

[165] LB Rorke, RJ Packer, and JA Biegel. Central nervous system atypical teratoid/rhabdoid tumors of infancy and childhood: definition of an entity. *J Neurosurg*, 85(1):56-65., 1996.

[166] F. D'Arco, S. Culleton, L. J. L. De Cocker, K. Mankad, J. Davila, and B. Tamrazi. Current concepts in radiologic assessment of pediatric brain tumors during treatment, part 1. *Pediatr Radiol*, 48(13):1833–1843, 12 2018.

[167] AR Judkins, J Mauger, A Ht, LB Rorke, and JA Biegel. Immunohistochemical analysis of hSNF5/INI1 in pediatric CNS neoplasms. *Am J Surg Pathol*, 28(5):644-50., 2004.

[168] B. R. Pawel. SMARCB1-deficient Tumors of Childhood: A Practical Guide. *Pediatr Dev Pathol*, 21(1):6–28, 2018.

[169] D. Holdhof, P. D. Johann, M. Spohn, M. Bockmayr, S. Safaei, P. Joshi, J. Masliah-Planchon, B. Ho, M. Andrianteranagna, F. Bourdeaut, A. Huang, M. Kool, S. A. Upadhyaya, A. E. Bendel, D. Indenbirken, W. D. Foulkes, J. W. Bush, D. Creytens,

U. Kordes, M. C. Frühwald, M. Hasselblatt, and U. Schüller. Atypical teratoid/rhabdoid tumors (ATRTs) with SMARCA4 mutation are molecularly distinct from SMARCB1-deficient cases. *Acta Neuropathol*, 141(2):291–301, 02 2021.

[170] HL Müller. Childhood craniopharyngioma–current concepts in diagnosis, therapy and follow-up. *Nat Rev Endocrinol*, 6(11):609-18., 2010.

[171] SW Clark, TJ Kenning, and JJ Evans. Recurrent ectopic craniopharyngioma in the sylvian fissure thirty years after resection through a pterional approach: a case report and review of the literature. *Nagoya J Med Sci*, 77(1-2):297-306., 2015.

[172] P Mortini, M Losa, G Pozzobon, R Barzaghi, M Riva, S Acerno, D Angius, G Weber, G Chiumello, and M Giovanelli. Neurosurgical treatment of craniopharyngioma in adults and children: early and long-term results in a large case series. *J Neurosurg*, 114(5):1350-9., 2011.

[173] D Horiuchi, T Shimono, S Doishita, T Goto, S Tanaka, and Y Miki. Ectopic clival craniopharyngioma with intratumoral hemorrhage: A case report. *Radiol Case Rep*, 14(8):977-980., 2019.

[174] H. L. Müller. Craniopharyngioma. *Endocr Rev*, 35(3):513–543, Jun 2014.

[175] L. O'steen and D. J. Indelicato. Advances in the management of craniopharyngioma. *F1000Res*, 7, 2018.

[176] AM Donson, J Apps, AM Griesinger, V Amani, DA Witt, RCE Anderson, TN Niazi, G Grant, M Souweidane, JM Johnston, EM Jackson, BK Kleinschmidt-DeMasters, MH Handler, AC Tan, L Gore, A Virasami, JM Gonzalez-Meljem, TS Jacques, JP Martinez-Barbera, NK Foreman, TC Hankinson, and Treatment for Pediatric Craniopharyngioma Consortium Advancing. Molecular Analyses Reveal Inflammatory Mediators in the Solid Component and Cyst Fluid of Human Adamantinomatous Craniopharyngioma. *J Neuropathol Exp Neurol*, 76(9):779-788., 2017.

[177] C Gao, Y Wang, R Broaddus, L Sun, F Xue, and W Zhang. Exon 3 mutations of CTNNB1 drive tumorigenesis: a review. *Oncotarget*, 9(4):5492-5508., 2017.

[178] E Oikonomou, DC Barreto, B Soares, Marco L De, M Buchfelder, and EF Adams.

Beta-catenin mutations in craniopharyngiomas and pituitary adenomas. *J Neurooncol*, 73(3):205-9., 2005.

[179] S Sartoretti-Schefer, W Wichmann, A Aguzzi, and A Valavanis. MR differentiation of adamantinous and squamous-papillary craniopharyngiomas. *AJNR Am J Neuroradiol*, 18(1):77-87., 1997.

[180] ML Garrè and A Cama. Craniopharyngioma: modern concepts in pathogenesis and treatment. *Curr Opin Pediatr*, 19(4):471-9., 2007.

[181] P Mortini, F Gagliardi, N Boari, and M Losa. Surgical strategies and modern therapeutic options in the treatment of craniopharyngiomas. *Crit Rev Oncol Hematol*, 88(3):514-29., 2013.

[182] S. Cavalheiro, P. A. Dastoli, N. S. Silva, S. Toledo, H. Lederman, and M. C. da Silva. Use of interferon alpha in intratumoral chemotherapy for cystic craniopharyngioma. *Childs Nerv Syst*, 21(8-9):719–724, Aug 2005.

[183] S Cavalheiro, C Di Rocco, S Valenzuela, PA Dastoli, G Tamburrini, L Massimi, JM Nicacio, IV Faquini, DF Ierardi, NS Silva, BL Pettorini, and SR Toledo. Craniopharyngiomas: intratumoral chemotherapy with interferon-alpha: a multicenter preliminary study with 60 cases. *Neurosurg Focus*, 28(4):E12., 2010.

[184] H Ma, W Yang, L Zhang, S Liu, M Zhao, G Zhou, L Wang, S Jin, Z Zhang, and J Hu. Interferon-alpha promotes immunosuppression through IFNAR1/STAT1 signalling in head and neck squamous cell carcinoma. *Br J Cancer*, 120(3):317-330., 2019.

[185] AD Rapidis and GT Wolf. Immunotherapy of head and neck cancer: current and future considerations. *J Oncol*, 2009:346345., 2009.

[186] J R Apps, J C Hutchinson, S Shelmerdine, A Virasami, E Winter, T S Jacques, J Martinez-Barbera, O Arthurs, and T Czech. CRAN-32. Case Based Learning: Three centuries of lessons from two craniopharyngeoma patients. *Neuro-Oncology*, 20:i43-i43., 2018.

[187] BL Pettorini, R Inzitari, L Massimi, G Tamburrini, M Caldarelli, C Fanali, T Cabras, I Messana, M Castagnola, and Rocco C Di. The role of inflammation in the genesis

of the cystic component of craniopharyngiomas. *Childs Nerv Syst*, 26(12):1779-84., 2010.

[188] JT Yeung, IF Pollack, A Panigrahy, and RI Jakacki. Pegylated interferon-α-2b for children with recurrent craniopharyngioma. *J Neurosurg Pediatr*, 10(6):498-503., 2012.

[189] R. I. Jakacki, B. H. Cohen, C. Jamison, V. P. Mathews, E. Arenson, D. C. Longee, J. Hilden, A. Cornelius, M. Needle, D. Heilman, J. C. Boaz, and T. G. Luerssen. Phase II evaluation of interferon-alpha-2a for progressive or recurrent craniopharyngiomas. *J Neurosurg*, 92(2):255–260, Feb 2000.

[190] CR Freeman and JP Farmer. Pediatric brain stem gliomas: a review. *Int J Radiat Oncol Biol Phys*, 40(2):265-71., 1998.

[191] A Ramos, A Hilario, A Lagares, E Salvador, A Perez-Nuñez, and J Sepulveda. Brainstem gliomas. *Semin Ultrasound CT MR*, 34(2):104-12., 2013.

[192] P Buczkowicz, U Bartels, E Bouffet, O Becher, and C Hawkins. Histopathological spectrum of paediatric diffuse intrinsic pontine glioma: diagnostic and therapeutic implications. *Acta Neuropathol*, 128(4):573-81., 2014.

[193] I. F. Pollack, S. Agnihotri, and A. Broniscer. Childhood brain tumors: current management, biological insights, and future directions. *J Neurosurg Pediatr*, 23(3):261–273, 03 2019.

[194] D Hargrave, U Bartels, and E Bouffet. Diffuse brainstem glioma in children: critical review of clinical trials. *Lancet Oncol*, 7(3):241-8, 2006.

[195] KE Warren. Diffuse intrinsic pontine glioma: poised for progress. *Front Oncol*, 2:205., 2012.

[196] KR Taylor, A Mackay, N Truffaux, Y Butterfield, O Morozova, C Philippe, D Castel, CS Grasso, M Vinci, D Carvalho, AM Carcaboso, Torres C de, O Cruz, J Mora, N Entz-Werle, WJ Ingram, M Monje, D Hargrave, AN Bullock, S Puget, S Yip, C Jones, and J Grill. Recurrent activating ACVR1 mutations in diffuse intrinsic pontine glioma. *Nat Genet*, 46(5):457-461., 2014.

[197] P Buczkowicz and C Hawkins. Pathology, Molecular Genetics, and Epigenetics of Diffuse Intrinsic Pontine Glioma. *Front Oncol*, 5:147., 2015.

[198] G Wu, AK Diaz, BS Paugh, SL Rankin, B Ju, Y Li, X Zhu, C Qu, X Chen, J Zhang, J Easton, M Edmonson, X Ma, C Lu, P Nagahawatte, E Hedlund, M Rusch, S Pounds, T Lin, A Onar-Thomas, R Huether, R Kriwacki, M Parker, P Gupta, J Becksfort, L Wei, HL Mulder, K Boggs, B Vadodaria, D Yergeau, JC Russell, K Ochoa, RS Fulton, LL Fulton, C Jones, FA Boop, A Broniscer, C Wetmore, A Gajjar, L Ding, ER Mardis, RK Wilson, MR Taylor, JR Downing, DW Ellison, J Zhang, and SJ Baker. The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. *Nat Genet*, 46(5):444-450., 2014.

[199] S Haase, MB Garcia-Fabiani, S Carney, D Altshuler, FJ Núñez, FM Méndez, F Núñez, PR Lowenstein, and MG Castro. Mutant ATRX: uncovering a new therapeutic target for glioma. *Expert Opin Ther Targets*, 22(7):599-613., 2018.

[200] D Carvalho, KR Taylor, NG Olaciregui, V Molinari, M Clarke, A Mackay, R Ruddle, A Henley, M Valenti, A Hayes, AH Brandon, SA Eccles, F Raynaud, A Boudhar, M Monje, S Popov, AS Moore, J Mora, O Cruz, M Vinci, PE Brennan, AN Bullock, AM Carcaboso, and C Jones. ALK2 inhibitors display beneficial effects in preclinical models of ACVR1 mutant diffuse intrinsic pontine glioma. *Commun Biol*, 2:156., 2019.

[201] LM Hoffman, M DeWire, S Ryall, P Buczkowicz, J Leach, L Miles, A Ramani, M Brudno, SS Kumar, R Drissi, P Dexheimer, R Salloum, L Chow, T Hummel, C Stevenson, QR Lu, B Jones, D Witte, B Aronow, CE Hawkins, and M Fouladi. Spatial genomic heterogeneity in diffuse intrinsic pontine and midline high-grade glioma: implications for diagnostic biopsy and targeted therapeutics. *Acta Neuropathol Commun*, 4:1., 2016.

[202] NJ Robison and MW Kieran. Diffuse intrinsic pontine glioma: a reassessment. *J Neurooncol*, 119(1):7-15., 2014.

[203] N Gupta, LC Goumnerova, P Manley, SN Chi, D Neuberg, M Puligandla, J Fangusaro, S Goldman, T Tomita, T Alden, A DiPatri, JB Rubin, K Gauvain, D Limbrick, J Leonard, JR Geyer, S Leary, S Browd, Z Wang, S Sood, A Bendel, M Nagib,

S Gardner, MA Karajannis, D Harter, K Ayyanar, W Gump, DC Bowers, B Weprin, TJ MacDonald, D Aguilera, B Brahma, NJ Robison, E Kiehna, M Krieger, E Sandler, P Aldana, Z Khatib, J Ragheb, S Bhatia, S Mueller, A Banerjee, AL Bredlau, S Gururangan, H Fuchs, KJ Cohen, G Jallo, K Dorris, M Handler, M Comito, M Dias, K Nazemi, L Baird, J Murray, N Lindeman, JL Hornick, H Malkin, C Sinai, L Greenspan, KD Wright, M Prados, P Bandopadhayay, KL Ligon, and MW Kieran. Prospective feasibility and safety assessment of surgical biopsy for patients with newly diagnosed diffuse intrinsic pontine glioma. *Neuro Oncol*, 20(11):1547-1555., 2018.

[204] E Pfaff, Damaty A El, GP Balasubramanian, M Blattner-Johnson, BC Worst, S Stark, H Witt, KW Pajtler, Tilburg CM van, R Witt, T Milde, M Jakobs, P Fiesel, MC Frühwald, Driever P Hernáiz, UW Thomale, MU Schuhmann, M Metzler, K Bochennek, T Simon, M Dürken, M Karremann, S Knirsch, M Ebinger, Bueren AO von, T Pietsch, C Herold-Mende, DE Reuss, K Kiening, P Lichter, A Eggert, CM Kramm, SM Pfister, DTW Jones, H Bächli, and O Witt. Brainstem biopsy in pediatric diffuse intrinsic pontine glioma in the era of precision medicine: the INFORM study experience. *Eur J Cancer*, 114:27-35., 2019.

[205] F Martínez-Ricarte, R Mayor, E Martínez-Sáez, C Rubio-Pérez, E Pineda, E Cordero, M Cicuéndez, MA Poca, N López-Bigas, Y Cajal S Ramon, M Vieito, J Carles, J Tabernero, A Vivancos, S Gallego, F Graus, J Sahuquillo, and J Seoane. Molecular Diagnosis of Diffuse Gliomas through Sequencing of Cell-Free Circulating Tumor DNA from Cerebrospinal Fluid. *Clin Cancer Res*, 24(12):2812-2819., 2018.

[206] A. M. Saratsis, S. Yadavilli, S. Magge, B. R. Rood, J. Perez, D. A. Hill, E. Hwang, L. Kilburn, R. J. Packer, and J. Nazarian. Insights into pediatric diffuse intrinsic pontine glioma through proteomic analysis of cerebrospinal fluid. *Neuro Oncol*, 14 (5):547–560, May 2012.

[207] V. M. Lu, E. A. Power, L. Zhang, and D. J. Daniels. Liquid biopsy for diffuse intrinsic pontine glioma: an update. *J Neurosurg Pediatr*, pages 1–8, Sep 2019.

[208] M. I. Vanan and D. D. Eisenstat. DIPG in Children - What Can We Learn from the Past? *Front Oncol*, 5:237, 2015.

[209] E Panditharatna, LB Kilburn, MS Aboian, M Kambhampati, H Gordish-Dressman, SN Magge, N Gupta, JS Myseros, EI Hwang, C Kline, JR Crawford, KE Warren, S Cha, WS Liang, ME Berens, RJ Packer, AC Resnick, M Prados, S Mueller, and J Nazarian. Clinically Relevant and Minimally Invasive Tumor Surveillance of Pediatric Diffuse Midline Gliomas Using Patient-Derived Liquid Biopsy. *Clin Cancer Res*, 24(23):5850-5859., 2018.

[210] S Stallard, MG Savelieff, K Wierzbicki, B Mullan, Z Miklja, A Bruzek, T Garcia, R Siada, B Anderson, BH Singer, R Hashizume, AM Carcaboso, KQ McMurray, J Heth, K Muraszko, PL Robertson, R Mody, S Venneti, H Garton, and C Koschmann. CSF H3F3A K27M circulating tumor DNA copy number quantifies tumor growth and in vitro treatment response. *Acta Neuropathol Commun*, 6(1):80., 2018.

[211] TY Huang, A Piunti, RR Lulla, J Qi, CM Horbinski, T Tomita, CD James, A Shilatifard, and AM Saratsis. Detection of Histone H3 mutations in cerebrospinal fluid-derived tumor DNA from children with diffuse midline glioma. *Acta Neuropathol Commun*, 5(1):28., 2017.

[212] M Murtaza, SJ Dawson, DW Tsui, D Gale, T Forshew, AM Piskorz, C Parkinson, SF Chin, Z Kingsbury, AS Wong, F Marass, S Humphray, J Hadfield, D Bentley, TM Chin, JD Brenton, C Caldas, and N Rosenfeld. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature*, 497(7447):108-12., 2013.

[213] MA Quail, H Swerdlow, and DJ Turner. Improved Protocols for Illumina Sequencing. *Curr Protoc Hum Genet*, Chapter 18:Unit 18.2., 2020.

[214] N Pallisgaard, KL Spindler, RF Andersen, I Brandslund, and A Jakobsen. Controls to validate plasma samples for cell free DNA quantification. *Clin Chim Acta*, 446:141-6., 2015.

[215] QIAGEN. *QIAamp Circulating Nucleic Acid Handbook*, 2013. 3rd Edition.

[216] C Guichard, G Amaddeo, S Imbeaud, Y Ladeiro, L Pelletier, IB Maad, J Calderaro, P Bioulac-Sage, M Letexier, F Degos, B Clément, C Balabaud, E Chevet, A Laurent, G Couchy, E Letouzé, F Calvo, and J Zucman-Rossi. Integrated analysis of somatic

mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat Genet*, 44(6):694-8., 2012.

[217] SA Forbes, D Beare, P Gunasekaran, K Leung, N Bindal, H Boutselakis, M Ding, S Bamford, C Cole, S Ward, CY Kok, M Jia, T De, JW Teague, MR Stratton, U McDermott, and PJ Campbell. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*, 43(Databaseissue):D805-11., 2015.

[218] JG Tate, S Bamford, HC Jubb, Z Sondka, DM Beare, N Bindal, H Boutselakis, CG Cole, C Creatore, E Dawson, P Fish, B Harsha, C Hathaway, SC Jupe, CY Kok, K Noble, L Ponting, CC Ramshaw, CE Rye, HE Speedy, R Stefancsik, SL Thompson, S Wang, S Ward, PJ Campbell, and SA Forbes. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*, 47(D1):D941-D947., 2019.

[219] A. D. Yates, P. Achuthan, W. Akanni, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, J. C. Marugán, C. Cummins, C. Davidson, K. Dodiya, R. Fatima, A. Gall, C. G. Giron, L. Gil, T. Grego, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, M. Kay, I. Lavidas, T. Le, D. Lemos, J. G. Martinez, T. Maurel, M. McDowall, A. McMahon, S. Mohanan, B. Moore, M. Nuhn, D. N. Oheh, A. Parker, A. Parton, M. Patricio, M. P. Sakthivel, A. I. Abdul Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, M. Sycheva, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, A. Vullo, B. Walts, A. Winterbottom, A. Zadissa, M. Chakiachvili, B. Flint, A. Frankish, S. E. Hunt, G. IIsley, M. Kostadima, N. Langridge, J. E. Loveland, F. J. Martin, J. Morales, J. M. Mudge, M. Muffato, E. Perry, M. Ruffier, S. J. Trevanion, F. Cunningham, K. L. Howe, D. R. Zerbino, and P. Flicek. Ensembl 2020. *Nucleic Acids Res*, 48(D1):D682–D688, 01 2020.

[220] NA O'Leary, MW Wright, JR Brister, S Ciufo, D Haddad, R McVeigh, B Rajput, B Robbertse, B Smith-White, D Ako-Adjei, A Astashyn, A Badretdin, Y Bao, O Blinkova, V Brover, V Chetvernin, J Choi, E Cox, O Ermolaeva, CM Farrell, T Goldfarb, T Gupta, D Haft, E Hatcher, W Hlavina, VS Joardar, VK Kodali, W Li, D Maglott, P Masterson, KM McGarvey, MR Murphy, K O'Neill, S Pujar, SH Rangwala, D Rausch, LD Riddick, C Schoch, A Shkeda, SS Storz, H Sun, F Thibaud-Nissen, I Tolstoy, RE Tully, AR Vatsan, C Wallin, D Webb, W Wu, MJ Landrum,

A Kimchi, T Tatusova, M DiCuccio, P Kitts, TD Murphy, and KD Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44(D1):D733-45., 2016.

[221] P. Sin-Chan, I. Mumal, T. Suwal, B. Ho, X. Fan, I. Singh, Y. Du, M. Lu, N. Patel, J. Torchia, D. Popovski, M. Fouladi, P. Guilhamon, J. R. Hansford, S. Leary, L. M. Hoffman, J. M. Mulcahy Levy, A. Lassaletta, P. Solano-Paez, E. Rivas, A. Reddy, G. Y. Gillespie, N. Gupta, T. E. Van Meter, H. Nakamura, T. T. Wong, Y. S. Ra, S. K. Kim, L. Massimi, R. G. Grundy, J. Fangusaro, D. Johnston, J. Chan, L. Lafay-Cousin, E. I. Hwang, Y. Wang, D. Catchpoole, J. Michaud, B. Ellezam, R. Ramanujachar, H. Lindsay, M. D. Taylor, C. E. Hawkins, E. Bouffet, N. Jabado, S. K. Singh, C. L. Kleinman, D. Barsyte-Lovejoy, X. N. Li, P. B. Dirks, C. Y. Lin, S. C. Mack, J. N. Rich, and A. Huang. A C19MC-LIN28A-MYCN Oncogenic Circuit Driven by Hijacked Super-enhancers Is a Distinct Therapeutic Vulnerability in ETMRs: A Lethal Brain Tumor. *Cancer Cell*, 36(1):51–67, 07 2019.

[222] CL Kleinman, N Gerges, S Papillon-Cavanagh, P Sin-Chan, A Pramatarova, DA Quang, V Adoue, S Busche, M Caron, H Djambazian, A Bemmo, AM Fontebasso, T Spence, J Schwartzentruber, S Albrecht, P Hauser, M Garami, A Klekner, L Bognar, JL Montes, A Staffa, A Montpetit, P Berube, M Zakrzewska, K Zakrzewski, PP Liberski, Z Dong, PM Siegel, T Duchaine, C Perotti, A Fleming, D Faury, M Remke, M Gallo, P Dirks, MD Taylor, R Sladek, T Pastinen, JA Chan, A Huang, J Majewski, and N Jabado. Fusion of TTYH1 with the C19MC microRNA cluster drives expression of a brain-specific DNMT3B isoform in the embryonal brain tumor ETMR. *Nat Genet*, 46(1):39-44., 2014.

[223] M Li, KF Lee, Y Lu, I Clarke, D Shih, C Eberhart, VP Collins, Meter T Van, D Picard, L Zhou, PC Boutros, P Modena, ML Liang, SW Scherer, E Bouffet, JT Rutka, SL Pomeroy, CC Lau, MD Taylor, A Gajjar, PB Dirks, CE Hawkins, and A Huang. Frequent amplification of a chr19q13.41 microRNA polycistron in aggressive primitive neuroectodermal brain tumors. *Cancer Cell*, 16(6):533-46., 2009.

[224] M Parker, KM Mohankumar, C Punchihewa, R Weinlich, JD Dalton, Y Li, R Lee, RG Tatevossian, TN Phoenix, R Thiruvenkatam, E White, B Tang, W Orisme, K Gupta, M Rusch, X Chen, Y Li, P Nagahawhatte, E Hedlund, D Finkelstein, G Wu,

S Shurtleff, J Easton, K Boggs, D Yergeau, B Vadodaria, HL Mulder, J Becksfort, P Gupta, R Huether, J Ma, G Song, A Gajjar, T Merchant, F Boop, AA Smith, L Ding, C Lu, K Ochoa, D Zhao, RS Fulton, LL Fulton, ER Mardis, RK Wilson, JR Downing, DR Green, J Zhang, DW Ellison, and RJ Gilbertson. C11orf95-RELA fusions drive oncogenic NF-$\varkappa$B signalling in ependymoma. *Nature*, 506(7489):451-5., 2014.

[225] AR Lawson, GF Hindley, T Forshew, RG Tatevossian, GA Jamie, GP Kelly, GA Neale, J Ma, TA Jones, DW Ellison, and D Sheer. RAF gene fusion breakpoints in pediatric brain tumors are characterized by significant enrichment of sequence microhomology. *Genome Res*, 21(4):505-14., 2011.

[226] Rubicon Genomics. *Targeted Capture of ThruPLEX Libraries with Agilent SureSelect XT Target Enrichment System*, 2016. RDM-152-002.

[227] Agilent Technologies Inc. *SureSelect XT Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library*, 2015. B4.

[228] R. Cronn, B. J. Knaus, A. Liston, P. J. Maughan, M. Parks, J. V. Syring, and J. Udall. Targeted enrichment strategies for next-generation plant biology. *Am J Bot*, 99(2): 291–311, Feb 2012.

[229] T. Suchan, C. Pitteloud, N. S. Gerasimova, A. Kostikova, S. Schmid, N. Arrigo, M. Pajkovic, M. Ronikier, and N. Alvarez. Hybridization Capture Using RAD Probes (hyRAD), a New Tool for Performing Genomic Analyses on Collection Specimens. *PLoS One*, 11(3):e0151651, 2016.

[230] R. Chen, H. Im, and M. Snyder. Whole-Exome Enrichment with the Agilent SureSelect Human All Exon Platform. *Cold Spring Harb Protoc*, 2015(7):626–633, Mar 2015.

[231] AR Quinlan and IM Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841-2., 2010.

[232] Horizon Discovery Group PLC. *cfDNA Reference Standard Set Product Specification*, 2015. 6068 PSS-01 (V-01).

[233] Rubicon Genomics. *ThruPLEX Tag-seq Kit Index Guide*, 2016. QAM-331-001.

[234] Rubicon Genomics. *ThruPLEX Tag-seq Kit Instruction Manual*, 2016. QAM-328-001.

[235] Agilent Technologies Inc. *SureSelect XT Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library*, 2016. Version B5.

[236] AM Newman, SV Bratman, J To, JF Wynne, NC Eclov, LA Modlin, CL Liu, JW Neal, HA Wakelee, RE Merritt, JB Shrager, BW Jr Loo, AA Alizadeh, and M Diehn. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med*, 20(5):548-54., 2014.

[237] Illumina Inc. *NextSeq 500 System WGS Solution*, 2014.

[238] Covaris Inc. *DNA Shearing with E220 Focused-ultrasonicator*, 2017. 010308 Rev N.

[239] W Rychlik, WJ Spencer, and RE Rhoads. Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res*, 18(21):6409–6412., 1990.

[240] MR Green and J Sambrook. Touchdown Polymerase Chain Reaction (PCR). *Cold Spring Harb Protoc*, 2018(5), 2018.

[241] H Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997v1, 2013.

[242] H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, and Genome Project Data Processing Subgroup 1000. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078-9., 2009.

[243] Broad Institute. Picard - A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF., 2016. `http://broadinstitute.github.io/picard` [Accessed: 2016-12-05].

[244] DC Koboldt, Q Zhang, DE Larson, D Shen, MD McLellan, L Lin, CA Miller, ER Mardis, L Ding, and RK Wilson. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*, 22(3):568-76., 2012.

[245] K Wang, M Li, and H Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16):e164., 2010.

[246] Z Chen, Y Yuan, X Chen, J Chen, S Lin, X Li, and H Du. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci Rep*, 10(1):3501., 2020.

[247] S McCarthy, S Das, W Kretzschmar, O Delaneau, AR Wood, A Teumer, HM Kang, C Fuchsberger, P Danecek, K Sharp, Y Luo, C Sidore, A Kwong, N Timpson, S Koskinen, S Vrieze, LJ Scott, H Zhang, A Mahajan, J Veldink, U Peters, C Pato, CM van Duijn, CE Gillies, I Gandin, M Mezzavilla, A Gilly, M Cocca, M Traglia, A Angius, JC Barrett, D Boomsma, K Branham, G Breen, CM Brummett, F Busonero, H Campbell, A Chan, S Chen, E Chew, FS Collins, LJ Corbin, GD Smith, G Dedoussis, M Dorr, AE Farmaki, L Ferrucci, L Forer, RM Fraser, S Gabriel, S Levy, L Groop, T Harrison, A Hattersley, OL Holmen, K Hveem, M Kretzler, JC Lee, M McGue, T Meitinger, D Melzer, JL Min, KL Mohlke, JB Vincent, M Nauck, D Nickerson, A Palotie, M Pato, N Pirastu, M McInnis, JB Richards, C Sala, V Salomaa, D Schlessinger, S Schoenherr, PE Slagboom, K Small, T Spector, D Stambolian, M Tuke, J Tuomilehto, LH Van den Berg, W Van Rheenen, U Volker, C Wijmenga, D Toniolo, E Zeggini, P Gasparini, MG Sampson, JF Wilson, T Frayling, PI de Bakker, MA Swertz, S McCarroll, C Kooperberg, A Dekker, D Altshuler, C Willer, W Iacono, S Ripatti, N Soranzo, K Walter, A Swaroop, F Cucca, CA Anderson, RM Myers, M Boehnke, MI McCarthy, R Durbin, and Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*, 48(10):1279-83., 2016.

[248] 1000 Genomes Project Consortium, A Auton, LD Brooks, RM Durbin, EP Garrison, HM Kang, JO Korbel, JL Marchini, S McCarthy, GA McVean, and GR Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68-74., 2015.

[249] G Glusman, J Caballero, DE Mauldin, L Hood, and JC Roach. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics*, 27(22):3216-7., 2011.

[250] JA Tennessen, AW Bigham, TD O'Connor, W Fu, EE Kenny, S Gravel, S McGee, R Do, X Liu, G Jun, HM Kang, D Jordan, SM Leal, S Gabriel, MJ Rieder,

G Abecasis, D Altshuler, DA Nickerson, E Boerwinkle, S Sunyaev, CD Bustamante, MJ Bamshad, JM Akey, Broad GO, Seattle GO, and NHLBI Exome Sequencing Project. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64-9., 2012.

[251] M Lek, KJ Karczewski, EV Minikel, KE Samocha, E Banks, T Fennell, AH O'Donnell-Luria, JS Ware, AJ Hill, BB Cummings, T Tukiainen, DP Birnbaum, JA Kosmicki, LE Duncan, K Estrada, F Zhao, J Zou, E Pierce-Hoffman, J Berghout, DN Cooper, N Deflaux, M DePristo, R Do, J Flannick, M Fromer, L Gauthier, J Goldstein, N Gupta, D Howrigan, A Kiezun, MI Kurki, AL Moonshine, P Natarajan, L Orozco, GM Peloso, R Poplin, MA Rivas, V Ruano-Rubio, SA Rose, DM Ruderfer, K Shakir, PD Stenson, C Stevens, BP Thomas, G Tiao, MT Tusie-Luna, B Weisburd, HH Won, D Yu, DM Altshuler, D Ardissino, M Boehnke, J Danesh, S Donnelly, R Elosua, JC Florez, SB Gabriel, G Getz, SJ Glatt, CM Hultman, S Kathiresan, M Laakso, S McCarroll, MI McCarthy, D McGovern, R McPherson, BM Neale, A Palotie, SM Purcell, D Saleheen, JM Scharf, P Sklar, PF Sullivan, J Tuomilehto, MT Tsuang, HC Watkins, JG Wilson, MJ Daly, DG MacArthur, and Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285-91., 2016.

[252] H Li and R Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754-60., 2009.

[253] B Langmead and SL Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357-9., 2012.

[254] H Thorvaldsdóttir, JT Robinson, and JP Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, 14(2):178-92., 2013.

[255] Bio-Rad Laboratories Inc. *ddPCR Supermix for Probes*. 10026235 Rev C.

[256] SR Kennedy and B Kohrn. Duplex Sequencing Pipeline, 2020. `https://github.com/Kennedy-Lab-UW/Duplex-Seq-Pipeline` [accessed: 2021-05-21].

[257] TaKaRa Bio Inc. ThruPLEX Tag-seq detection sensitivity and specificity using Horizon cfDNA reference standards, 2017. https://www.takarabio.com/learning-centers/next-generation-sequencing/technical-notes/dna-seq/tag-seq-variant-detection [accessed: 2021-05-16].

[258] T Goschzik, M Gessi, V Dreschmann, U Gebhardt, L Wang, S Yamaguchi, DA Wheeler, L Lauriola, CC Lau, HL Müller, and T Pietsch. Genomic Alterations of Adamantinomatous and Papillary Craniopharyngioma. *J Neuropathol Exp Neurol*, 76(2):126-134., 2017.

[259] PA Northcott, I Buchhalter, AS Morrissy, V Hovestadt, J Weischenfeldt, T Ehrenberger, S Gröbner, M Segura-Wang, T Zichner, VA Rudneva, HJ Warnatz, N Sidiropoulos, AH Phillips, S Schumacher, K Kleinheinz, SM Waszak, S Erkek, DTW Jones, BC Worst, M Kool, M Zapatka, N Jäger, L Chavez, B Hutter, M Bieg, N Paramasivam, M Heinold, Z Gu, N Ishaque, C Jäger-Schmidt, CD Imbusch, A Jugold, D Hübschmann, T Risch, V Amstislavskiy, FGR Gonzalez, UD Weber, S Wolf, GW Robinson, X Zhou, G Wu, D Finkelstein, Y Liu, FMG Cavalli, B Luu, V Ramaswamy, X Wu, J Koster, M Ryzhova, YJ Cho, SL Pomeroy, C Herold-Mende, M Schuhmann, M Ebinger, LM Liau, J Mora, RE McLendon, N Jabado, T Kumabe, E Chuah, Y Ma, RA Moore, AJ Mungall, KL Mungall, N Thiessen, K Tse, T Wong, SJM Jones, O Witt, T Milde, Deimling A Von, D Capper, A Korshunov, ML Yaspo, R Kriwacki, A Gajjar, J Zhang, R Beroukhim, E Fraenkel, JO Korbel, B Brors, M Schlesner, R Eils, MA Marra, SM Pfister, MD Taylor, and P Lichter. The whole-genome landscape of medulloblastoma subtypes. *Nature*, 547(7663):311-317., 2017.

[260] Illumina Inc. Performance specifications for the HiSeq 2500 System, 2021. https://emea.illumina.com/systems/sequencing-platforms/hiseq-2500/specifications.html [accessed: 2021-04-15].

[261] B. Ricciuti, G. Jones, M. Severgnini, J. V. Alessi, G. Recondo, M. Lawrence, T. Forshew, C. Lydon, M. Nishino, M. Cheng, and M. Awad. Early plasma circulating tumor DNA (ctDNA) changes predict response to first-line pembrolizumab-based

therapy in non-small cell lung cancer (NSCLC). *J Immunother Cancer*, 9(3), Mar 2021.

[262] N. B. Leighl, R. D. Page, V. M. Raymond, D. B. Daniel, S. G. Divers, K. L. Reckamp, M. A. Villalona-Calero, D. Dix, J. I. Odegaard, R. B. Lanman, and V. A. Papadimitrakopoulou. Clinical Utility of Comprehensive Cell-free DNA Analysis to Identify Genomic Biomarkers in Patients with Newly Diagnosed Metastatic Non-small Cell Lung Cancer. *Clin Cancer Res*, 25(15):4691–4700, Aug 2019.

[263] GA Auwera and BD Geraldine. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, Inc, Farnham, United Kingdom, 2020. ISBN 9781491975190.

[264] C. D. Warden, A. W. Adamson, S. L. Neuhausen, and X. Wu. Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ*, 2:e600, 2014.

[265] S. Tian, H. Yan, M. Kalmbach, and S. L. Slager. Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics*, 17(1):403, Oct 2016.

[266] V Potapov and JL Ong. Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLoS One*, 12(1):e0169774., 2017.

[267] NHLBI GO Exome Sequencing Project. *Exome Variant Server (URL: http://evs.gs.washington.edu/EVS/)*. [(09, 2019) accessed].

[268] MJ Landrum, JM Lee, M Benson, GR Brown, C Chao, S Chitipiralla, B Gu, J Hart, D Hoffman, W Jang, K Karapetyan, K Katz, C Liu, Z Maddipatla, A Malheiro, K McDaniel, M Ovetsky, G Riley, G Zhou, JB Holmes, BL Kattman, and DR Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*, 46(D1):D1062-D1067., 2018.

[269] KW Eaton, LS Tooke, LM Wainwright, AR Judkins, and JA Biegel. Spectrum of SMARCB1/INI1 mutations in familial and sporadic rhabdoid tumors. *Pediatr Blood Cancer*, 56(1):7-15., 2011.

[270] LF Johansson, Dijk F van, Boer EN de, KK van Dijk-Bos, JD Jongbloed, AH van der Hout, H Westers, RJ Sinke, MA Swertz, RH Sijmons, and B Sikkema-Raddatz. CoNVaDING: Single Exon Variation Detection in Targeted NGS Data. *Hum Mutat*, 37(5):457-64., 2016.

[271] V Plagnol, J Curtis, M Epstein, KY Mok, E Stebbings, S Grigoriadou, NW Wood, S Hambleton, SO Burns, AJ Thrasher, D Kumararatne, R Doffinger, and S Nejentsev. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, 28(21):2747-54., 2012.

[272] G Povysil, A Tzika, J Vogt, V Haunschmid, L Messiaen, J Zschocke, G Klambauer, S Hochreiter, and K Wimmer. panelcn.MOPS: Copy-number detection in targeted NGS panel data for clinical diagnostics. *Hum Mutat*, 38(7):889-897., 2017.

[273] A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*, 21(6):974–984, Jun 2011.

[274] E. Talevich, A. H. Shain, T. Botton, and B. C. Bastian. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol*, 12(4):e1004873, Apr 2016.

[275] I Sanchez-Navarro, J da Silva L R, F Blanco-Kelly, O Zurita, N Sanchez-Bolivar, C Villaverde, MI Lopez-Molina, B Garcia-Sandoval, S Tahsin-Swafiri, P Minguez, R Riveiro-Alvarez, I Lorda, R Sanchez-Alcudia, R Perez-Carro, D Valverde, Y Liu, L Tian, H Hakonarson, A Avila-Fernandez, M Corton, and C Ayuso. Combining targeted panel-based resequencing and copy-number variation analysis for the diagnosis of inherited syndromic retinopathies and associated ciliopathies. *Sci Rep*, 8(1):5285., 2018.

[276] M. Alcaide, M. Cheung, J. Hillman, S. R. Rassekh, R. J. Deyell, G. Batist, A. Karsan, A. W. Wyatt, N. Johnson, D. W. Scott, and R. D. Morin. Evaluating the quantity, quality and size distribution of cell-free DNA by multiplex droplet digital PCR. *Sci Rep*, 10(1):12564, 07 2020.

[277] N. P. Carter. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet*, 39(7 Suppl):16–21, Jul 2007.

[278] P. v. Ijssel and B. Ylstra. Oligonucleotide array comparative genomic hybridization. *Methods Mol Biol*, 396:207–221, 2007.

[279] W. W. Lockwood, R. Chari, B. Chi, and W. L. Lam. Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *Eur J Hum Genet*, 14(2):139–148, Feb 2006.

[280] J. P. Schouten, C. J. McElgunn, R. Waaijer, D. Zwijnenburg, F. Diepvens, and G. Pals. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res*, 30(12):e57, Jun 2002.

[281] A. Benard-Slagter, I. Zondervan, K. de Groot, F. Ghazavi, V. Sarhadi, P. Van Vlierberghe, B. De Moerloose, C. Schwab, K. Vettenranta, C. J. Harrison, S. Knuutila, J. Schouten, T. Lammens, and S. Savola. Digital Multiplex Ligation-Dependent Probe Amplification for Detection of Key Copy Number Alterations in T- and B-Cell Lymphoblastic Leukemia. *J Mol Diagn*, 19(5):659–672, 09 2017.

[282] ZY Han, W Richer, P Fréneaux, C Chauvin, C Lucchesi, D Guillemot, C Grison, D Lequin, G Pierron, J Masliah-Planchon, A Nicolas, D Ranchère-Vince, P Varlet, S Puget, I Janoueix-Lerosey, O Ayrault, D Surdez, O Delattre, and F Bourdeaut. The occurrence of intracranial rhabdoid tumours in mice depends on temporal control of Smarcb1 inactivation. *Nat Commun*, 7:10421., 2016.

[283] A Klochendler-Yeivin, L Fiette, J Barra, C Muchardt, C Babinet, and M Yaniv. The murine SNF5/INI1 chromatin remodeling factor is essential for embryonic development and tumor suppression. *EMBO Rep*, 1(6):500-6., 2000.

[284] CJ Guidi, AT Sands, BP Zambrowicz, TK Turner, DA Demers, W Webster, TW Smith, AN Imbalzano, and SN Jones. Disruption of Ini1 leads to peri-implantation lethality and tumorigenesis in mice. *Mol Cell Biol*, 21(10):3598-603., 2001.

[285] C Girardot, J Scholtalbers, S Sauer, SY Su, and EE Furlong. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinformatics*, 17(1):419., 2016.

[286] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods*, 11(2):163–166, Feb 2014.

[287] N. Stoler, B. Arbeithuber, W. Guiblet, K. D. Makova, and A. Nekrutenko. Streamlined analysis of duplex sequencing data with Du Novo. *Genome Biol*, 17(1):180, 08 2016.

[288] S. Kosugi, Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*, 20(1):117, 06 2019.

[289] LE MacConaill, RT Burns, A Nag, HA Coleman, MK Slevin, K Giorda, M Light, K Lai, M Jarosz, MS McNeill, MD Ducar, M Meyerson, and AR Thorner. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics*, 19(1):30., 2018.

[290] Rahul Sinha, Geoff Stanley, Gunsagar S. Gulati, Camille Ezran, Kyle J. Travaglini, Eric Wei, Charles K.F. Chan, Ahmad N. Nabhan, Tianying Su, Rachel M. Morganti, Stephanie D. Conley, Hassan Chaib, Kristy Red-Horse, Michael T. Longaker, Michael P. Snyder, Mark A. Krasnow, and Irving L. Weissman. Index switching causes "spreading-of-signal" among multiplexed samples in illumina hiseq 4000 dna sequencing. *bioRxiv*, 2017. URL `https://www.biorxiv.org/content/early/2017/04/09/125724`.

[291] D. T. Cheng, T. N. Mitchell, A. Zehir, R. H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z. Y. Liu, H. H. Won, S. N. Scott, A. R. Brannon, C. O'Reilly, J. Sadowska, J. Casanova, A. Yannes, J. F. Hechtman, J. Yao, W. Song, D. S. Ross, A. Oultache, S. Dogan, L. Borsu, M. Hameed, K. Nafa, M. E. Arcila, M. Ladanyi, and M. F. Berger. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J Mol Diagn*, 17 (3):251–264, May 2015.

[292] Thermo Fisher Scientific. *Oncomine Comprehensive Assay Plus User Guide*, 2021. C.0.

[293] Thermo Fisher Scientific. Oncomine Comprehensive Assay Plus, manual library preparation, 2021. `https://www.thermofisher.com/order/catalog/product/A48577?SID=srch-hj-A48577` [accessed: 2021-06-16].

[294] E. Samorodnitsky, B. M. Jewell, R. Hagopian, J. Miya, M. R. Wing, E. Lyon, S. Damodaran, D. Bhatt, J. W. Reeser, J. Datta, and S. Roychowdhury. Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Hum Mutat*, 36(9):903–914, Sep 2015.

[295] S. R. Head, H. K. Komori, S. A. LaMere, T. Whisenant, F. Van Nieuwerburgh, D. R. Salomon, and P. Ordoukhanian. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, 56(2):61–64, 2014.

[296] Agilent Technologies Inc. *SureSelect XT HS2 DNA System*, 2021. D0.

[297] I. Yamaguchi, T. Watanabe, O. Ohara, and Y. Hasegawa. PCR-free whole exome sequencing: Cost-effective and efficient in detecting rare mutations. *PLoS One*, 14 (9):e0222562, 2019.

[298] Integrated DNA Technologies Inc. Reverse Complement Adapters for the Mitigation of UMI Hopping, May 2018. URL `\url{https://patents.justia.com/patent/20180340216}`. US Patent 20,180,340,216.

[299] D. C. García-Olmo, L. Gutiérrez-González, R. Ruiz-Piqueras, M. G. Picazo, and D. García-Olmo. Detection of circulating tumor cells and of tumor DNA in plasma during tumor progression in rats. *Cancer Lett*, 217(1):115–123, Jan 2005.

[300] C. Rago, D. L. Huso, F. Diehl, B. Karim, G. Liu, N. Papadopoulos, Y. Samuels, V. E. Velculescu, B. Vogelstein, K. W. Kinzler, and L. A. Diaz. Serial assessment of human tumor burdens in mice by the analysis of circulating DNA. *Cancer Res*, 67 (19):9364–9370, Oct 2007.

[301] A. A. Kamat, F. Z. Bischoff, D. Dang, M. F. Baldwin, L. Y. Han, Y. G. Lin, W. M. Merritt, C. N. Landen, C. Lu, D. M. Gershenson, J. L. Simpson, and A. K. Sood.

Circulating cell-free DNA: a novel biomarker for response to therapy in ovarian carcinoma. *Cancer Biol Ther*, 5(10):1369–1374, Oct 2006.

[302] F. E. Boas, F. Nurili, A. Bendet, C. Cheleuitte-Nieves, O. Basturk, G. Askan, A. O. Michel, S. Monette, E. Ziv, C. T. Sofocleous, A. W. P. Maxwell, L. B. Schook, S. B. Solomon, D. P. Kelsen, A. Scherz, and H. Yarmohammadi. Induction and characterization of pancreatic cancer in a transgenic pig model. *PLoS One*, 15(9):e0239391, 2020.

[303] L. B. Schook, T. V. Collares, W. Hu, Y. Liang, F. M. Rodrigues, L. A. Rund, K. M. Schachtschneider, F. K. Seixas, K. Singh, K. D. Wells, E. M. Walters, R. S. Prather, and C. M. Counter. A Genetic Porcine Model of Cancer. *PLoS One*, 10(7):e0128864, 2015.

[304] N. Robertson, L. B. Schook, and K. M. Schachtschneider. Porcine cancer models: potential tools to enhance cancer drug trials. *Expert Opin Drug Discov*, 15(8):893–902, 08 2020.

[305] N. Nakamura, E. Hatano, K. Iguchi, M. Sato, H. Kawaguchi, I. Ohtsu, T. Sakurai, N. Aizawa, H. Iijima, S. Nishiguchi, T. Tomono, Y. Okuda, S. Wada, S. Seo, K. Taura, S. Uemoto, and M. Ikegawa. Elevated levels of circulating ITIH4 are associated with hepatocellular carcinoma with nonalcoholic fatty liver disease: from pig model to human study. *BMC Cancer*, 19(1):621, Jun 2019.